

Guohua Li · Susan P. Baker *Editors*

Injury Research

Theories, Methods, and Approaches

Injury Research

Guohua Li • Susan P. Baker
Editors

Injury Research

Theories, Methods, and Approaches

Editors

Guohua Li, MD, DrPH
Department of Anesthesiology
College of Physicians and Surgeons
Department of Epidemiology
Mailman School of Public Health
Columbia University
New York, NY10032, USA
gl2240@columbia.edu

Susan P. Baker, MPH, ScD (Hon.)
Center for Injury Research and Policy
Johns Hopkins Bloomberg School
of Public Health
Baltimore, MD 21205, USA
sbaker@jhsph.edu

ISBN 978-1-4614-1598-5 e-ISBN 978-1-4614-1599-2
DOI 10.1007/978-1-4614-1599-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011943885

© Springer Science+Business Media, LLC 2012

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

*To Dr. William Haddon Jr. and other pioneers who made injury no longer
an accident.*

*G.L.
S.P.B.*

Foreword

This book is a milestone in the field of injury and violence prevention in that it provides a comprehensive look at the various theories and methods that are used to perform injury research. Beginning with the building of data systems to conduct injury surveillance for identifying and monitoring injury, and documenting methods for examining injury causation and injury outcomes, it gives a state-of-the-art picture of where the field of injury research stands. By documenting analytical approaches to injury research, it provides guidance in the various methods that may be used to assess injury events and interventions and then describes the methodological approaches to decreasing injury burden.

Dr. Li and Professor Baker continue to be leaders in the field of injury research, and have assembled an internationally recognized cadre of injury researchers who have contributed to the book. The selection of authors from multiple disciplines highlights the breadth and diversity of the disciplines involved in the field of injury research. From epidemiologists to clinicians and economists, from basic scientists to legal experts and behavioral scientists, the need for a multidisciplinary approach to the problem of injury is made clear. And, unlike many other fields, where each discipline speaks its own language, the authors of the text speak in a common language – that of the field of injury prevention and control.

One of the remarkable features of this book is the way the information is presented. The writing and information are such that the content can be understood by someone who is entering the field of injury research as a student or an early-career scientist, but is also valuable to the senior researcher who has already made significant contributions to the knowledge base of injury research. The focus is not on a single method or phase of injury research, but moves from the laboratory setting to the community and policy environments, and targets the translation and dissemination of injury research as critical to building the field.

The first section of the book, which focuses on surveillance, provides a strong foundation for the remainder of the methodological discussions, and also for anyone who is interested in injury surveillance. The discussions and descriptions of injury causation research methods, including explanations of forensic issues and qualitative and quantitative methods, allow a comprehensive approach to exploring the factors that contribute to injury, from individual behavior, human body tolerance to forces, and the physical environment in which injuries occur. Outcomes, ranging from anatomic injury severity, clinical outcomes and system impacts, are discussed in sufficient detail to aid the reader in understanding the wide range of outcomes that are important in injury research. Analytic approaches include approaches that are emerging because of advancing technology or social interactions. Finally, the injury reduction approaches, when taken together, give us a picture of the true nature of what is needed to solve injury problems.

The richness of the text is in the explanations of various theories and research methods, and in the descriptions of how research methods are successfully applied in injury research. The book serves as a guide that will not remain on the shelf, but will be referenced time and time again by injury researchers, students, and others who are interested in injury and its toll.

Linda C. Degutis, DrPH, MSN
Director
National Center for Injury Prevention and Control
Centers for Disease Control and Prevention
Atlanta, GA, USA

Preface

In 1964, William Haddon, Jr., Edward A. Suchman, and David Klein forged a book, titled *Accident Research: Methods and Approaches*, to foster the establishment of accident research as a scientific discipline. Their book was extraordinary for its time because it brought together a variety of applied research methods for understanding the causes and prevention of accidents, illustrated through illuminating examples from published studies. For many years, it served as the only resource book on research methodology available in the field. Since then, the field of accident research has witnessed tremendous transformations and growth in both scope and depth. Among the most profound changes is the increasing acceptance of the view that injury is no accident. For centuries, the fatalistic view that injuries were accidents resulting from bad luck, malevolence, or simply “acts of god” prevailed. Research in the past four decades, however, has provided undisputable evidence that injury is predictable, preventable, and treatable, and that even in an event such as a crash, fall, or shooting, the risk, severity, and outcome of injury is modifiable through effective interventions. As a result, injury is now widely recognized as a health problem, and in the field of public health and medicine, the word *accident* is generally replaced by *injury*. The purpose of this edited volume is to provide the reader with a contemporary reference text on injury research methods.

This book consists of 36 individual chapters written by some of the most accomplished injury researchers in the world. These chapters are organized in five parts. Part I contains four chapters concerning injury surveillance. Systematic collection, analysis, and dissemination of mortality, morbidity, and exposure data based on well-established health information systems are essential for monitoring the trends and patterns of injury epidemiology and for developing and evaluating intervention programs. As a basic epidemiologic method and an imperative public health function, surveillance plays a pivotal role in injury research. These four chapters discuss major methodological and technical issues in injury surveillance, including data systems, injury classifications, applications of information technology and innovative methods, special populations, and high-impact topical areas.

Part II comprises eight chapters covering a wide range of theories and methods for understanding the causes of injury. Contributed by experts from forensic pathology, ergonomics, engineering, psychology, epidemiology, and behavioral science, these chapters provide a multidisciplinary exposition of the various concepts and methods used by injury researchers and practitioners working in different fields. Among the topics discussed in this section are experimental and observational designs and qualitative methods.

Part III is made up of seven chapters on research methods pertinent to injury consequences. It begins with an introduction to the Barell matrix for standardized multiple injury profiling, proceeds to explain methods for measuring injury severity, triaging and managing injury patients in emergency care settings, and evaluating diagnostic and prognostic biomarkers in trauma care. The section concludes with explorations of the conceptual and theoretical frameworks underlying the

International Classification of Function and the methods for quantifying the economic costs of injury. This section should be especially informative and relevant to clinical and translational researchers as well as health services researchers.

Part IV features seven chapters on statistical and analytical techniques especially relevant to injury research, including video data analysis, age–period–cohort modeling, multilevel modeling, geographic information systems and spatial regression, and social network analysis. These chapters are not meant to provide an exhaustive presentation of quantitative methods. Rather, they highlight the advances in a few select analytical techniques readily applicable to injury data.

Part V contains ten chapters discussing the theories and methods underpinning various approaches to injury prevention and control. The first two chapters in this section provide an overview of the legal and economic frameworks for improving public safety through policy interventions. The subsequent four chapters explain the environmental, technological, behavioral, and medical approaches to injury control. The final four chapters address methodological and technical issues in injury research related to medical error, resource constraints, and program evaluation. The reader will find these chapters intellectually stimulating and practically instructive.

Despite the remarkable growth in recent decades, injury research has been largely insulated by invisible disciplinary boundaries, and scientific advances are hindered by limited understanding and collaboration across disciplines. Given the complexity of injury causation and prevention, an interdisciplinary approach is imperative for the future of injury research. By drawing on expertise from different disciplines, we hope that this book will serve as a reference resource as well as a bridge to interdisciplinary and transdisciplinary understanding and collaboration among injury researchers.

We thank the contributing authors for their expertise and collegiality. All of them are active researchers with many competing responsibilities. It is no small undertaking to write the chapter manuscripts and go through several rounds of revisions. Their cooperation and commitment are greatly appreciated. We also thank Ms. Kristine Queja, publishing editor at Springer, for her trust, guidance, and support. She first approached us to discuss the book project at the annual meeting of the American Public Health Association in Philadelphia in 2009 and since has helped us at every step along the way to the finish line. Finally, we would like to thank Ms. Barbara H. Lang for her administrative and editorial assistance. Without her organizational and coordinating skills, we might never see this project come to fruition.

New York, NY, USA
Baltimore, MD, USA

Guohua Li, MD, DrPH
Susan P. Baker, MPH, ScD (Hon.)

Contents

Part I Injury Surveillance

1 Surveillance of Injury Mortality	3
Margaret Warner and Li-Hui Chen	
2 Surveillance of Injury Morbidity	23
Li-Hui Chen and Margaret Warner	
3 Injury Surveillance in Special Populations	45
R. Dawn Comstock	
4 Surveillance of Traumatic Brain Injury	61
Jean A. Langlois Orman, Anbesaw W. Selassie, Christopher L. Perdue, David J. Thurman, and Jess F. Kraus	

Part II Injury Causation

5 Forensic Pathology	89
Ling Li	
6 Determination of Injury Mechanisms	111
Dennis F. Shanahan	
7 Ergonomics	139
Steven Wiker	
8 Experimental Methods	187
Jonathan Howland and Damaris J. Rohsenow	
9 Epidemiologic Methods	203
Guohua Li and Susan P. Baker	
10 Qualitative Methods	221
Shannon Frattaroli	
11 Environmental Determinants	235
Shanthi Ameratunga and Jamie Hosking	
12 Behavioral Determinants	255
Deborah C. Girasek	

Part III Injury Outcome

13 Injury Profiling	269
Limor Aharonson-Daniel	
14 Injury Severity Scaling	281
Maria Seguí-Gómez and Francisco J. Lopez-Valdes	
15 Triage	297
Craig Newgard	
16 Clinical Prediction Rules	317
James F. Holmes	
17 Biomarkers of Traumatic Injury	337
Cameron B. Jeter, John B. Redell, Anthony N. Moore, Georgene W. Hergenroeder, Jing Zhao, Daniel R. Johnson, Michael J. Hylin, and Pramod K. Dash	
18 Functional Outcomes	357
Renan C. Castillo	
19 Injury Costing Frameworks	371
David Bishai and Abdulgafoor M. Bachani	

Part IV Analytical Approaches

20 Statistical Considerations	383
Shrikant I. Bangdiwala and Baishakhi Banerjee Taylor	
21 Video Data Analysis	397
Andrew E. Lincoln and Shane V. Caswell	
22 Age–Period–Cohort Modeling	409
Katherine M. Keyes and Guohua Li	
23 Multilevel Modeling	427
David E. Clark and Lynne Moore	
24 Geographical Information Systems	447
Becky P.Y. Loo and Shenjun Yao	
25 Spatial Regression	465
Jurek Grabowski	
26 Social Network Analysis	475
Paul D. Juarez and Lorien Jasny	

Part V Approaches to Injury Reduction

27 Legal Approach	495
Tom Christoffel	
28 Public Policy	507
David Hemenway	

29 Environmental Approach 519
 Leon S. Robertson

30 Technological Approach 529
 Flaura K. Winston, Kristy B. Arbogast, and Joseph Kaniathra

31 Behavioral Approach 549
 Andrea Carlson Gielen, Eileen M. McDonald, and Lara B. McKenzie

32 EMS and Trauma Systems 569
 Lenora M. Olson and Stephen M. Bowman

33 Systems Approach to Patient Safety 583
 Sneha Shah, Michelle Patch and Julius Cuong Pham

34 Intervention in Low-Income Countries 599
 Samuel N. Forjuoh

35 Implementing and Evaluating Interventions 619
 Caroline F. Finch

36 Economic Evaluation of Interventions 641
 Ted R. Miller and Delia Hendrie

Index 667

Abbreviations

ADR	Adverse Drug Reaction
AHRQ	Agency for Healthcare Research and Quality
AIS	Abbreviated Injury Scale
AL	Action Limit
ARRA	American Recovery and Reinvestment Act of 2009
ATD	Anthropomorphic Test Device
BA	Biochemical Analysis
BBB	Blood-Brain Barrier
BCR	Benefit–Cost Ratio
BRFSS	Behavioral Risk Factor Surveillance System
BSI	Bloodstream Infection
CAPI	Computer-Assisted Personal Interviewing
CASI	Computer-Assisted Self-Interviewing
CBA	Cost–Benefit Analysis
CDC	Centers for Disease Control and Prevention
CEA	Cost–Effectiveness Analysis
CER	Cost–Effectiveness Ratio
CFOI	Census of Fatal Occupational Injury
CIREN	Crash Injury Research and Engineering Network
CODES	Crash Outcome Data Evaluation System
COF	Coefficient of Friction
CPSC	Consumer Product Safety Commission
CR	Cardiac Rate
CSF	Cerebrospinal Fluid
CT	Computed Tomography
CUA	Cost–Utility Analysis
DALY	Disability-Adjusted Life Year
DB	Dry Bulb
E-Code	ICD External Cause of Injury and Poisoning Code
ED	Emergency Department
EMR/EHR	Electronic Medical Records/Electronic Health Records
EMS	Emergency Medical Services
FARS	Fatality Analysis Reporting System
GDP	Gross Domestic Product
HCUP	Healthcare Cost and Utilization Project
HCUP-NIS	Healthcare Cost and Utilization Project Nationwide Inpatient Sample
HDDS	Electronic Hospital Discharge Data System

HEDDS	Hospital ED Data System
HR	Heart Rate
HSR	Harm Susceptibility Ratio
ICD	International Classification of Diseases
ICD-9-CM	Ninth Revision of the International Classification of Diseases, Clinical Modification
ICD-10-AM	Tenth Revision of the International Classification of Diseases, Australian Modification
ICD-10-CM	Tenth Revision of the International Classification of Diseases, Clinical Modification
ICE	International Collaborative Effort
ICECI	International Classification of External Causes of Injury
ICER	Incremental Cost–Effectiveness Ratio
ICF	International Classification of Function
ICISS	International Classification of Diseases–Based Injury Severity Score
ICU	Intensive Care Unit
ISS	Injury Severity Score
LMF	Localized Muscle Fatigue
LMICs	Low- and Middle-Income Countries
LOC	Loss of Consciousness
MCD	Multiple Cause of Death Data
ME	Medical Examiner
MEP	Metabolic Energy Prediction
MEPS	Medical Expenditure Panel Survey
MPL	Maximum Permissible Limit
MR	Metabolic Rate
MRI	Magnetic Resonance Imaging
NAMCS	NCHS Ambulatory Medical Care Survey
NASS	National Automotive Sampling System
NASS-GES	National Automotive Sampling System–General Estimates System
NCAP	New Car Assessment Program
NC DETECT	North Carolina disease event tracking and epidemiologic collection
NCHS	National Center for Health Statistics
NCIS	National Coroners Information System
NDI	National Death Index
NEISS	National Electronic Injury Surveillance System
NEISS-AIP	National Electronic Injury Surveillance System–All Injury Program
NEMSIS	National Emergency Medical Services Information System
NFIRS	National Fire Incident Reporting System
NFPA	National Fire Protection Association
NHAMCS	National Hospital Ambulatory Medical Care Survey
NHANES	National Health and Nutrition Examination Survey
NHDS	National Hospital Discharge Survey
NHIS	National Health Interview Survey
NHTSA	National Highway Traffic Safety Administration
NIOSH	National Institute for Occupational Safety and Health
NIS	Nationwide Inpatient Sample
NISS	New Injury Severity Score
NSCOT	National Study of the Costs and Outcomes of Trauma
NSDUH	National Survey on Drug Use and Health
NTDB	National Trauma Data Bank
NVDRS	National Violent Death Reporting System
NVSS	National Vital Statistics System

NWB	Natural Wet-Bulb Temperature
PMHS	Postmortem Human Subject
PPPA	Poison Prevention Packaging Act
PRR	Proportional Reporting Ratio
PTA	Post-Traumatic Amnesia
PV	Present Value
QALY	Quality-Adjusted Life Year
QI	Quality Improvement
RSE	Relative Standard Error
SDT	Signal Detection Theory
STIPDA	State and Territorial Injury Prevention Directors Association
TBI	Traumatic Brain Injury
TRISS	Trauma and Injury Severity Score
UB-04	2004 Uniform Billing Form
USA	United States of America
UV	Ultraviolet
VSL	Value of Statistical Life
WBGT	Wet-Bulb-Globe-Thermometer
WHO	World Health Organization
WISQARS	Web-Based Injury Statistics Query and Reporting System
YPLL	Years of Potential Life Lost

Contributors

Limor Aharonson-Daniel, PhD Department of Emergency Medicine, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

PREPARED Center for Emergency Response Research, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Shanthi Ameratunga, MBChB, PhD Section of Epidemiology and Biostatistics, School of Population Health, University of Auckland, Auckland, New Zealand

Kristy B. Arbogast, PhD Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abdulgafoor M. Bachani, PhD, MHS International Injury Research Unit, Health Systems Program, Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Susan P. Baker, MPH, ScD (Hon.) Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Shrikant I. Bangdiwala, PhD Department of Biostatistics and Injury Prevention Research Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

David Bishai, MD, PhD, MPH Center for Injury Research and Policy and International Injury Research Unit, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Stephen M. Bowman, PhD Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Renan C. Castillo, PhD Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Shane V. Caswell, PhD George Mason University, Manassas, VA, USA

Li-Hui Chen, PhD Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Tom Christoffel, JD Boulder, CO, USA

David E. Clark, MD Maine Medical Center, Portland, ME, USA

Harvard Injury Control Research Center, Harvard School of Public Health, Boston, MA, USA

R. Dawn Comstock, PhD Center for Injury Research and Policy, The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA
Department of Pediatrics, College of Medicine, The Ohio State University, Columbus, OH, USA
Division of Epidemiology, College of Public Health, The Ohio State University, Columbus, OH, USA

Pramod K. Dash, PhD Department of Neurobiology & Anatomy, The University of Texas Medical School at Houston, Houston, TX, USA

Caroline F. Finch, PhD Australian Center for Research into Injury in Sport and its Prevention, Monash Injury Research Institute, Monash University, Clayton, VIC, Australia

Samuel N. Forjuoh, MD, DrPH, MPH Department of Family & Community Medicine, Scott & White Healthcare, Texas A&M Health Science Center College of Medicine, Temple, TX, USA

Shannon Frattaroli, PhD, MPH Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Andrea Carlson Gielen, ScD, ScM Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Deborah C. Girasek, PhD, MPH Department of Preventive Medicine & Biometrics, Uniformed Services University of the Health Sciences, Bethesda, MD, USA

Jurek Grabowski, PhD AAA Foundation for Traffic Safety, Washington, DC, USA

David Hemenway, PhD Harvard Injury Control Research Center, Harvard School of Public Health, Boston, MA, USA

Delia Hendrie, MA Population Health Research, Curtin Health Innovation Research Institute (CHIRI), Curtin University, Perth, WA, Australia

Georgene W. Hergenroeder, RN, MHA The Vivian L. Smith Department of Neurosurgery, The University of Texas Medical School at Houston, Houston, TX, USA

James F. Holmes, MD, MPH Department of Emergency Medicine, University of California at Davis School of Medicine, Sacramento, CA, USA

Jamie Hosking, MBChB, MPH Section of Epidemiology and Biostatistics, School of Population Health, University of Auckland, Auckland, New Zealand

Jonathan Howland, PhD, MPH, MPA Department of Emergency Medicine, Boston Medical Center, Boston University School of Medicine, Boston, MA, USA

Michael J. Hylin, PhD Department of Neurobiology & Anatomy, The University of Texas Medical School at Houston, Houston, TX, USA

Lorien Jasny, MA, PhD Department of Sociology, University of California at Irvine, Irvine, CA, USA

Cameron B. Jeter, PhD Department of Neurobiology & Anatomy, The University of Texas Medical School at Houston, Houston, TX, USA

Daniel R. Johnson, PhD Department of Neurobiology & Anatomy, The University of Texas Medical School at Houston, Houston, TX, USA

Paul D. Juarez, PhD Department of Family and Community Medicine,
Meharry Medical College, Nashville, TN, USA

Joseph Kaniathra, PhD Active Safety Engineering, LLC, Ashburn, VA, USA

Katherine M. Keyes, PhD Department of Epidemiology, Columbia University
Mailman School of Public Health, New York, NY, USA

Jess F. Kraus, PhD, MPH Department of Epidemiology, University of California at
Los Angeles, Los Angeles, CA, USA

Jean A. Langlois Orman, ScD, MPH Statistics and Epidemiology,
US Army Institute of Surgical Research, Houston, TX, USA

Guohua Li, MD, DrPH Department of Epidemiology, Columbia University Mailman School
of Public Health, New York, NY, USA

Department of Anesthesiology, Columbia University College of Physicians and Surgeons,
New York, NY, USA

Ling Li, MD Office of the Chief Medical Examiner, State of Maryland, Baltimore, MD, USA

Andrew E. Lincoln, ScD, MS MedStar Sports Medicine Research Center,
MedStar Health Research Institute, Union Memorial Hospital, Baltimore, MD, USA

Becky P.Y. Loo, PhD Department of Geography, The University of Hong Kong,
Hong Kong, China

Francisco J. Lopez-Valdes, BEng Center for Applied Biomechanics, University of Virginia,
Charlottesville, VA, USA

Eileen M. McDonald, MS Center for Injury Research and Policy,
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Lara B. McKenzie, PhD, MA Center for Injury Research and Policy,
The Research Institute at Nationwide Children's Hospital, The Ohio State University,
Columbus, OH, USA

Ted R. Miller, PhD Center for Public Health Improvement and Innovation,
Pacific Institute for Research and Evaluation, Calverton, MD, USA

Anthony N. Moore, BS Department of Neurobiology & Anatomy,
The University of Texas Medical School at Houston, Houston, TX, USA

Lynne Moore, PhD Département de Médecine Sociale et Préventive, Université Laval,
Québec City, QC, Canada

Centre Hospitalier Affilié Universitaire de Québec, Pavillon Enfant-Jésus, Quebec City, QC, Canada

Craig Newgard, MD, MPH Department of Emergency Medicine, Center for Policy
and Research in Emergency Medicine, Oregon Health and Science University, Portland, OR, USA

Lenora M. Olson, PhD Intermountain Injury Control Research Center, University of Utah
Department of Pediatrics, Salt Lake City, UT, USA

Michelle Patch, MSN, RN Department of Emergency Medicine, The Johns Hopkins Hospital,
Baltimore, MD, USA

Christopher L. Perdue, MD, MPH Armed Forces Health Surveillance Center,
Silver Spring, MD, USA

Julius Cuong Pham, MD, PhD Department of Emergency Medicine,
Johns Hopkins University School of Medicine, Baltimore, MD, USA

Department of Anesthesiology and Critical Care Medicine,
Johns Hopkins University School of Medicine, Baltimore, MD, USA

John B. Redell, PhD Department of Neurobiology & Anatomy, The University
of Texas Medical School at Houston, Houston, TX, USA

Leon S. Robertson, PhD Yale University, New Haven, CT, USA
Green Valley, AZ, USA

Damaris J. Rohsenow, PhD Center for Alcohol and Addiction Studies, Brown University,
Providence, RI, USA

Maria Seguí-Gómez, MD, ScD European Center for Injury Prevention, Facultad de Medicina,
Universidad de Navarra, Pamplona, Spain

Anbesaw W. Selassie, DrPH Department of Biostatistics, Bioinformatics and Epidemiology,
Medical University of South Carolina, Charleston, SC, USA

Sneha Shah, MD Department of Emergency Medicine, The Johns Hopkins Hospital,
Baltimore, MD, USA

Dennis F. Shanahan, MD, MPH Injury Analysis, LLC, Carlsbad, CA, USA

Baishakhi Banerjee Taylor, PhD Trinity College of Arts and Sciences, Duke University,
Durham, NC, USA

David J. Thurman, MD, MPH National Center for Chronic Disease Prevention
and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA

Margaret Warner, PhD Office of Analysis and Epidemiology, National Center for Health
Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

Steven Wiker, PhD, CPE Ergonomic Design Institute, Seattle, WA, USA

Flaura K. Winston, MD, PhD Department of Pediatrics, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA, USA

Center for Injury Research and Prevention, The Children's Hospital of Philadelphia,
Philadelphia, PA, USA

Shenjun Yao, PhD Department of Geography, The University of Hong Kong, Hong Kong, China

Jing Zhao, MD, PhD Department of Neurobiology & Anatomy, The University
of Texas Medical School at Houston, Houston, TX, USA

Part I
Injury Surveillance

Chapter 1

Surveillance of Injury Mortality

Margaret Warner and Li-Hui Chen

Introduction

Tracking injury mortality is fundamental to injury surveillance because death is both a severe and an easily measured outcome. Injury mortality has been monitored for a variety of purposes. For instance, the decline in motor vehicle crash death rates over time was used to document that improvement in motor vehicle safety was one of the ten greatest achievements in public health of the twentieth century (Centers for Disease Control and Prevention 1999).

However, mortality surveillance also has some limitations. As discussed in the chapter on injury morbidity, many injuries are nonfatal, and death is not necessarily a surrogate for the most serious injuries. Risk of death may be influenced by factors other than severity (e.g., comorbid conditions, distance to the hospital). In addition, some injuries, such as internal organ injuries, are very serious, but if survived, these injuries may not result in long-term limitations. Some injuries are less likely to result in death but may have very serious long-term outcomes (e.g., lower-leg fractures).

This chapter focuses on surveillance of fatal injuries using existing data systems, primarily from the United States of America (USA), although aspects of systems from some other countries are discussed. The chapter includes details for monitoring all injury deaths and subgroups of injury deaths. This includes surveillance needs by intent of injury (e.g., homicide), mechanism of injury (e.g., motor vehicle crash), nature of injury (e.g., hip fracture), activity when injured (e.g., occupational injuries), or place of injury (e.g., in the home).

The chapter describes data sources for injury mortality surveillance with a focus on vital statistics data, provides an overview of major classification systems for injury mortality, summarizes issues related to defining cases in injury mortality data systems, presents ways that injury mortality data are disseminated, provides methods to evaluate injury mortality surveillance systems, and concludes with a discussion of future directions for injury mortality surveillance.

M. Warner, PhD (✉)

Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Room 6424, 3311 Toledo Road, Hyattsville, MD 20782, USA
e-mail: mwarner@cdc.gov

L.-H. Chen, PhD

Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention, Room 6423, 3311 Toledo Road, Hyattsville, MD 20782, USA
e-mail: eyx5@cdc.gov

Data Sources

Vital records are the oldest and most commonly used source for injury mortality surveillance. Other sources which can supplement vital records or can serve as the primary source for countries that do not maintain vital records are presented in brief.

Vital Records

Vital records are the main source of mortality data for all causes in the USA, as well as in many other countries, and provide the most complete counts of deaths. Vital records generally include the cause or causes of death, and injury deaths can be selected from among these causes. Vital records also include demographic information about the decedent, and date and place of death. In the USA, vital records are collected by the States and then compiled into the National Vital Statistics System by the National Center for Health Statistics. A detailed description of the system can be found elsewhere (Xu et al. 2010). In many countries, including the USA, the source document for vital records is the death certificate.

A death certificate is a medicolegal form which includes demographic information on the decedent as well as the circumstances and causes of the death. The World Health Organization (WHO) has set guidelines for the cause-of-death section of the death certificate in an attempt to standardize the reporting of death (Anderson 2011).

In the USA, demographic information is completed by the funeral director as reported by the “best qualified person” who is usually a family member or friend (National Center for Health Statistics 2003a). Demographic information includes name, age, sex, race, and place of residence. The cause-of-death section of the death certificate must be completed by the attending physician, medical examiner, or coroner (National Center for Health Statistics 2003b).

The cause-of-death section of the US standard death certificate is shown in Fig. 1.1. The cause-of-death section is divided into two parts. In Part I of the death certificate, those responsible for certifying the cause of death are asked to provide a description of the chain of events leading to death, beginning with the condition most proximate to death (i.e., the immediate cause) and working backward to the underlying cause of death. In Part II, the certifier is asked to report other conditions that may have contributed to death but were not in the causal chain. For injuries, certifiers are prompted to describe how the injury occurred in “Box 43” and the place of injury in “Box 40.” The sequence of events leading to death as certified on the death certificate using Part I and Part II plays an important role in determining the underlying cause of death.

There is wide variation in the way that the cause-of-death portion of death certificates is completed in the USA, which is not surprising, given the range of experience of the certifiers completing this section of the death certificate. Although the written protocol suggests that the death certificate should include as much detail as possible, some certifiers provide more detail than others. For instance, in the case of a drug poisoning death, some certifiers provide little detail (e.g., drug intoxication); some certifiers provide more detail (e.g., methadone overdose), while others provide even more detailed information (e.g., decedent took methadone prescribed for pain relief and overdosed accidentally).

In the USA, death certificates must be filed within 3–5 days after a death in most states, with the cause of death supplied to the best of the certifier’s ability. However, if the certifier is unsure of the cause of death, the certificate will be marked as pending further investigation. In the USA, injury deaths account for a high proportion of pending certificates, including those for homicide, suicide, and poisoning (Minino et al. 2006).

In the USA, the information provided in the cause-of-death portion of the death certificate is coded according to the International Classification of Diseases (ICD) (see *Classification* section in

To Be Completed By: MEDICAL CERTIFIER	CAUSE OF DEATH (See instructions and examples)			Approximate interval: Onset to death
	<p>32. PART I. Enter the <u>chain of events</u> -- diseases, injuries, or complications --that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.</p> <p>IMMEDIATE CAUSE (Final disease or condition resulting in death)</p> <p>a. _____ Due to (or as a consequence of):</p> <p>b. _____ Due to (or as a consequence of):</p> <p>c. _____ Due to (or as a consequence of):</p> <p>d. _____</p> <p>Sequentially list conditions, if any, leading to the cause listed on line a. Enter the UNDERLYING CAUSE (disease or injury that initiated the events resulting in death) LAST</p>			<p>Part I Lines 1-4 Causes of death are entered sequentially starting with immediate cause and ending with the underlying cause.</p>
	<p>PART II. Enter other <u>significant conditions contributing to death</u> but not resulting in the underlying cause given in Part I.</p>			
	<p>35. DID TOBACCO USE CONTRIBUTE TO DEATH?</p> <p><input type="checkbox"/> Yes <input type="checkbox"/> Probably <input type="checkbox"/> No <input type="checkbox"/> Unknown</p>		<p>36. IF FEMALE:</p> <p><input type="checkbox"/> Not pregnant within past year <input type="checkbox"/> Pregnant at time of death <input type="checkbox"/> Not pregnant but pregnant within 42 days of death <input type="checkbox"/> Not pregnant but pregnant 43 days to 1 year before death <input type="checkbox"/> Unknown if pregnant within the past year</p>	
	<p>38. DATE OF INJURY (Mo/Day/Yr)(Spell Month)</p>		<p>39. TIME OF INJURY</p>	
<p>42. LOCATION OF INJURY: State: Street & Number:</p>		<p>40. PLACE OF INJURY (e.g., Decedent's home, construction site, restaurant, wooded area)</p>		<p>41. INJURY AT WORK? <input type="checkbox"/> Yes <input type="checkbox"/> No</p>
<p>43. DESCRIBE HOW INJURY OCCURRED</p>			<p>Box 43. How injury occurred Generally determines external cause of death.</p>	

Fig. 1.1 The cause-of-death section of the death certificate (US Standard Certificate of Death – Rev 11/2003 available at <http://www.cdc.gov/nchs/data/dvs/death11-03final-acc.pdf>)

the chapter for details on ICD) using an automated coding system with some records still coded by hand. In order to accommodate an automated coding system, the text as written on the death certificate is transcribed into an electronic format, in the case of paper certificates, or retained, in the case of electronic certification.

Data from Coroners and Medical Examiners

In the USA, a death certificate for injury and other sudden and unnatural deaths must be certified by a coroner or medical examiner (ME) and typically requires further investigation into the cause of death. The investigation into the cause of death may include both a medical and legal component. The medical component focuses on the cause of death, while the legal focuses on whether the death was unintentional, self-inflicted, or inflicted by another person.

The distinction between medical examiners and coroners is linked to the tasks required for death investigation. Medical examiners are board-certified physicians with training specifically in medical investigation and are appointed to their posts. Coroners traditionally were involved in the legal aspects of the investigation, did not have a medical background, and were often elected officials. However, more recently, some coroners have medical degrees and have also been appointed to their posts (Hickman et al. 2007).

The medical component of death investigations includes reviewing the medical history of the deceased and may include an autopsy. For all causes of death, autopsy rates are decreasing in the USA; the rate was 7.7% in 2003 (Hoyert et al. 2007). For injuries, autopsy rates vary by intent, cause, and type of injury. For instance, in 2003, while over 90% of homicides and over 75% of external cause of deaths with an undetermined intent were autopsied, only 52% of suicides and less than half (44%) of unintentional injuries were autopsied.

For drug-related deaths, an important component of the medical investigation is the toxicological tests employed to determine the types of drugs involved. The tests and the substances tested may vary from case to case as well as among jurisdictions and over time. Testing for specific drugs is conducted after the drug has been identified as a problem and only if the test is not cost prohibitive. For instance, because it was not included in the standard drug screening tests and it was very expensive, testing for fentanyl was not routine until around 2002.

In the USA, medical examiners and coroners do not have a standard format for recording death investigation data. Some states and jurisdictions have created their own format and store the information electronically for research purposes. For instance, states participating in the National Violent Death Reporting System (NVDRS) (Weiss et al. 2006; Paulozzi et al. 2004), which is described later in the chapter, must report information to this system in a standard form and have created systems to store the data electronically. In addition, some offices have created systems for reporting and disseminating information on specific causes of injury deaths in their state. For example, Florida releases a report annually on the drugs involved in drug-related deaths.

In Australia, coroners perform all death investigations, and the coroners reports are used to compile details about every death reported to coroners in a national system referred to as the National Coroners Information System (NCIS) (Driscoll et al. 2003). To supplement the information from the coroners' reports, police reports, autopsy reports, and toxicology reports are used to gather further details on the causes and circumstances of the death. The full reports from the coroners and the other source documents are available with restricted access. The system was designed not only as a surveillance and injury prevention tool but also as a resource for coroners to monitor the consistency of death investigations. It has proved useful for injury prevention and control, as well as other purposes (Driscoll et al. 2003; [National Coroners Information System](#)).

Systems Based on Multiple Data Sources

Surveillance systems can be developed with information from more than one source. Death certificates or vital records often serve as the primary source for these systems. These records are supplemented with needed details from other sources. However, data from different sources may not agree and, thus, present some challenges for analysis (Karch and Logan 2008). In the USA, examples of databases which capture information on fatal injury from many sources include Fatal Analysis Reporting System (FARS), Census of Fatal Occupational Injury (CFOI), and NVDRS.

FARS is produced by the National Highway Traffic Safety Administration and tracks deaths from fatal car crashes in the USA ([National Highway Traffic Safety Administration](#)). Source documents include vital statistics, reports from the police, the state highway department, the coroner/medical examiner, the hospital, and the emergency medical service, as well as the state vehicle registration files and driver-licensing files.

CFOI is produced by the US Department of Labor and tracks all occupational injury fatalities in the USA (Bureau of Labor Statistics 2007). Source documents include death certificates, news accounts, workers' compensation reports, and Federal and State agency administrative records.

NVDRS is produced by the Centers for Disease Control and Prevention and tracks homicides, suicides, deaths by legal intervention, and deaths of undetermined intent, as well as unintentional firearm injury deaths in 17 states in the USA (Weiss et al. 2006; Paulozzi et al. 2004). Source documents include records from law enforcement, coroners and medical examiners, and crime laboratories.

Supplementary Data Sources

Newspapers and other news sources have been used to collect data on specific causes of injury death both in the USA and around the world. In the past decade, the number of online news media has increased, and the capability to search for news reports has improved. These improvements may eliminate some of the barriers to using news as a data source for injury surveillance. Even prior to these improvements, studies have found news reports to be a useful tool for injury surveillance (Rainey and Runyan 1992; Barss et al. 2009; Rosales and Stallones 2008; Genovesi et al. 2010). One study found that newspapers covered more than 90% of fire fatalities and over three quarters of the drownings in North Carolina (Rainey and Runyan 1992). The researchers found that the newspaper included more information than medical examiner records on several factors, including the cause of the fire, the presence of smoke detectors, pool fences, warning signs, and supervision of children. A study of drowning in the United Arab Emirates found that newspaper clippings were able to provide more information about drowning than ministry reports (Barss et al. 2009). However, relying solely on newspaper reports may give an incomplete (Rosales and Stallones 2008) and even misleading picture (Genovesi et al. 2010) because news media tend to include unusual stories rather than the usual causes of death.

Police reports can also be useful for capturing information about events leading up to the death (Logan et al. 2009). In the USA, FARS is based in part on police reports because of the information gleaned on the circumstances of the crash. In developing countries, where little or no data on injury deaths exist, police reports may provide some data (Rahman et al. 2000; Bhalla et al. 2009). However, limitations of police reports include inconsistent reporting (Agran et al. 1990).

Modeling and surveys can be used to estimate death rates for countries or regions of the world that do not have the resources or political power to compile censuses of fatalities (Hill et al. 2007). For example, the Global Burden of Diseases modeled injury death rates for many countries in its World Report (Mathers et al. 2008). Models use data from many sources, and the quality of the estimates varies by the reliability of the sources. Many techniques are being developed to make the models more robust (Hill et al. 2007; Mathers et al. 2008; Patton et al. 2009; Lawoyin et al. 2004; Sanghavi et al. 2009; Fottrell and Byass 2010). In some countries, only the fact of death is known, not the cause. When this is the case, methods to estimate cause, based on interviews with lay respondents on the signs and symptoms experienced by the deceased before death, referred to as a verbal autopsy, have been developed (Lawoyin et al. 2004; Fottrell and Byass 2010; Baiden et al. 2007). Results of verbal autopsies can be used to estimate the portion of deaths due to specific causes and are used to supplement models. Modeling and verbal autopsies were used to estimate the magnitude of burn injuries in India (Sanghavi et al. 2009).

Classification of Injury Deaths

Mortality data in surveillance systems are stored and retrieved using classification systems that can be used to identify deaths from injuries or specific types of injuries (Fingerhut and McLoughlin 2001). Injury deaths in mortality surveillance systems are usually classified according to cause of the

injury, including any objects, substances, or person involved; intent of injury; and physical trauma to the body. In addition, place of injury and the activity engaged in at the time of injury are often included in data systems. These, along with an identifier, age, and sex, are considered the core minimum data set for injury surveillance (Holder et al. 2001).

The cause of the injury describes the mode of transmission of external energy to the body. Knowing how the energy is transmitted can lead to prevention of the event leading to injury – primary prevention. The intent of the injury (sometimes referred to as manner) is also important as some interventions may vary depending on the intent, particularly interventions that are not strictly passive and require a behavioral component.

The body regions involved and nature of injury can assist with developing both secondary and tertiary prevention programs. For instance, knowing that the fatalities in many crashes were the result of crushing chest injuries from the steering wheel led to the development and implementation of air bags.

Place and activity at the time of injury provide more information about the environment in which the injury occurred. When used together with the external cause, they provide information that can be used to help inform prevention strategies.

International Classification of Diseases

The ICD is the most widely used classification system for all deaths of all causes worldwide (World Health Organization 2004). The WHO maintains the ICD in order to provide a common language for health conditions. In the USA, causes of death have been classified using the tenth revision of ICD since 1999 and using the ninth revision from 1979 to 1998. Since the first version of ICD, injuries have been separately identified using the classification system. Since ICD-6, injuries have been described in two ways: either (1) by “external cause of injury” which describes the cause and intent in a single code or (2) by the “nature of injury” which describes the body region and nature of injury in a single code.

The International Classification of External Causes of Injury (ICECI), which is also maintained by WHO and is compatible with the ICD, is a more detailed classification system designed specifically for injury (WHO Family of International Classifications). Although ICECI is not used in the USA, the system has many advantages for classifying injury deaths. ICECI has a multiaxial and hierarchical structure with a core module, including mechanism of injury, objects/substances producing injury, place of occurrence, activity when injured, the role of human intent, and the use of alcohol and psychoactive drugs. There are additional modules for classifying data on violence, transport, place, sports, and occupational injury. The 11th revision of ICD will be in part based on the ICECI.

External Cause of Injury

The external cause of injury describes the vector that transfers the energy to the body (e.g., fall, motor vehicle traffic accident, or poisoning) and the intent of the injury (e.g., unintentional, homicide/assault, suicide/self-harm, or undetermined). External cause codes are often referred to as E-codes. The terms cause and mechanism of injury are often used interchangeably, and intent and manner of death are used interchangeably, although there are slight differences in meaning depending on the discipline (e.g., medical examiner, epidemiologist). In ICD-10, external-cause-of-injury codes are in Chapter 20 and begin with the letter *U, V, X, W, and Y. In ICD-9, the external-cause-of-injury

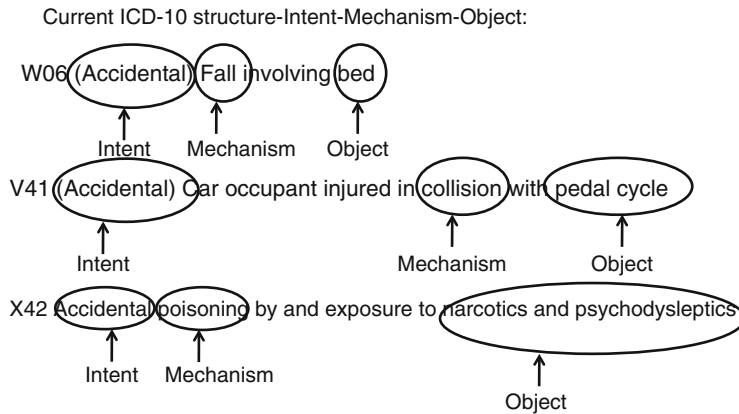


Fig. 1.2 External cause codes code structure shown with fall, motor vehicle, and poisoning ICD-10 codes

codes are included in the Supplemental Classification of External Causes of Injury and Poisoning and begin with the letter E.

External cause codes classify many dimensions of the cause of injury in a single code. Most external cause codes include at least the dimension of intent and cause, with intent as the primary axis and cause as the secondary axis. In addition, the external cause code often specifies the objects or substances involved. Figure 1.2 provides an example of external cause codes for falls, motor vehicle crashes, and poisoning.

The ICD-coding guidelines include a method to select an underlying cause of death which is used for many analyses (World Health Organization 2004; National Center for Health Statistics 2010a). The underlying cause of death in ICD is *the disease or injury that initiated the chain of events leading directly to death, or the circumstances of the accident or violence which produced the fatal injury*. For injury deaths, the underlying cause of death is always the external cause, in recognition that it is closest to the agent of injury; and the nature of injury (e.g., traumatic concussion) is included in the multiple causes of death. When more than one cause is involved in the death, the underlying cause is determined by (1) the sequence of conditions on the death certificate, (2) rules and guidelines of the ICD, and (3) associated ICD classification rules.

Nature of Injury

The nature-of-injury codes describe the body region that was injured (e.g., head) and the nature of injury (e.g., fracture and laceration). These codes are sometimes referred to as the diagnosis codes. In the USA, the diagnosis, however, in most cases is gleaned from the death certificate and may be little more than a lay description of the injuries. In ICD-10, the nature-of-injury codes are included in Chap. 19 and begin with the letter S or T. In ICD-9, the nature-of-injury codes are included in a Chap. 17 and are designated by codes 800–999. The primary axis for nature of injury is the body region in ICD-10 and was the type of injury (e.g., fracture and laceration) in ICD-9.

Nature-of-injury codes cannot be the underlying cause of death in ICD and are always included in the multiple causes of death. In the USA, up to 20 causes are recorded in the vital statistics data. Both ICD-10 and ICD-9 have rules to select a main injury from among the multiple causes. However, the methods suggested for both revisions are under debate.

Matrices Used to Present ICD-Coded Data

The ICD injury matrices are frameworks designed to organize ICD-coded injury data into meaningful groupings and were developed specifically to facilitate national and international comparability in the presentation of injury statistics (Minino et al. 2006; Bergen et al. 2008; Centers for Disease Control and Prevention 1997; Fingerhut and Warner 2006). The external-cause-of-injury matrix is a two-dimensional array which presents both the cause and intent of the injury using the ICD codes. The injury mortality diagnosis matrix for ICD-10 codes is a two-dimensional array describing both the body region and nature of the injury.

The matrices cross classify the codes so that it is easier to examine deaths using the secondary axis. For instance, using the external-cause-of-injury matrix, one can quickly identify all deaths by drowning, regardless of intent. Since the burden of proof for intent may vary by jurisdiction or over time, this ability to conduct surveillance on causes, regardless of intent, may be important for unbiased research.

Place and Activity at Time of Injury

Information about the place of occurrence and the activity engaged in at the time of injury is useful for prevention. In addition, some injury researchers focus only on certain activities (e.g., occupational) or places (e.g., schools). ICD has a limited classification scheme for place and activity; ICECI, however, has more detailed place of injury occurrence classification. In ICECI, the codes are designed to reflect the place of occurrence and the activity engaged in, as well as “an area of responsibility” for prevention. For instance, ICECI activity codes include “paid work,” and place codes include “school, education area” because for injuries at work and in school, prevention efforts may be a shared responsibility.

Activity, which describes what the injured person was doing when the injury occurred, may be difficult to classify because a person may be engaged in more than one type of activity. For example, a bus driver may be injured while engaged in paid work driving a bus. ICECI has established precedence rules for selecting a primary and a secondary activity.

Place of injury, which describes where the person was when he or she was injured, may be easier to classify than activity or cause of injury. ICD includes broad categories for use when all causes of injury are under study and ICECI has more detailed place categories. However, for specific causes such as drowning, more detailed place classification schemes have been suggested (Brenner et al. 2001). For instance, knowing whether the drowning occurred in a pool, pond, or bucket will inform prevention.

Issues to Consider in Operationally Defining Injury Deaths

There are many issues to consider in operationally defining injury deaths that meet the purposes of surveillance within the context of an existing data system. Surveillance can be used to monitor all injury deaths or for subgroups by intent of injury (e.g., homicide), mechanism of injury (e.g., motor vehicle crash), and nature of injury [e.g., Traumatic Brain Injury (TBI) or hip fractures] during a specified activity (e.g., occupational injuries) or in a specified place (e.g., in the home). When using an existing surveillance system, injury deaths of interest for a specific surveillance objective may need to be selected from other deaths. The existing surveillance system may be limited in its ability to address the specific surveillance objective exactly, and an operational definition based on the existing system needs to be developed. Consideration should be given to how the operational defini-

tion and the definition of interest differ and, ultimately, the effect on the estimates produced by the surveillance system. This section will discuss some issues to consider in defining injury deaths and subgroups of injury deaths including using the ICD, using data on multiple causes of death, and considering deaths that do not occur immediately after the injury.

Operational Definitions of Injury Using the ICD

If the data are ICD coded, as is the case for vital records in the USA, the external-cause-of-injury matrix is often used to define groupings of injuries by major categories of causes or intents. Definitions of injury deaths based on the ICD matrix exclude deaths from complications of surgical and medical care. These deaths are often excluded from injury deaths because they are seen as out of the purview of traditional injury prevention and control. The issue has been debated in the literature (Langley 2004). In addition, injuries resulting from minor assaults on the body over long periods of time are not included in the injury chapter of the ICD. For instance, deaths caused by chronic exposure to drugs or alcohol that results in liver disease are not included in the external cause chapter of the ICD.

Unit of Analysis

The unit of analysis for most analyses involving mortality data is deaths. For statistical analyses, the deaths are assumed to be independent events. However, for injuries, the deaths are not always independent of one another and may even be correlated (e.g., motor vehicle passengers). If the deaths are correlated, then the unit of analysis should be at the event level (e.g., car and plane crashes, house fires), or special statistical techniques that take into account the correlation are required. Both FARS and the NVDRS allow analyses at both the decedent and the event level.

Underlying vs. Multiple Causes

Injury deaths are multifaceted, and there may be more than one cause and more than one comorbid condition involved, as mentioned in the section describing the ICD. For some purposes, such as ranking causes of death, it is important to define mutually exclusive causes of death. Official rankings of all causes of death in the USA are based on a mutually exclusive list of causes defined using the underlying cause of death. For other purposes, a broad net is cast for a particular cause, and data on multiple causes of death should be used in the analysis. For example, an analysis with the goal of tracking all drowning-related deaths would have a broader definition of drowning than an analysis designed to rank the leading causes of injury, and therefore would require the use of multiple causes of death (Smith and Langley 1998). There may be large differences in numbers of deaths identified for a specific injury cause when using the underlying cause of death compared to the multiple causes of death (Kresfeld and Harrison 2007; Redelings et al. 2007).

When more than one injury is listed as contributing to the death, there are many methods to define injury deaths using the data on multiple causes of injury (Minino et al. 2006; Bergen et al. 2008; Fingerhut and Warner 2006; Aharonson-Daniel et al. 2003). Five methods are briefly described here. Selecting a main injury to analyze is a method that may be the easiest to explain. However, currently, there is no consensus on a method to select the main injury. Another common method is selecting deaths with a particular diagnosis mentioned at least once, sometimes referred to as “any mention.” This method is often used when analyzing a particular type of injury (e.g., TBI deaths). Another

method is to use the injury diagnoses as the unit of analysis, and all injury diagnoses mentioned are counted once; this is sometimes referred to as “total mentions.” With this strategy, each diagnosis mentioned is given equal weight in the analysis. A fourth method, sometimes referred to as weighted total mentions, assigns each injury diagnosis recorded for a death equal weight within a death so that each death is counted equally. For example, if a death includes mention of a superficial injury and a TBI, each is given a weight of $\frac{1}{2}$. A fifth method, referred to as multiple injury profiles, uses the injury diagnosis matrix to show combinations or profiles of injuries involved in deaths. Chapter 13 of this book is devoted to multiple injury profiles.

Late Deaths

Definitions of injury mortality should include some consideration of deaths which do not occur immediately after the traumatic event, referred to here as late deaths. Research has shown that the period of time between an injury and death may be years after the injury (Mann et al. 2005; Cameron et al. 2005; Probst et al. 2009). Injuries resulting in death after discharge from the hospital are of particular interest to those in the trauma field. Research has shown that injury is less likely to be included as a cause of death on the death certificate as the time between injury and deaths increases. For instance, fatality from hip fracture is suspected to be underreported because deaths can occur several days to weeks after the fracture (Cameron et al. 2005). One study found that even when the period between injury and death is as short as 3 days, the injury information was not recorded on the death certificate (Langlois et al. 1995).

The goal of surveillance should be considered when determining whether late injury deaths are defined as injury-related or whether they should be attributed to another cause of death. The guidance on this decision is limited (Cryer et al. 2010). For instance, in FARS, the operational definition used is that the death must have occurred within 30 days of the crash. Practically, data systems vary in the ability to identify late deaths from injury. In US vital statistics data, there is no time limit on reporting the cause of the death as injury-related as long as an injury cause is written on the death certificate. However, the cause may be ICD coded using a sequela code rather than an external cause code if the death occurred more than 1 year after the injury, or if the words “healed” or “history of” were mentioned (National Center for Health Statistics 2010a).

Dissemination

Dissemination is integral to a surveillance system because a goal of surveillance is to inform stakeholders, such as policymakers and those who design prevention programs, of changes in trends and emerging issues. This section includes a brief description of injury indicators used to disseminate injury mortality data, followed by a discussion of analytic issues, and standard publications and web-based dissemination.

Injury Indicators

An injury indicator describes a health outcome or a factor known to be associated with an injury among a specified population (Davies et al. 2001). Injury deaths are often used as indicators to monitor the general health of a population and to monitor injury occurrence. They can also be used for disseminating data from surveillance systems. Since indicators are a tool for measuring progress in

health outcomes, a good injury indicator should be free of bias and reflect variation and trends in injuries or injury-related phenomena. Injury indicators are being developed for international comparisons of injury statistics (Cryer et al. 2005). More information on injury indicators is available in the chapter on the surveillance of injury morbidity.

Analytic Issues

There are several analytic issues to consider in disseminating injury mortality data. This section describes variation and reliability of mortality data and common statistics and methods for disseminating mortality data, including death rates and ranking.

Variation and Reliability

Even though most vital statistics data are complete or near-complete enumerations of deaths and are not subject to sampling variation, there may be variation in the number of deaths that occurs randomly over time. If the number of deaths for specific causes of injury death is small or if the population at risk is small, the reliability of injury statistics generated from mortality data should be considered. Detailed methods for estimating variance for mortality statistics can be found elsewhere (Xu et al. 2010).

For US vital statistics mortality data, the National Center for Health Statistics recommends that in analyses of groups with less than 100 deaths, variation should be considered, and in analysis of groups with less than 20 deaths, rates should be considered unreliable. This is based on the assumption that the underlying distribution in the number of deaths follows a Poisson or negative binomial distribution (Brillinger 1986).

Rates and Population Estimates

Rates are a common measure of the risk of death. Typically, for rate calculations, the population data are from the Census Bureau. The decennial census of the population has been held in the USA every 10 years since 1790. It has enumerated the resident population as of April 1 of the census year since 1930. Postcensal population estimates are estimates made for the years following a census, before the next census has been taken. The Census Bureau produces a postcensal series of estimates of the July 1 resident population of the USA annually. Each annual series of postcensal estimates is referred to as a vintage. Population estimates for the same year differ by vintage (National Center for Health Statistics 2010b). For example, population estimates for 2002 of Vintage 2003 differ from estimates for 2002 of Vintage 2004. Analysts who wish to benchmark their rates estimates from the National Vital Statistics System need to consider the populations used to calculate death rates. Death rates in standard reports from the NVSS are calculated with population estimates as of July 1 of the year of the death data. For example, 2007 death rates are calculated using population estimates as of July 1, 2007, Vintage 2007 (Xu et al. 2010).

Ranking Causes of Death

Ranking causes of death by the numbers of deaths in a set of mutually exclusive cause groups is often used in the dissemination of mortality data since it conveys the relative importance of specific causes of death by providing a method to compare the relative burden of each cause. Injuries are

included in the ranking of causes of death for most official national statistics; however, official rankings are usually by intent and, in some cases, by cause within intent. For instance, in the USA in 2007, suicide was ranked as the eighth leading cause of death, and homicide, the ninth among all causes of death (Xu et al. 2010). Among injury deaths, causes of injury death have been ranked using the standard groupings in the ICD external-cause-of-injury matrix and focusing on the mechanism of injury rather than the intent of injury (Minino et al. 2006; Anderson et al. 2004).

Measures of premature mortality such as years of potential life lost (YPLL) can provide a summary measure of the burden of injury (Segui-Gomez and MacKenzie 2003). Years of life lost for each decedent is estimated as age at death minus a set age (e.g., 75). YPLL is derived by summing years of life lost for decedents of all ages less than the set age selected (e.g., 75). The statistic highlights the fact that injuries disproportionately affect the young. YPLL can also be used to show trends in injury mortality.

Standard Publications and Web-Based Dissemination

Mortality data are disseminated using both standard publications and the Internet. In the USA, publications of mortality data were available as early as 1890 (Department of the Interior Census Office 1896). For many years, mortality data were disseminated in a bound volume referred to as the Vital Statistics of the USA. These volumes are useful resources for historical injury mortality statistics in the USA and are available at many state and university libraries and on the web. Since 1997, tabulated statistics for deaths have been published annually in National Vital Statistics Reports (<http://www.cdc.gov/nchs/products/nvsr.htm#vol53>).

Data may be disseminated on the web both as statistical reports, such as those described above, and as interactive query systems designed to tabulate the data as needed. The online interactive query system WONDER includes interactive methods to analyze mortality data by underlying causes and by multiple causes of death (Centers for Disease Control and Prevention). WONDER uses the external-cause-of-injury matrix to categorize injuries by injury intent and mechanism.

The web-based injury statistics query and reporting system, WISQARS, includes several interactive modules for reporting fatal injury statistics in the USA, including Injury Mortality Reports, Leading Causes of Death Reports, YPLL Reports, Fatal Injury Mapping, and Cost of Injury Reports (National Center for Injury Prevention and Control). In WISQARS, the injury cause categories are based on the external-cause-of-injury matrix, and the nature-of-injury categories in the cost of injury module are based on the injury mortality diagnosis matrix.

In the USA, the multiple-cause-of-death microdata files from the National Vital Statistics System are available for downloading starting with the year 1968 on the NCHS vital statistics web site (National Center for Health Statistics). The NCHS injury data and resource web site have more information and tools for analyzing injury mortality data in the USA (<http://www.cdc.gov/nchs/injury.htm>).

Surveillance Systems Evaluation and Enhancements

Surveillance systems can be improved by periodically evaluating the systems. General criteria on which to evaluate injury surveillance systems have been developed (Mitchell et al. 2009; Macarthur and Pless 1999) and are described in Chapter 2. This section includes a brief description of the quality of the information on the death certificate and in the vital statistics data. In addition, possible methods to enhance the systems with supplemental data are discussed.

Quality of Vital Statistics Mortality Data

Vital statistics mortality data have many recognized strengths and limitations (Committee for the Workshop on the Medicolegal Death Investigation System 2003). A major strength for the surveillance of injury deaths is that it includes all deaths in the USA over a long time period. Other strengths are the standardization of format, content, and coding of the data.

The universal coverage allows for surveillance of all causes of mortality and the ability to make statistical inferences regarding trends and subgroup differences for relatively uncommon causes of death and small geographic areas and population groups. Standard forms for the collection of the data and model procedures for the uniform registration of the events have been developed and recommended for nationwide use. Material is available to assist persons in completing the death certificate. Software is available to automate coding of medical information on the death certificate, following WHO rules specified in the ICD. The ICD has codes for classifying all types of diseases and injuries and is used around the world and updated regularly.

Limitations of the vital statistics data for injury mortality surveillance include lack of detail on some death certificates resulting in nonspecific causes of injury death codes; lack of standardization in determining intent; and improper certification of the sequence of events leading to the death.

The quality of the vital statistics death data is limited by the quality of the certification of death. Injury deaths may be certified with little detail on the external cause of death and nature of injury on the death certificate. For instance, some death certificates may be filled out with little more than “MVA” or “drug intoxication” as the external cause and “multiple injuries” as the description of the nature of injury. This lack of detail on the death certificate leads to nonspecific cause-of-death codes which are not useful for injury mortality surveillance or for injury prevention and control (Breiding and Wiersema 2006; Lu et al. 2007; Romano and McLoughlin 1992). For example, a study found that TBI may be underestimated in Oklahoma by as much as 25% because the injury description on the death certificates did not provide detail needed to identify brain injuries (Rodriguez et al. 2006).

Measuring the proportion of nonspecific and ill-defined causes is a method to assess the quality of vital statistics death data (Bhalla et al. 2010). For injury data, the focus is on the proportions of unknown and ill-defined causes of injury. Using this method, at least 20 countries were identified as having data of high enough quality that they can be used to monitor trends in death from injury. For other countries, the number of deaths with an imprecise or partially specified cause of death or cause of injury was so high that the distribution of deaths with causes enumerated would be of questionable accuracy.

Methods of evaluating intent may differ among certifiers, leading to inconsistencies in the vital statistics data (Breiding and Wiersema 2006). For example, in determining whether the cause of death was self-inflicted, one certifier might conclude that a mildly depressed person who went for an early morning swim in the ocean was intending to commit suicide, whereas another may require more conclusive proof such as a suicide note and, in its absence, certify the death as undetermined. Figure 1.3 shows an example of the variation in the reported intent for poisoning deaths between different states and years. For states with a state ME’s office, there is usually a standardized approach, or at least a standard philosophy for assessing intent. For states without a centralized office, there may be more variation in methods of determining intent within a state. In other countries, the intent is a medicolegal decision. For instance, in England and Wales, after the coroner completes the inquest, a final ruling on injury intent is made.

The certifier’s description of the sequence of events or conditions leading to death as reported on the death certificate is a key factor in determining the underlying cause of death and may result in injuries being omitted from death certificates or included as a contributing cause rather than the underlying cause. For injuries, the improper sequencing of events is more likely to occur if the death is not immediate and other health conditions related to the injury contribute to the death. For example, with hip fractures or spinal cord injuries, other health conditions may contribute to death, but

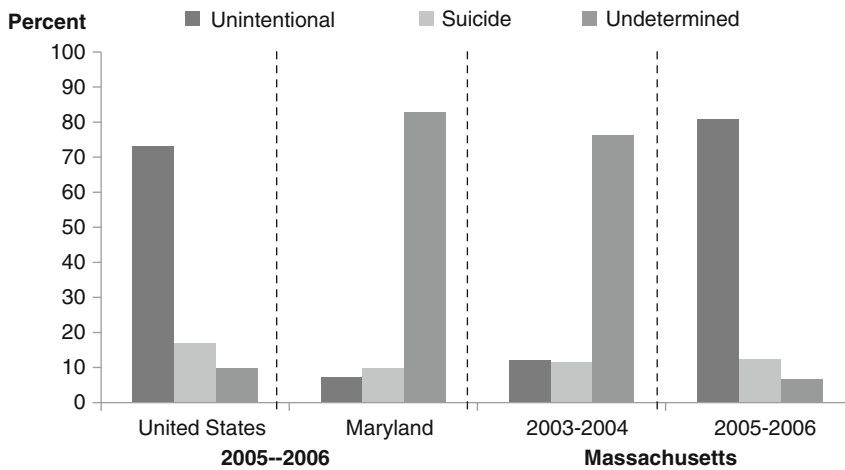


Fig. 1.3 Percent distribution of poisoning deaths by intent

without the hip fracture or the spinal cord injury, the death would not have occurred. If the certifier incorrectly lists the external cause in Part II of the death certificate, it may be included as a contributing factor and not the underlying cause of death. In contrast, if the external cause had been properly listed at the end of the sequence in Part I, it will be selected as the underlying cause of death.

Despite these limitations, vital statistics data have great value for injury mortality surveillance due to their universal coverage over a long time period, and standardization of content, format, and cause-of-death classification.

Supplements to Surveillance Systems

Common methods for supplementing routinely available injury mortality surveillance data are described in this section, including retaining source data used for coding (e.g., death certificates), linking vital statistics data with other sources, and conducting follow-back surveys.

Retaining the Source Data

The source data used to classify the deaths in mortality surveillance systems are sometimes retained and incorporated into the system. By allowing access to the source data, the free-form data that were used to classify deaths can be further mined for details. For example, NCIS, the coroner's data system in Australia, includes copies of many of the reports used to classify deaths. With prior approval, researchers may be able to review these reports for details that may have been lost during the classification of deaths.

The source data for the coded causes of death in vital statistics are the narrative text written in the cause-of-death section. These narratives have been used for surveillance purposes to provide details that can supplement the coded data. For instance, the location of drowning has been further evaluated using a review of death certificates (Brenner et al. 2001). However, analyzing these data has traditionally required a manual review of death certificates. More recently, with the advent of automated coding software, the narrative text is routinely transcribed from paper death certificate or entered directly on electronic death certificates for use as the input data to code the multiple-cause-of-death data. The electronic form of the narrative text data, sometimes referred to in the USA as the

“literal text,” has been used for surveillance purposes since 2003. In the USA, these data have been analyzed to describe deaths involving motor vehicles that were not on the road (Austin 2009). In England and Wales, and New Zealand, a field for additional notes on the death is available. This field has been used to help identify drugs involved in deaths (Flanagan and Rooney 2002).

Data Linkage

In the USA, national surveys of health, such as the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES), are routinely linked with mortality data using the National Death Index (NDI). The NDI is an indexing system for locating death records for cohorts under study by epidemiologists and other health and medical investigators (National Center for Health Statistics). Multiple cause-of-death data, including injuries, are available through linked data files.

Both the NHIS and NHANES are data sources rich in information about the health and demographic characteristics of the persons surveyed. Using the survey data linked with mortality data, it is possible to study risk factors for injury death using data previously collected. For instance, using linked data researchers compared the risk of suicide among veterans to the risk among the general population (Kaplan et al. 2007). Socioeconomic factors and neighborhood factors related to injury mortality have also been studied using survey data linked with mortality data (Cubbin et al. 2000a, b).

Follow-Back Surveys

Mortality follow-back surveys have been used to supplement vital statistics data with information supplied by the next of kin or another person familiar with the decedent. Unlike health survey data linked with mortality data, which provide baseline information for the decedents and others in the population, follow-back surveys provide additional information about the circumstance of the injury that led to the person's deaths. For example, follow-back surveys have been used to find information on alcohol use (Sorock et al. 2006; Baker et al. 2002; Li et al. 1994; Chen et al. 2005) and firearm and other violent deaths (Kung et al. 2003; Conner et al. 2001; Schneider and Shenassa 2008; Wiebe 2003; Dahlberg et al. 2004).

Future Directions

Timely Vital Data for Surveillance of Emerging Threats

A major purpose of vital statistics data in the USA is statistical reporting, and an emphasis has been placed on accuracy, at the expense of timeliness. To ensure accuracy, the data are released after all death certificate queries have been returned and quality reviews and consistency checks have been completed. Injury deaths, and in particular poisoning deaths, are some of the last certificates to be resolved, and therefore, for high-quality injury data, waiting is necessary. However, in the USA, the National Vital Statistics System is being reengineered to improve the speed with which the data are processed and released. Additional improvements to the system include the ability to monitor the literal text from death certificates even before the data are processed, and these data may be useful for the surveillance of emerging threats.

Quality of Vital Statistics Data

Electronic death registration is expected to improve both the quality and timeliness of the data. Electronic death registration allows for help screens, explanation of the death certificate, and automating queries on ill-defined causes of death. In addition, it speeds the process of certifying deaths. In the USA, death registration is a state responsibility, but there is a federal effort to assist in the transition from paper to electronic medical records. In 2007, 15 states were registering some deaths electronically with many more states in various stages of transition. The National Association for Public Health Statistics and Information Systems web site includes more details on electronic death registration (see <http://www.naphsis.org>).

Narrative Text

Narrative text has been used successfully in injury surveillance for many years (McKenzie et al. 2010a). In the future, the increased storage capacity of most data systems allows for retention of narrative text as well as source documents used in data collection. Methods to analyze the text and source documents are improving, and computer software packages specifically designed for text analysis are available (McKenzie et al. 2010b). Incoming narrative text on causes of death that have not yet been classified could help to detect emerging mortality threats. The retention of the text combined with improvements in the ability to rapidly abstract data from these nonstandard sources may lead to improvements in surveillance.

Linking Surveillance Data

Linking data sources capitalizes on existing resources, and mortality data are often a component of linkages. In the last decade, methods of linking data sources and the software to perform data linkage have advanced, making linking sources easier and more accurate. Surveillance which utilizes linked morbidity and mortality data creates a more complete picture of the burden of injury. Linked survey data and mortality data will provide better understanding of the injury mortality risk factors.

Advances in Injury Mortality Surveillance in Less Resourced Environments

Internationally there are ongoing efforts to increase the reliability of death registration systems (AbouZahr et al. 2007; Setel et al. 2007). In addition, the Global Burden of Diseases project is ranking causes of death globally. One outcome of this effort has been to show through statistics the relative importance of injuries, and road traffic accidents in particular. This systematic review of international mortality surveillance data facilitates improving the quality of mortality data for all.

References

- AbouZahr, C., Cleland, J., Coullare, F., et al. (2007). Who counts? 4 – the way forward. *The Lancet*, 370(9601), 1791–1799.
- Agran, P. F., Castillo, D. N., & Winn, D. G. (1990). Limitations of data compiled from police reports on pediatric pedestrian and bicycle motor vehicle events. *Accident Analysis and Prevention*, 22(4), 361–370.
- Aharonson-Daniel, L., Boyko, V., Ziv, A., Avitzour, M., & Peleg, K. (2003). A new approach to the analysis of multiple injuries using data from a national trauma registry. *Injury Prevention*, 9(2), 156–162.

- Anderson, R. N. (2011). Adult mortality. In R. Rogers & E. Crimmins (Eds.), *International handbook of adult mortality*. New York, NY: Springer Science.
- Anderson, R. N., Minino, A. M., Fingerhut, L. A., Warner, M., & Heinen, M. A. (2004). Deaths: Injuries, 2001. *National Vital Statistics Reports*, 52(21), 1–86.
- Austin, R. (2009). Not-in-traffic surveillance 2007 – children. In *A brief statistical summary*. Washington, DC: National Highway Traffic Safety Administration
- Baiden, F., Bawah, A., Biai, S., et al. (2007). Setting international standards for verbal autopsy. *Bulletin of the World Health Organization*, 85(8), 570–571.
- Baker, S. P., Braver, E. R., Chen, L. H., Li, G., & Williams, A. F. (2002). Drinking histories of fatally injured drivers. *Injury Prevention*, 8(3), 221–226.
- Barss, P., Subait, O. M., Ali, M. H., & Grivna, M. (2009). Drowning in a high-income developing country in the Middle East: Newspapers as an essential resource for injury surveillance. *Journal of Science and Medicine in Sport*, 12(1), 164–170.
- Bergen, G., Chen, L., Warner, M., & Fingerhut, L. (2008). *Injury in the United States: 2007 Chartbook*. Hyattsville, MD: National Center for Health Statistics.
- Bhalla, K., Shahraz, S., Bartels, D., & Abraham, J. (2009). Methods for developing country level estimates of the incidence of deaths and non-fatal injuries from road traffic crashes. *International Journal of Injury Control and Safety Promotion*, 16(4), 239–248.
- Bhalla, K., Harrison, J. E., & LA Saeid, F. (2010). Availability and quality of cause-of-death data for estimating the global burden of injuries. *Bulletin of the WHO*, 88, 831–838C.
- Breiding, M. J., & Wiersema, B. (2006). Variability of undetermined manner of death classification in the US. *Injury Prevention*, 12(Suppl. 2), ii49–ii54.
- Brenner, R. A., Trumble, A. C., Smith, G. S., Kessler, E. P., & Overpeck, M. D. (2001). Where children drown, United States, 1995. *Pediatrics*, 108(1), 85–89.
- Brillinger, D. R. (1986). The natural variability of vital-rates and associated statistics. *Biometrics*, 42(4), 693–712.
- Bureau of Labor Statistics. (2007). *Fatal workplace injuries in 2006: A collection of data and analysis*. Washington, DC: Author.
- Cameron, C. M., Purdie, D. M., Kliewer, E. V., & McClure, R. J. (2005). Long-term mortality following trauma: 10 year follow-up in a population-based sample of injured adults. *The Journal of Trauma*, 59(3), 639–646.
- Centers for Disease Control and Prevention. (1997). Recommended framework for presenting injury mortality data. *MMWR Recommendations and Reports*, 46(RR-14), 1–30.
- Centers for Disease Control and Prevention. (1999). Ten great public health achievements – United States, 1900–1999. *MMWR*, 48(12), 241–243.
- Centers for Disease Control and Prevention. *CDC WONDER*. Accessed March 24, 2011, from <http://wonder.cdc.gov/>.
- Chen, L. H., Baker, S. P., & Li, G. H. (2005). Drinking history and risk of fatal injury: Comparison among specific injury causes. *Accident Analysis and Prevention*, 37(2), 245–251.
- Committee for the Workshop on the Medicolegal Death Investigation System. (2003). *Medicolegal Death Investigation System: Workshop summary*. Washington, DC: The National Academies Press.
- Conner, K. R., Cox, C., Duberstein, P. R., Tian, L. L., Nisbet, P. A., & Conwell, Y. (2001). Violence, alcohol, and completed suicide: A case-control study. *The American Journal of Psychiatry*, 158(10), 1701–1705.
- Cryer, C., Langley, J. D., Jarvis, S. N., Mackenzie, S. G., Stephenson, S. C., & Heywood, P. (2005). Injury outcome indicators: The development of a validation tool. *Injury Prevention*, 11(1), 53–57.
- Cryer, C., Gulliver, P., Samaranayaka, A., Davie, G., & Langley, J. (2010). *New Zealand Injury Prevention Strategy indicators of injury death: Are we counting all the cases?* Dunedin, New Zealand: University of Otago, Injury Prevention Research Unit.
- Cubbin, C., LeClere, F. B., & Smith, G. S. (2000a). Socioeconomic status and injury mortality: Individual and neighbourhood determinants. *Journal of Epidemiology and Community Health*, 54(7), 517–524.
- Cubbin, C., LeClere, F. B., & Smith, G. S. (2000b). Socioeconomic status and the occurrence of fatal and nonfatal injury in the United States. *American Journal of Public Health*, 90(1), 70–77.
- Dahlberg, L. L., Ikeda, R. M., & Kresnow, M. J. (2004). Guns in the home and risk of a violent death in the home: Findings from a national study. *American Journal of Epidemiology*, 160(10), 929–936.
- Davies, M., Connolly, A., & Horan, J. (2001). *State injury indicators report*. Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.
- Department of the Interior Census Office. (1896). *Vital and social statistics in the United States at the eleventh census 1890. Part I. Analysis and rate tables* (1078 pp). Washington, DC: Government Printing Office.
- Driscoll, T., Henley, G., & Harrison, J. E. (2003). *The National Coroners Information System as an information tool for injury surveillance*. Canberra, Australia: Australian Institute of Health and Welfare.
- Fingerhut, L. A., & McLoughlin, E. (2001). Classifying and counting injury. In F. P. Rivara, P. Cummings, T. D. Koepsell, D. C. Grossman, & R. V. Maier (Eds.), *Injury control: A guide to research and program evaluation*. Cambridge: Cambridge University Press.
- Fingerhut, L. A., & Warner, M. (2006). The ICD-10 injury mortality diagnosis matrix. *Injury Prevention*, 12(1), 24–29.

- Flanagan, R. J., & Rooney, C. (2002). Recording acute poisoning deaths. *Forensic Science International*, 128(1–2), 3–19.
- Fottrell, E., & Byass, P. (2010). Verbal autopsy: Methods in transition. *Epidemiologic Reviews*, 32(1), 38–55.
- Genovesi, A. L., Donaldson, A. E., Morrison, B. L., & Olson, L. M. (2010). Different perspectives: A comparison of newspaper articles to medical examiner data in the reporting of violent deaths. *Accident Analysis and Prevention*, 42(2), 445–451.
- Hickman, M. J., Hughes, K. A., Strom, K. J., & Roper-Miller, J. D. (2007). *Medical examiners and coroners' offices, 2004*. Washington, DC: Bureau of Justice Statistics.
- Hill, K., Lopez, A. D., Shibuya, K., & Jha, P. (2007). Who counts? 3 – Interim measures for meeting needs for health sector data: Births, deaths, and causes of death. *The Lancet*, 370(9600), 1726–1735.
- Holder, Y., Peden, M., Krug, E., Lund, J., Gururaj, G., & Kobusingye, O. (2001). *Injury surveillance guidelines*. Geneva, Switzerland: World Health Organization.
- Hoyert, D. L., Kung, H. C., & Xu, J. (2007). *Autopsy patterns in 2003* (Vital and health statistics). Hyattsville, MD: National Center for Health Statistics.
- Kaplan, M. S., Hugueta, N., McFarland, B. H., & Newsom, J. T. (2007). Suicide among male veterans: A prospective population based study. *Journal of Epidemiology and Community Health*, 61(7), 619–624.
- Karch, D. L., & Logan, J. E. (2008). Data consistency in multiple source documents – findings from homicide incidents in the National Violent Death Reporting System, 2003–2004. *Homicide Studies*, 12(3), 264–276.
- Kresfeld, R. S., & Harrison, J. E. (2007). *Use of multiple causes of death data for identifying and reporting injury mortality*. *Injury Technical Papers*. Canberra, Australia: Flinders University.
- Kung, H. C., Pearson, J. L., & Liu, X. H. (2003). Risk factors for male and female suicide decedents ages 15–64 in the United States – results from the 1993 National Mortality Followback Survey. *Social Psychiatry and Psychiatric Epidemiology*, 38(8), 419–426.
- Langley, J. (2004). Challenges for surveillance for injury prevention. *Injury Control and Safety Promotion*, 11(1), 3–8.
- Langlois, J. A., Smith, G. S., Baker, S. P., & Langley, J. D. (1995). International comparisons of injury mortality in the elderly – issues and differences between New Zealand and the United States. *International Journal of Epidemiology*, 24(1), 136–143.
- Lawoyin, T. O., Asuzu, M. C., Kaufman, J., et al. (2004). Using verbal autopsy to identify and proportionally assign cause of death in Ibadan, southwest Nigeria. *The Nigerian Postgraduate Medical Journal*, 11(3), 182–186.
- Li, G. H., Smith, G. S., & Baker, S. P. (1994). Drinking behavior in relation to cause of death among US adults. *American Journal of Public Health*, 84(9), 1402–1406.
- Logan, J. E., Karch, D. L., & Crosby, A. E. (2009). Reducing “Unknown” data in violent death surveillance: A study of death certificates, Coroner/Medical Examiner and Police Reports from the National Violent Death Reporting System, 2003–2004. *Homicide Studies*, 13(4), 385–397.
- Lu, T. H., Walker, S., Anderson, R. N., McKenzie, K., Bjorkenstam, C., & Hou, W. H. (2007). Proportion of injury deaths with unspecified external cause codes: A comparison of Australia, Sweden, Taiwan and the US. *Injury Prevention*, 13(4), 276–281.
- Macarthur, C., & Pless, I. B. (1999). Evaluation of the quality of an injury surveillance system. *American Journal of Epidemiology*, 149(6), 586–592.
- Mann, N. C., Knight, S., Olson, L. M., & Cook, L. J. (2005). Underestimating injury mortality using Statewide databases. *The Journal of Trauma*, 58(1), 162–167.
- Mathers, C., Fat, D. M., & Boerma, J. T. (2008). *The global burden of disease: 2004 update*. Geneva, Switzerland: World Health Organization.
- McKenzie, K., Scott, D. A., Campbell, M. A., & McClure, R. J. (2010). The use of narrative text for injury surveillance research: A systematic review. *Accident Analysis and Prevention*, 42(2), 354–363.
- McKenzie, K., Campbell, M. A., Scott, D. A., Discoll, T. R., Harrison, J. E., & McClure, R. J. (2010). Identifying work related injuries: Comparison of methods for interrogating text fields. *BMC Medical Informatics and Decision Making*, 10, 19.
- Minino, A. M., Anderson, R. N., Fingerhut, L. A., Boudreault, M. A., & Warner, M. (2006). Deaths: Injuries, 2002. *National Vital Statistics Reports*, 54(10), 1–124.
- Mitchell, R. J., Williamson, A. M., & O'Connor, R. (2009). The development of an evaluation framework for injury surveillance systems. *BMC Public Health*, 9, 260.
- National Center for Health Statistics. (2003a). *Funeral director's handbook on death registration and fatal death reporting*. Hyattsville, MD: Author.
- National Center for Health Statistics. (2003b). *Physicians' handbook on medical certification of death*. Hyattsville, MD: Author.
- National Center for Health Statistics. (2010a). *Instructions for classifying the multiple causes of death, ICD-10 manual 2b*. Hyattsville, MD: Author.

- National Center for Health Statistics. (2010b). *Health, United States, 2009: With special feature on medical technology*. Hyattsville, MD: Author.
- National Center for Health Statistics. *About the National Death Index*. Accessed March 24, 2011, from http://www.cdc.gov/nchs/data_access/ndi/about_ndi.htm.
- National Center for Health Statistics. *National Vital Statistics System, Mortality Data*. Accessed March 24, 2011, from <http://www.cdc.gov/nchs/deaths.htm>.
- National Center for Injury Prevention and Control. *WISQARS (Web-based Injury Statistics Query and Reporting System)*. Accessed March 24, 2011, from <http://www.cdc.gov/injury/wisqars/index.html>.
- National Coroners Information System. *National Coroners Information System Annual Report 2008–09*. Accessed March 24, 2011, from <http://www.ncis.org.au/index.htm>.
- National Highway Traffic Safety Administration. *Fatality Analysis Reporting System (FARS)*. Accessed March 24, 2011, from <http://www.nhtsa.gov/FARS>.
- Patton, G. C., Coffey, C., Sawyer, S. M., et al. (2009). Global patterns of mortality in young people: A systematic analysis of population health data. *The Lancet*, *374*(9693), 881–892.
- Paulozzi, L. J., Mercy, J., Frazier, L., & Anest, J. L. (2004). CDC's National Violent Death Reporting System: Background and methodology. *Injury Prevention*, *10*(1), 47–52.
- Probst, C., Zelle, B. A., Sittaro, N. A., Lohse, R., Krettek, C., & Pape, H. C. (2009). Late death after multiple severe trauma: When does it occur and what are the causes? *The Journal of Trauma*, *66*(4), 1212–1217.
- Rahman, F., Andersson, R., & Svanstrom, L. (2000). Potential of using existing injury information for injury surveillance at the local level in developing countries: Experiences from Bangladesh. *Public Health*, *114*(2), 133–136.
- Rainey, D. Y., & Runyan, C. W. (1992). Newspapers: A source for injury surveillance? *American Journal of Public Health*, *82*(5), 745–746.
- Redelings, M. D., Wise, M., & Sorvillo, F. (2007). Using multiple cause-of-death data to investigate associations and causality listed on the death certificate. *American Journal of Epidemiology*, *166*(1), 104–108.
- Rodriguez, S. R., Mallonee, S., Archer, P., & Gofton, J. (2006). Evaluation of death certificate-based surveillance for traumatic brain injury – Oklahoma 2002. *Public Health Reports*, *121*(3), 282–289.
- Romano, P. S., & McLoughlin, E. (1992). Unspecified injuries on death certificates – a source of bias in injury research. *American Journal of Epidemiology*, *136*(7), 863–872.
- Rosales, M., & Stallones, L. (2008). Coverage of motor vehicle crashes with injuries in U.S. newspapers, 1999–2002. *Journal of Safety Research*, *39*(5), 477–482.
- Sanghavi, P., Bhalla, K., & Das, V. (2009). Fire-related deaths in India in 2001: A retrospective analysis of data. *The Lancet*, *373*(9671), 1282–1288.
- Schneider, K. L., & Shenassa, E. (2008). Correlates of suicide ideation in a population-based sample of cancer patients. *Journal of Psychosocial Oncology*, *26*(2), 49–62.
- Segui-Gomez, M., & MacKenzie, E. J. (2003). Measuring the public health impact of injuries. *Epidemiologic Reviews*, *25*, 3–19.
- Setel, P. W., Macfarlane, S. B., Szreter, S., et al. (2007). Who counts? 1 – a scandal of invisibility: Making everyone count by counting everyone. *The Lancet*, *370*(9598), 1569–1577.
- Smith, G. S., & Langley, J. D. (1998). Drowning surveillance: How well do E codes identify submersion fatalities. *Injury Prevention*, *4*(2), 135–139.
- Sorock, G. S., Chen, L. H., Gonzalgo, S. R., & Baker, S. P. (2006). Alcohol-drinking history and fatal injury in older adults. *Alcohol*, *40*(3), 193–199.
- Weiss, H. B., Gutierrez, M. I., Harrison, J., & Matzopoulos, R. (2006). The US National Violent Death Reporting System: Domestic and international lessons for violence injury surveillance. *Injury Prevention*, *12*(Suppl 2), ii58–ii62.
- WHO Family of International Classifications. *International Classification of External Causes of Injuries (ICECI)*. Accessed March 24, 2011, from <http://www.rivm.nl/who-fic/ICECIeng.htm>.
- Wiebe, D. J. (2003). Homicide and suicide risks associated with firearms in the home: A national case-control study. *Annals of Emergency Medicine*, *41*(6), 771–782.
- World Health Organization. (2004). *International statistical classification of diseases and related health problems, tenth revision* (2nd ed.). Geneva, Switzerland: Author.
- Xu, J. Q., Kochanek, K. D., Murphy, S. L., & Tejada-Vera, B. (2010). *Deaths: Final data for 2007*. Hyattsville, MD: National Center for Health Statistics.

Chapter 2

Surveillance of Injury Morbidity

Li-Hui Chen and Margaret Warner

Introduction

The Centers for Disease Control and Prevention (CDC) defines surveillance as “the ongoing systematic collection, analysis, and interpretation of health data, essential to the planning, implementation, and evaluation of health practice, closely integrated with the timely dissemination of these data to those who need to know” (Centers for Disease Control and Prevention 1996). The surveillance of injury morbidity shares many of the same characteristics as surveillance for other causes of morbidity (Horan and Mallonee 2003; Johnston 2009; Pless 2008).

Injury surveillance data can be analyzed for a variety of purposes including: detecting injury trends, measuring the size of the problem, identifying high-risk populations, projecting resource needs, establishing priorities, developing prevention strategies, supporting prevention activities, and evaluating prevention efforts.

Nonfatal injury contributes much to the burden of injury, as only a small proportion of injuries result in death. For each injury death, there are over ten injury hospitalizations and nearly 200 emergency department (ED) visits (Bergen et al. 2008). Nonfatal injuries differ from fatal injuries not only in their magnitude but also in their attributes. For example, nonfatal injuries have a wide range of outcomes from transient to lifelong effects.

This chapter complements the chapter on surveillance of injury mortality. It emphasizes issues relevant to the surveillance of nonfatal injuries and focuses on existing data systems. Several issues concerning injury morbidity surveillance are described and discussed including: data sources, classification, definitions, presentation and dissemination, evaluation and improvement, and future directions. Methods to establish surveillance systems can be found elsewhere (Holder et al. 2001; Sethi et al. 2004). Examples and issues described in the chapter, unless otherwise noted, focus on surveillance in the USA.

L.-H. Chen, PhD (✉) • M. Warner, PhD

National Center for Health Statistics, Centers for Disease Control and Prevention, Office of Analysis and Epidemiology, 3311 Toledo Road, Room 6423, Hyattsville, MD 20782, USA
e-mail: li-hui.chen@cdc.hhs.gov; mwarner@cdc.gov

Data Sources for Injury Morbidity Surveillance

Several types of data sources are available for injury morbidity surveillance. The data systems covered in this chapter are organized into three sections: health care provider-based data, population-based data, and other sources of injury data. Many common sources of injury data can be categorized into two groups: data collected from administrative or medical records at locations where injuries are treated, referred to here as health care provider-based data; and data collected from people who may or may not have had an injury and are respondents in a survey of a defined population, referred to here as population-based data.

This section focuses on general methods and analytic issues for selected data sources and provides examples of established data systems based on the data sources. More exhaustive information on data systems for injury morbidity surveillance is available elsewhere. For instance, the CDC provides a list of 44 national data systems for injury research in the USA ([National Center for Injury Prevention and Control 2011](#)) and a review of the data sources used for monitoring the objectives of Healthy People 2010 and Healthy People 2020 (US Department of Health and Human Services 2000, 2020, 2011).

Health Care Provider-Based Data

Health care facilities, where people receive medical treatment for injuries, provide a source of injury data. Health care provider-based data can be used for routine surveillance as well as to obtain information on serious and rare injuries. Data collected at health care facilities are usually based on administrative or medical records and provide more detail and higher quality medical information than data collected from a population-based survey. However, compared with population-based data, health care provider-based data generally have relatively little detail on demographic characteristics and cause of injury and even less on injury risk factors. Health care provider-based data systems may collect information from all records from all facilities, a sample of records from all facilities, all records from a sample of facilities, a sample of records from a sample of facilities, or may use an even more complex sampling strategy.

The number of health care events per person in a population, the utilization rate, is often calculated using health care provider-based data and a population count obtained from another data source. Defining the population denominator for rate calculations involves careful consideration. Details on defining populations are addressed in the *Rates and Population Coverage* section of this chapter.

The three main types of health care provider-based injury data, ED data, hospital inpatient data, and trauma registries, are described in more detail below. Other health care provider-based data, such as data from visits to physician offices and hospital outpatient departments ([National Center for Health Statistics 2011](#); Schappert and Rechsteiner 2008), pre-hospital emergency medical services (NEMSIS Technical Assistance Center 2011), and poisoning control centers (Bronstein et al. 2009), are not covered in this chapter but should be considered for analysis.

Emergency Department Data

About 20% of the US population seeks medical care in EDs at least once each year ([National Center for Health Statistics 2010](#)). Injuries account for about 30% of initial visits to EDs for medical care

(Bergen et al. 2008). ED visit data include visits for injuries with a wide spectrum of severity since people may seek primary care for some minor injuries at EDs (Institute of Medicine 2007) and people may enter the health care system for major trauma through EDs. Therefore, the ED is a logical place to collect information for a basic understanding of medically attended injuries.

The ED component of the National Hospital Ambulatory Medical Care Survey (NHAMCS) and the National Electronic Injury Surveillance System-All Injury Program (NEISS-AIP) are two examples of federal data systems that provide national estimates of injury-related visits based in the ED. NHAMCS collects data on ED visits using a national probability sample of visits to the EDs of nonfederal, general, and short-stay hospitals (National Center for Health Statistics 2011). NEISS-AIP collects information on initial visits for injuries treated in a nationally, representative sample of 66 hospital EDs that have at least six beds and provide 24-h emergency services (Schroeder and Ault 2001; National Center for Injury Prevention and Control 2011). In 2007, 27 states and the District of Columbia (DC) had a hospital ED data system (HEDDS) and 18 states mandated E-coding in their statewide HEDDS (Annest et al. 2008).

Data from EDs generally have more detail on the cause of injury but have less detail on injury diagnosis and outcome than data from an inpatient setting. Data from the ED are often collected using a simple form requiring few details to minimize the impact on health care providers in the time-sensitive ED environment. In addition, since the ED is often the first place of treatment and patients are either transferred or discharged quickly, the outcome of the injury may be unknown.

Hospital Inpatient Data

Injuries account for about 6% of hospital discharges in the USA (Bergen et al. 2008). Hospital inpatient data are often used for injury morbidity surveillance since they include injuries that are severe enough to require hospitalization. Because patients admitted to the hospital usually have longer stays than those treated in the ED, hospital inpatient records usually contain more detailed and accurate information about the diagnosis of injury than ED visit records (Farchi et al. 2007).

Examples of federal data sources based on hospital inpatient records include: the National Hospital Discharge Survey (NHDS) (Hall et al. 2010), which is an annual national probability sample survey of discharges from nonfederal, general, and short-stay hospitals; and the Healthcare Cost and Utilization Project Nationwide Inpatient Sample (HCUP-NIS) (Agency for Healthcare Research and Quality 2011), which includes all discharge records collected from a subset of hospitals. In 2008, the HCUP-NIS included discharges from hospitals located in 42 States and represented approximately 90% of all hospital discharges in the USA. In 2007, 45 states and DC had a statewide electronic hospital discharge data system (HDDS) and 26 states and DC mandated E-coding in their statewide HDDS database (Annest et al. 2008).

Hospital inpatient data often have limited information on the cause of injury. The cause of injury may not be included in the medical record of the injured person, and if it is included, it may not be collected and coded in hospital settings (McKenzie et al. 2008; McKenzie et al. 2009; McKenzie et al. 2006; Langley et al. 2007). More detail on cause of injury reporting in the inpatient data can be found in the *Classification* section of this chapter.

Trauma Registries

Trauma registries collect data on patients who receive hospital care for trauma-related injuries and usually come from trauma centers, which are often located in or affiliated with hospitals. The data are primarily used in studies of the quality of trauma care and of outcomes in individual institutions

and trauma systems, but can also be used for the surveillance of injury morbidity ([National Highway Traffic Safety Administration 2011](#); [Moore and Clark 2008](#)). Trauma registries usually involve records from a large number of patients with wide variation in data quality. However, the data are specialized for trauma research and often include more clinical details about the injuries including classification of the injury using the Abbreviated Injury Scale (AIS) ([Gennarelli and Wodzin 2006](#)). Trauma data may include limited information on the circumstances or causes of injury.

The National Trauma Data Bank (NTDB) is the largest trauma registry in the USA, and in 2009, included data on more than 4 million patients treated at more than 600 registered trauma centers ([American College of Surgeons Committee on Trauma 2011](#)). Trauma centers voluntarily participate in the NTDB by submitting data. For patient records to be included in the NTDB, the record must include at least one injury condition as defined by ICD-9-CM and the patient must have been treated at a participating trauma center.

Determining the population covered by trauma centers is especially challenging since trauma registries typically collect data from trauma centers that participate voluntarily ([Moore and Clark 2008](#)). When using trauma registry data, it is also important to consider the characteristics of patients who are treated in trauma centers. Treatment in a trauma center is known to be related not only to the nature and severity of an injury but also to factors not related to the injury (e.g., distance to the trauma center). Trauma centers vary by many factors (e.g., region of the country and number of patients treated) ([MacKenzie et al. 2003](#)).

Population-Based Data

Population-based data are collected from survey respondents who may or may not have had an injury. Injury data from population-based surveys are not dependent on where medical care was sought, and thus can be used to monitor the full severity spectrum of nonfatal injury. In addition, the population is defined as part of the sample design, and this facilitates rate calculations. Data are usually gathered using questionnaires administered either by mail, by telephone, in-person, or using a combination of these modes. Injury data may be self-reported or reported by a proxy, who is generally a family member.

In contrast to information collected from medical records, information collected from people can provide details about the circumstances surrounding a specific injury (e.g., cause of injury, place where injury occurred, activity when injured) and more information about the demographics, income, preexisting health and environment of the injured person. Population-based data can also provide information about behaviors associated with injury (e.g., drinking and driving, wearing a helmet), and knowledge and beliefs about risky behaviors and preventive measures.

Unlike information collected from medical records, memory and other human factors affects the accuracy and completeness of data collected from people. Injury severity may influence memory. Minor injury is a common event, and may not be remembered by the person responding; whereas severe injury is a relatively infrequent event, and is less likely to be forgotten. In a population-based data source, the number of respondents needed to yield enough injuries to have an adequate sample for analysis is quite large. One way to increase the number of injury events reported by respondents is to increase the length of the time period over which the respondent is asked to report the event; however, increasing the length of the time period may result in asking respondents to report minor injuries that they have forgotten. This presents two measurement issues; first is the need to set a severity threshold for identifying injuries of interest, and second is the need to determine a time period over which the respondent is asked to remember injuries of interest.

An ideal severity threshold (i.e., the minimal level of injury severity covered) would be one that is influenced only by injury severity and not by other factors. In general, the more severe the injury,

the less the threshold will be influenced by extraneous factors. Typical severity thresholds on household surveys are defined by whether medical care was sought for the injury and/or by a time period of restricted activity (e.g., 1 day or 3 days) (Heinen et al. 2004). However, with these low severity thresholds, many factors other than injury severity (e.g., health insurance status and employment status) can influence whether the injury sustained meets the severity threshold, and therefore lead to variation in the severity of injuries reported.

The length of time over which persons will likely remember the injuries of interest is important because the longer the reference period (i.e., the length of time between injury and the interview specified in the questionnaire) the greater the number and variety of events captured for analysis. However, as events happen further in the past, people tend to forget more. Examples of periods of time over which respondents have been asked to report injuries in various household questionnaires include 1 month, 3 months, 1 year, and a lifetime (Heinen et al. 2004). Analysis of injury data from surveys suggests that reference periods used in survey questions of 1 year is too long and 3 months is more appropriate (Warner et al. 2005; Harel et al. 1994; Mock et al. 1999). Detailed analysis of recall periods (i.e., the length of time between the injury and the interview) shows that for less severe injuries, a shorter recall period such as 1 month is more appropriate (Warner et al. 2005).

Some surveys, such as the National Health Interview Survey (NHIS), which is described in the next section, collect enough information about the date of injury to allow subsetting the data by different recall periods. Therefore, analysts can choose to use a shorter time period as the time period for analyzing the data. This would be a good procedure, for example, when an analysis of relatively minor injuries is being conducted.

Respondents may be unwilling to report personal information on sensitive topics, such as domestic violence or drug use, using typical survey procedures. To address this issue, methodologists have designed questionnaires and techniques to administer surveys in a more sensitive manner using technology such as computer-assisted self-interviewing (CASI), which allow respondents to report sensitive information by themselves and in private with the interviewer blinded to the responses. These techniques are used to capture sensitive information on illicit drug use and related health in the National Survey on Drug Use and Health (NSDUH) (Substance Abuse and Mental Health Services Administration 2011).

Most population-based injury morbidity data are collected using cross-sectional surveys. Longitudinal surveys are less common than cross-sectional surveys but are important for some specific injury research objectives such as cost estimation or outcomes research. Cross-sectional and longitudinal surveys are described below.

Cross-Sectional Surveys

Cross-sectional surveys collect information on a defined population at a particular point in time (Last 2001) and can be used to estimate prevalence of health conditions. If the cross-sectional survey asks about new cases of health conditions within a specified period of time, then prevalence estimates may be used to approximate incidence estimates. For example, for some acute injuries (e.g., lower extremity fracture) resulting from events that are relatively rare and that occur at a defined point in time (e.g., motor vehicle crashes), prevalence estimates can be used to approximate incidence. For chronic injury (e.g., knee and back strain) resulting from events that are more common and that may not occur at a defined point, prevalence estimates cannot approximate incidence.

The NHIS, which collects detailed information on health, including injury events, is an example of a cross-sectional survey. NHIS (National Center for Health Statistics 1997) is a household in-person survey conducted using computer-assisted personal interviewing (CAPI) of a representative

sample of the US civilian, noninstitutionalized population. Many countries have population-based health surveys that include questions about injuries (McGee et al. 2004).

Longitudinal Surveys

Longitudinal surveys collect information on a defined population on more than one occasion over a period of time (Korn and Graubard 1999). Longitudinal surveys are sometimes referred to as panel surveys. Comparisons of longitudinal surveys and cross-sectional surveys can be found elsewhere (Korn and Graubard 1999). For injury, longitudinal surveys can be useful for obtaining information on injury outcomes, such as functional limitations or resulting disability, or information on details that may suffer from recall bias, such as medical expenditures for an injury event.

An example of a longitudinal survey is the Medical Expenditure Panel Survey (MEPS) (Agency for Healthcare Research and Quality 2011), which produces nationally representative estimates of health care use, expenditures, sources of payment, insurance coverage, and quality of care for the US civilian noninstitutionalized population. MEPS consists of three components: a household component, medical provider component, and insurance component. The household data are collected from a nationally representative subsample of previously interviewed NHIS households over a period of 2 years through several rounds of interviews and medical record reviews. MEPS was the data source for several cost of injury studies (Centers for Disease Control and Prevention 2004; Corso et al. 2006) and for the Cost of Injury Reports module of the on-line fatal and nonfatal injury data retrieval program Web-based Injury Statistics Query and Reporting System (WISQARS) (National Center for Injury Prevention and Control 2011).

Other Sources of Injury Data

Besides health care provider-based and population-based data sources, data collected from other sources can be used for injury surveillance. Data collected from police reports provide information for injury events that involve the police such as car crashes or violence involving firearms. For example, the National Automotive Sampling System-General Estimates System (NASS-GES) (National Highway Traffic Safety Administration 2010) is a nationally representative sample of police-reported motor vehicle crashes in the USA and was used to study the effect on teenage drivers of carrying passengers (Chen et al. 2000). Data collected from fire departments provide information on circumstances of fire-related injuries. For example, the National Fire Incident Reporting System (NFIRS) (US Fire Administration 2011) is an on-line system in the USA where fire departments report fires; it was used to identify fires started by children playing with lighters (Smith et al. 2002). Data collected from Workers' Compensation Claims provide information on the cause of the injury, occupation, and medical cost for work-related injuries in the USA. For example, information from worker's compensation was used to study work-related eye injuries (McCall et al. 2009) and medical cost of occupation injuries (Waehrer et al. 2004).

Data collected from syndromic surveillance systems, which are designed to identify clusters of health conditions early so that public health agencies can mobilize and provide rapid responses to reduce morbidity and mortality, can be used to monitor emerging injury problems such as natural disasters, terrorism events, and mass casualty events. For example, North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) is a syndromic surveillance system; it has been used to monitor heat waves (Rein 2010) and has recently added data from the poison control center to monitor poisonings.

Classification of Injury Morbidity

Morbidity data are classified for clinical and research applications and for billing. In health care provider-based data, diagnoses and procedures are classified and, in some cases, external causes of injury are also classified. In a population-based survey, the respondent's description of an injury and its circumstances may be classified. This section provides information on clinical modifications to the International Classification of Diseases (ICD) for classifying external cause of injury and nature of injury.

International Classification of Diseases, Clinical Modifications

Clinical modifications to the ICD provide the additional codes needed for classifying the clinical detail available in many medical settings for all causes of diseases and injuries. The clinical modifications to the Ninth Revision of ICD (ICD-9-CM) are updated annually to allow for new medical discoveries and medical advancements, and for other administrative reasons. ICD-9-CM coding is required for all Medicare claims and is used by many insurers for billing in the USA. It is also used for coding patient diagnoses and procedures for many health care provider-based surveys in the USA, including the NHAMCS and the NHDS. In addition, it is used in some US population-based surveys such as NHIS, for coding the respondents' answers to questions on the cause and nature of injuries.

Many countries have developed clinical modifications to ICD-10 to classify morbidity data in their countries (Jette et al. 2010). The Australian modifications (ICD-10-AM) have been adapted for use in several countries. In the USA, the clinical modifications to the 10th Revision of ICD (ICD-10-CM) is scheduled to be implemented in the fall of 2013 (National Center for Health Statistics 2011).

External Cause of Injury Codes

The ICD External Causes of Injury and Poisoning codes, commonly referred to as E-Codes, are used to describe both the intent of injury (e.g., suicide) and the mechanism of injury (e.g., motor vehicle crash). The External Cause of Injury Matrix cross-classifies the ICD-9-CM E-codes so that injuries can be analyzed either by intent or by mechanism of injury. The matrix is updated regularly and is available at: <http://www.cdc.gov/ncipc/osp/matrix2.htm>.

E-codes describe the cause of injury and, therefore, are critical for injury prevention (National Center for Injury Prevention and Control 2009). The quality of E-coding, both in terms of completeness and accuracy, should be assessed when analyzing injury morbidity data based on administrative or medical records. Because the primary purpose of many administrative records is billing, the importance of E-codes may not be apparent to all health care providers; therefore, health care records do not always include E-codes.

In the USA and internationally, the completeness and accuracy of E-codes have been evaluated, and the results vary by geographic region, age, injury diagnosis, and data source (Langley et al. 2007; Hunt et al. 2007; Langley et al. 2006; LeMier et al. 2001; MacIntyre et al. 1997). In the USA, states with laws or mandates requiring E-coding have more complete E-coding on average than states without the requirement (Abellera et al. 2005). However, some states without mandates have high completion of E-codes (Annest et al. 2008). E-coding can be improved by including a desig-

nated field for reporting E-codes on billing reports (Injury Surveillance Workgroup 2007; Burd and Madigan 2009). In the USA, the standard, uniform bill for health care providers, the 2004 Uniform Billing Form (UB-04), which is used throughout the country, includes three fields specifically for recording E-codes (Centers for Medicare and Medicaid Services 2010).

There have been many attempts to improve E-coding in the USA. Improving E-coding in state-based ED and hospital inpatient data systems are two of the Healthy People 2020 Objectives for the Nation (US Department of Health and Human Services 2020). A recent report on E-coding is *The Recommended Actions to Improve External-Cause-of-Injury Coding in State-Based Hospital Discharge and Emergency Department Data Systems* (National Center for Injury Prevention and Control 2009). Currently, efforts are underway in the USA to recommend complete reporting of E-codes for all injury-related ED visits and hospitalizations in the Electronic Health Record (EHR) System as part of the American Recovery and Reinvestment Act of 2009 (Centers for Medicare and Medicaid Services 2011).

Nature of Injury Codes

Nature of injury codes, sometimes referred to as diagnosis codes, are used to describe both the type of injury (e.g., fracture and burn) and the body region injured. The recording of nature of injury codes is usually more complete than E-codes in the health care setting and nature of injury codes are often used to define injuries in health care provider-based data.

The Barell matrix was developed to provide a standard format for reporting injury data by the nature of injury codes. The matrix is a two-dimensional array of ICD-9-CM diagnosis codes for injuries grouped by body region of the injury and the nature of the injury. The matrix assigns injury codes into clinically meaningful groups, referred to as cells of the matrix. The cells of the matrix were designed to allow comparisons across data sources as well as over time and by geographic locations. A more detailed description of the matrix, including guidelines for its use in presenting and analyzing data, is provided in Chapter 13 (Barell et al. 2002).

Factors Affecting Case Definitions of Injury Morbidity

The case definition of injury will differ by the purpose of the analysis. For example, the case definition for monitoring utilization of medical resources for injury will differ from one for monitoring the incidence of an injury. The case definition may also differ by the data source selected. In some analyses, there may be more than one possible data source and the source that includes the most relevant information should be selected. For example, many sources provide data on nonfatal motor vehicle injuries. If the objective of the analysis is to monitor the number of crashes by type of vehicle, then NASS-GES would be a reasonable choice. However, if the objective is to monitor the number of crashes by income of the injured person, then NHIS would be a reasonable choice.

Factors affecting definitions of all injury cases and specific injury types (e.g., specific external causes, intents, and body regions) that are related to injury morbidity surveillance are described in this section with references to Chapter 1 for some general issues.

Unit of Analysis

The unit of analysis used in common case definitions of nonfatal injury varies and may be at the level of the individual (e.g., injured person), the event (e.g., motor vehicle crash), the injury (e.g., body part injured), or the contact with the health care provider (e.g., ED visit). In some instances, the unit of analysis may be at the community level. For example, communities or nursing homes may be the unit of analysis when studying initiatives for reducing fall-related injuries among older people in a defined setting (McClure et al. 2005).

The case definition should specify the unit of analysis when more than one unit is possible. For instance, a person could have more than one injury event (e.g., multiple falls), more than one injury as a result of an event (e.g., injury to the head and the neck), or more than one contact with health care providers for the same injury. The case definition should include whether the person, the event, the injury, or the health care contact is the unit of analysis. The data may need to be manipulated to produce units that are appropriate for the analysis.

Injury Incidence

For primary prevention of injury, measures of injury incidence are of greater interest than measures of health care utilization or burden of injury. To approximate injury incidence from utilization data, methods to count only the initial visit for an injury have been developed (Gedeborg et al. 2008; Du et al. 2008). However, in many systems, it may be difficult to distinguish the initial visit from follow-up visits for the same injury because the data are deidentified for confidentiality. In addition, in some data systems, it may even be difficult to identify new patients from those who transfer to or from another department within the health facility.

Surveillance using health care provider-based data may also be problematic for injury incidence estimation because whether and where an injured individual receives health care may depend on many factors other than the nature of the injury, especially for less severe injuries. For example, for minor injuries, people without health insurance coverage or people in remote areas may be less likely to seek medical care than people with health insurance or those living in cities. In addition, because of the variety of health care options in the USA, a single type of health care provider, such as a hospital ED, may not be the only contact the injured has with the health care system. For example, some people seeking medical care for a fracture will go to a hospital ED, while others will go to an orthopedic specialist.

Identifying Injury Events

Some health care provider-based data sources such as NHAMCS (National Center for Health Statistics 2011) and HCUP-NIS (Agency for Healthcare Research and Quality 2011), collect data on all encounters with a health care provider, regardless of the reason for the encounter. Injury-related encounters must be selected when using such data sources. Other health care provider-based data sources such as NEISS-AIP (Schroeder and Ault 2001; National Center for Injury Prevention and Control 2011) or the NTDB (American College of Surgeons Committee on Trauma 2011), collect information only for injury-related encounters; therefore, the criteria used to differentiate injury-related encounters from others have already been defined and implemented in the data collection process. However, the criteria used to differentiate between injury and non-injury encounters should be evaluated to see whether the criteria used are appropriate for the objective of the analysis.

In many population-based data sources, injury cases are identified by asking respondents to a survey whether they were injured. The survey questions usually involve a severity threshold (e.g., requiring medical care or resulting in restricted activity days) and a recall period (e.g., a month, year, or lifetime) (Heinen et al. 2004). Understanding how injuries are identified in a survey is critical to interpreting analyses based on the survey (Sethi et al. 2004; Chen et al. 2009).

External Cause of Injury vs. Diagnosis of Injury

Injury-related cases can be identified based on a diagnosis of injury or an external cause of injury or both. In some data sources, such as inpatient data, the recording of the external cause may be incomplete and unspecified, thus the diagnosis of injury must be relied on for defining injury cases. In other data sources, the objective of the analysis will dictate whether the case is defined by external cause or diagnosis. For example, if the objective is to analyze certain external causes of injury (e.g., motor vehicle crashes) then external causes must be used to form the case definition. If the objective is to analyze a certain body region (e.g., traumatic brain injury) or a certain type of injury (e.g., fracture), then diagnoses should be used to identify the cases of interest. If the objective is to analyze all injuries, both external causes of injury and diagnoses of injury could be used to identify cases. For example, to identify injury-related ED visits in the NHAMCS, using both external causes of injury and injury diagnoses is recommended (Fingerhut 2011).

Primary vs. Multiple Diagnoses

Many health care provider-based data sources allow for more than one diagnosis or reason for visit. When more than one diagnosis or reason is available, the number of fields searched to select the injury of interest should be considered in the case definition. For instance, some case definitions are based on an injury diagnosis only in the primary diagnosis field, some in the first-listed diagnosis field and others in any of the diagnosis fields. The number of fields considered to define injury-related health care encounters will influence the number of cases identified.

A primary diagnosis is specified in some health care provider-based data sources, and when not specified, the first-listed diagnosis is often assumed to be the primary diagnosis. Because health care provider-based data are collected with billing as the primary purpose and public health surveillance as a secondary purpose, the diagnosis listed first may be related to the cost of the injury (Injury Surveillance Workgroup 2003). Because the external cause of injury cannot be the first-listed diagnosis and may even be listed in a separate field designated for external causes, diagnosis other than the first-listed diagnosis must be used when the objective is to determine the number of hospitalizations attributed to an external cause.

The State and Territorial Injury Prevention Directors Association (STIPDA) Injury Surveillance Workgroup recommended using the nature of injury codes in the principal diagnosis field to define injury hospital discharges because it is simple, and applicable in all states (Injury Surveillance Workgroup 2003). However, if injury discharges are defined using seven diagnosis fields, then at least 30% more discharges would be designated as injury discharges than by using only the principal diagnosis field (Heinen et al. 2005). One study suggests that three diagnosis fields be considered to identify injury hospital discharges (Lawrence et al. 2007).

There is wide variation from state to state and hospital to hospital on how many diagnoses are recorded and reported. In the USA, the number of fields used to report diagnoses and other injury-related information in hospital records is increasing. This increase may lead to more opportunities to identify injuries using hospital records; in addition, the chance that multiple injuries will be recorded for a discharge may increase as well.

When more than one external cause of injury or diagnosis is used in a case definition, the method used to take into account the multiple causes or diagnoses should be described. Multiple causes or diagnoses can be taken into account using any mention, total mentions, or weighted total mentions. These methods are similar to those for injury mortality and are described in Chapter 1.

Injury Severity

When forming a case definition for injury morbidity surveillance, injury severity should be considered because injury severity varies among nonfatal injuries from minor (e.g., paper cut) to severe (e.g., gunshot to the head). Many case definitions do not explicitly state the severity, but the place of treatment for an injury provides some information about the severity threshold. For example, when using data from health care providers, it is assumed that inpatient cases are more severe than ED cases, which are more severe than cases treated in physician office visits. This implicit severity assumption should be stated explicitly by researchers (Cryer and Langley 2008).

There are many ways to measure injury severity. Established systems for measuring injury severity such as those based on AIS (Gennarelli and Wodzin 2006) [e.g., Injury Severity Score (ISS) (Baker et al. 1974) and New Injury Severity Score (NISS) (Osler et al. 1997)] and those based on ICD-9-CM [e.g., ICDMAP (Mackenzie et al. 1989) and international classification of diseases-based injury severity score (ICISS) (Osler et al. 1996; Stephenson et al. 2004)], primarily focus on threat to life. The measures focusing on threat to life may not be good measures of threat to functional limitation or disability (Expert Group on Injury Severity Measurement 2011). A severity measure that focuses on threat to functional limitation or disability may be more appropriate for some case definitions.

In some data sources (e.g., trauma registry data), a severity measure such as AIS (Gennarelli and Wodzin 2006) is provided, so that severity can be more easily specified in the case definition. Severity measures for ICD-based systems can be empirically derived using a measure such as ICISS (Osler et al. 1996; Stephenson et al. 2004). More detail about injury severity can be found in Chapter 14.

Health care provider-based data reflect, in part, guidelines for utilization and delivery of care that are extraneous to disease or injury incidence and may change over time. Injuries that meet a high severity threshold will be less influenced by these extraneous factors than injuries of minor severity. Therefore, the use of health care provider-based data to measure trends for more severe injuries better reflect injury incidence trends than such data for less severe injuries (Cryer et al. 2002; Langley et al. 2003).

For example, injury hospital discharges among persons aged 25–64 decreased an average of 5% per year from 1988 to 2000 in the USA. However, when injury severity was included in the analysis, the rates declined most for the least severe injuries (Bergen et al. 2008). Discharge rates can change for many reasons. By examining the difference in trends by severity levels, one might conclude that the observed change has more to do with a change in health care practice than with injury incidence (Fig. 2.1) (Bergen et al. 2008).

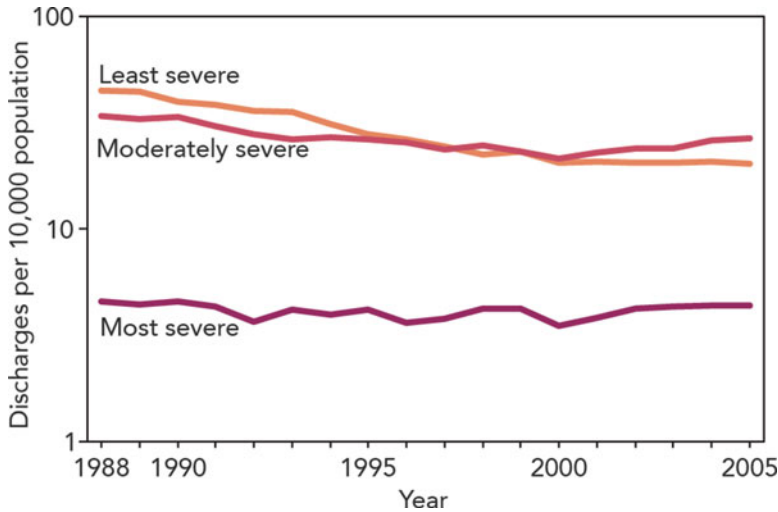


Fig. 2.1 Injury hospital discharge rates for persons 25–64 years, 1988–2005. Source: Centers for Disease Control and Prevention, National Center for Health Statistics, *Injury in the United States: 2007 Chartbook*, Figure 15.2

Data Presentation and Dissemination

The way that injury data are presented and disseminated affects their value for injury prevention and control. Stakeholders and policy makers, in particular, are more likely to use information provided in a quickly understood format. This requires careful synthesis and interpretation of the injury data. This section provides a brief description of standard sets of injury indicators developed to summarize injury morbidity surveillance data, followed by a discussion of analytical issues such as variance estimation and rate calculation, and the interpretation of trend data from surveillance systems. The section concludes with a description of common modes of dissemination including standard publications, micro-data, and other on-line resources.

Injury Indicators

According to the CDC “An injury indicator describes a health outcome of an injury, such as hospitalization or death, or a factor known to be associated with an injury, such as a risk or protective factor, among a specified population.” (Davies et al. 2001) Injury indicators can be used to identify emerging problems, show the magnitude of a problem, track trends, make comparisons among different geographic areas and different populations, and help to determine the effectiveness of interventions (Cryer and Langley 2008; Lyons et al. 2005). The International Collaborative Effort on Injury Statistics (ICE) and its members have developed a set of criteria to determine the validity of injury indicators. According to ICE, a good injury indicator should include: (1) a clear case definition, (2) a focus on serious injury, (3) unbiased case ascertainment, (4) source data that are representative of the target population, (5) availability of data to generate the indicator, and (6) the existence of a full written specification for the indicator (Cryer et al. 2005).

In the USA, injury indicators are included in the set of indicators used to monitor the health of the nation in the Healthy People initiative (US Department of Health and Human Services 2000; US Department of Health and Human Services 2020). In addition, the Council of State and Territorial Epidemiologists has identified a set of state injury indicators to monitor injuries and risk factors (Injury

Surveillance Workgroup 2007). The definitions of the indicators as well as recent statistics from the state injury indicator report can be found at: http://apps.nccd.cdc.gov/NCIPC_SII/Default/default.aspx.

In New Zealand, the standard set of injury indicators used to monitor that nation's health includes several that have explicit injury severity thresholds and include only severe injuries (Cryer et al. 2007). By eliminating the less severe injuries that may be more likely to be affected by changes in health care utilization, the indicators are more comparable over time.

Analytic Issues

There are several analytic issues to consider when presenting and disseminating injury morbidity data. This section describes some of these issues including sample weights and variance estimation, rates and population coverage, and trends.

Sample Weights and Variance Estimation

Morbidity data are usually based on sample surveys and estimates from the surveys are subject to sampling variation. Therefore, both sample weights and variance need to be considered when estimates are calculated. Sample weights take into account the sample design and adjust for nonresponse and are, therefore, usually developed by the data provider. Guidance on how to appropriately weight the sample and estimate the variance is usually addressed in the data documentation. If multiple years of survey data are analyzed, there may be further issues to consider in estimating variation since the design of the sample may change over time.

In some cases, estimates may be unreliable because the sample size is not large enough to provide a stable estimate. One measure of an estimate's reliability is its relative standard error (RSE), which is the standard error divided by the estimate, expressed as a percentage. For example, *Health, United States*, an annual report on the health of the US population, uses the following guidelines for presenting statistics from national surveys: estimates are considered unreliable if the RSE is greater than 20%; published statistics are preceded by an asterisk if the RSE is between 20 and 30% and are not shown if the RSE is greater than 30% (National Center for Health Statistics 2010).

Rates and Population Coverage

Rates are commonly disseminated and are usually estimated for a population (e.g., rate per 100,000 persons). This assumes that the entire population is at risk for the injury. However, in some cases, the entire population may not be at risk for injury. To address this with motor vehicle data, denominators such as number of miles driven or number of registered vehicles are also used for rate calculations.

For health care provider-based data sources, determining the most appropriate population for rate calculations may not be straightforward. If the health care provider-based data are nationally representative, national population estimates from the Census Bureau can be used as the denominator. If the data are not nationally representative, the population covered by health care facilities may be difficult to determine. For instance, should a state, city, or geographic boundary be used to define the population covered? A review of injury epidemiology in the UK and Europe found that mismatch between numerator and denominator is a common problem for research aiming to provide injury incidence rates for a population (Alexandrescu et al. 2009).

However, even in a national survey, selecting the most appropriate population for injury rate calculations may not be completely straightforward. For example, NHAMCS can be used to estimate the number of injury visits to EDs in nonfederal short-stay or general hospitals, which is the numerator for the rate of injury ED visits. The population used in the denominator may be the noninstitutionalized civilian population or, alternatively, the total civilian population. Some institutionalized persons (e.g., people living in nursing homes) may use the EDs, particularly for injuries, so including these persons by using the total civilian population as the denominator may be more appropriate than using the noninstitutionalized civilian population (Fingerhut 2011).

For population-based data sources, the sample is usually drawn from a well-defined population, so both the numerator and denominator for rate calculations can be estimated directly from the same data source. For example, injury rates from NHIS can be calculated as the number of injuries estimated using NHIS divided by the population estimated using NHIS (Chen et al. 2009).

Trends

Surveillance systems are often used to measure trends in injury morbidity. This section describes some issues to consider when interpreting these trends. When changes in trends are detected by surveillance, analysts need to consider all possible reasons why a change in injury morbidity estimates may have occurred. Many factors could artificially influence the trend including factors related to: data collection such as a change in questionnaire; classification systems such as coding changes; dissemination such as a change in injury definition; and/or utilization of medical care such as a change in the setting of care. If possible, other data sources measuring a similar trend should be examined.

Two examples illustrate the importance of considering possible factors that might influence a trend. The first example is from injury estimates based on NHIS. The injury rates based on NHIS were lower during 2000–2003 compared with 1997–1999 and 2004 and beyond (Chen et al. 2009). However, because the questionnaire was revised, it is likely that the change reflects changes in the questionnaire. The second example is from injury-related visits to EDs reported in *Health, United States* (National Center for Health Statistics 2010). The reported injury-related ED visit rate was 1,267 per 10,000 persons in 1999–2000 and 994 per 10,000 persons in 2006–2007. However, a footnote indicates that the estimates starting with 2005–2006 were limited to initial visits for the injury. Since there was a change in injury definition, one cannot conclude that there is a decrease in injury-related ED visit rates.

When interpreting trends produced from cross-sectional survey data, one needs to be aware that changes in the population may affect trends (Flegal and Pamuk 2007). Cross-sectional surveys such as NHIS provide information about a population at a certain point in time. Possible changes in the population, such as an increase in the immigrant population or an increase in the percentage of people who are over age 65, need to be considered. Age adjustment can be used to eliminate differences in rates due to differences in the age composition of the population over time or across population subgroups, and should be considered when examining trends across time or across subgroups defined by sex and race/ethnicity groups or by geographic location.

Standard Publications, Micro-data and Online Resources

Standard annual or periodic publications to present summarized data and inform stakeholders of key results are often produced for large, on-going data systems. For example, National Center for Health Statistics (NCHS) publishes summary health statistics for the US population based on NHIS data

every year and the reports include: (1) injury-related tables and (2) technical notes on methods and definitions (Adams et al. 2009). Some standard publications, such as *Health, USA*, contain statistics based on multiple data sources (National Center for Health Statistics 2010) and can be used not only to monitor statistics but also as a reference for brief descriptions of many data sources and methods used to produce the reported statistics.

Electronic micro-data files are available for many large, on-going data systems, including many of the data sources mentioned in this chapter. These data and associated documentation may be available free of charge or at cost, and are often available for downloading from the web. Some data systems such as NHIS, provide statistical software code for use with the micro-data (National Center for Health Statistics 1997). In addition, some on-line data resources provide analytic guidance and statistical software code for injury analysis. Two examples are the NCHS injury data and resource web site (<http://www.cdc.gov/nchs/injury.htm>) and the ICD Programs for Injury Categorization (ICDPIC) (<http://ideas.repec.org/c/boc/bocode/s457028.html>).

Injury morbidity data are disseminated through many on-line resources. Some on-line resources provide interactive querying capabilities so the data can be tabulated as the user requires; examples include WISQARS (<http://www.cdc.gov/injury/wisqars/index.html>), HCUPnet (<http://hcup.ahrq.gov/HCUPnet.asp>), and BRFSS (<http://www.cdc.gov/brfss/>). Other on-line resources such as the Health Indicators Warehouse (<http://healthindicators.gov/>) include pretabulated statistics for initiatives such as Healthy People 2020.

Surveillance Systems Evaluation and Enhancements

Surveillance systems should be evaluated periodically with the goal of improving the systems' quality, efficiency, and usefulness as well as determining the quality, completeness, and timeliness of the data (German et al. 2001). An evaluation of a system can also be useful for analysts because it provides information about the characteristics of the data system. This section includes a brief description of important features to evaluate and possible methods to enhance the systems.

Surveillance System Evaluation

According to the CDC, important surveillance system attributes to evaluate include simplicity, flexibility, data quality, acceptability, sensitivity, positive predictive value, representativeness, timeliness, and stability (German et al. 2001). The relative importance of these attributes depends on the objectives of the system. Because resources are usually limited, improvement in one attribute might be at the expense of another. For example, to improve data quality, more time is needed for quality control and therefore the timeliness of the system might decrease.

An evaluation framework specifically developed for injury surveillance systems includes 18 characteristics to assess the injury surveillance system's data quality, the system's operation, and the practical capability of the injury surveillance system (Mitchell et al. 2009). The framework also includes criteria for rating those characteristics.

If emerging threats are of interest, the timeliness of the data is of utmost importance. If emerging threats are not the concern, data quality may be the most important attribute. Five characteristics in the framework developed for injury surveillance systems (Mitchell et al. 2009) assess data quality: data completeness, sensitivity, specificity, positive predictive value, and representativeness. Examining the percentage of "unknown," "blank," "other specified," and "unspecified" responses to items is a simple way to examine the completeness of data (German et al. 2001; Mitchell et al. 2009).

Supplements to Surveillance Systems

Surveillance systems can be enhanced by including additional information such as narrative text and links with other data sources.

Narrative Text

Many injury surveillance systems collect narrative text that describes the injury and injury circumstances. Information on injury events from narrative text can provide more specific information than coded data (McKenzie et al. 2010; Mikkelsen and Aasly 2003). A variety of techniques including manual review, automated text search methods, and statistical tools have been used to extract data from narrative text and translate the text into formats typically used by injury epidemiologists (McKenzie et al. 2010). With the increasing capacity to store electronic data, narrative text data are becoming more available for analysis. For example, NHIS has released a file including narrative text that describes injury episodes annually since 1997. The Consumer Product Safety Commission uses narrative text in the National Electronic Injury Surveillance System to monitor emerging consumer product-related hazards. NHAMCS includes a cause of injury text box which is used for injury surveillance.

Data Linkage

Multiple data sources can be linked to provide a more comprehensive picture of an injury event than can be provided by a single data source. Linked data sources can also be used to study the risk factors for injury. The linkages can be between different sources collecting data on the same event; or they can be between a source providing data on risk factors and a source providing data on injury outcomes that are collected at a later time for the same person and produce a cohort-like dataset.

An example of data linkage from multiple data sources for the same injury event is the Crash Outcome Data Evaluation System (CODES). This system links crash records to injury records to follow persons involved in motor vehicle crashes to obtain data on injuries sustained in the crash (National Highway Traffic Safety Administration 2000). Typically crash records (e.g., police reports) include detail about the characteristics of the crash, the vehicle and the surrounding environments, but usually include limited information about the cost and outcome of the crash. On the other hand, outcome data such as emergency medical services records, hospital records, and vital statistics data, usually have limited data about the scene of crash. By linking data sources for an injury event, CODES provides data with detailed crash and outcome information (National Highway Traffic Safety Administration 2011).

An example of a cohort-like dataset is the National Health Interview Survey linked with mortality files (National Center for Health Statistics 2011). To produce such files, consenting NHIS survey participants in specific years are periodically matched with the National Death Index. The linked data can be used to investigate the association of a wide variety of risk factors from NHIS with injury mortality (National Center for Health Statistics 2011). NHIS data files have also been linked to Centers for Medicare and Medicaid Services Medicare enrollment and claims files (National Center for Health Statistics 2011) and social security benefit history data (National Center for Health Statistics 2011).

Future Directions

Timeliness of data release. Data timeliness is an attribute commonly included in surveillance system evaluations. Efforts to improve timeliness are a challenge for many large, on-going morbidity surveillance systems, as timeliness may be sacrificed to obtain higher data quality. However, technology and more standardized methods for data quality control have improved timeliness for many data sources. For example, 2009 NHIS micro-data were released for analysis in mid-2010, just 6 months after the completion of interviews. In addition, electronic methods for releasing data provide analysts more immediate data access. For example, the NEISS-AIP injury morbidity data are updated annually in WISQARs within days of the data release. With these improvements, it is becoming more realistic to use survey data to monitor emerging threats.

Electronic medical records/electronic health records. The use of electronic medical records/electronic health records (EMR/EHR) is increasing in the USA (Hsiao et al. 2011). The White House has set a goal that every American have an EMR/EHR by 2014, and financial incentives from the American Recovery and Reinvestment Act of 2009 (ARRA) will help in meeting that goal. Increasing use of electronic storage of medical information may lead to opportunities to provide more complete and accurate injury information than is available currently. However, issues of confidentiality as well as control of data quality will be important. In addition, methods of text analysis will need to be improved to capture the needed data.

Confidentiality concerns. Protection of personal information is critical for the viability of surveillance systems. With increasing amounts of data available electronically, even greater efforts are needed to protect confidentiality. The technology for protecting confidentiality during data collection, transmittal, and storage is advancing. For example, encryption can protect data from unauthorized use. However, some technology, such as transmitting data over the internet, increases the chance of leaking personal information. In addition, data linkages and increased electronic storage of data increase the amount of information available on a particular event or person, which increases the risk of identifying an event or an individual. Confidentiality concerns have led to more data being released in an aggregate form or as restricted micro-data available for monitored analysis in a place such as the NCHS Research Data Center (National Center for Health Statistics 2011). This practice is likely to increase in the future.

References

- Abellera, J., Conn, J., Annet, L., et al. (2005). How states are collecting and using cause of injury data: 2004 update to the 1997 report, Council of State and Territorial Epidemiologists. Atlanta, GA.
- Adams, P. F., Heyman, K. M., Vickerie, J. L. (2009). Summary health statistics for the U.S. population: National Health Interview Survey, 2008. *Vital Health Stat* 10(243). Hyattsville, MD: National Center for Health Statistics.
- Agency for Healthcare Research and Quality (2011). Healthcare Cost and Utilization Project (HCUP) databases. Available at: <http://www.hcup-us.ahrq.gov/databases.jsp>. Accessed 23 Mar 2011.
- Agency for Healthcare Research and Quality (2011). The Medical Expenditure Panel Survey (MEPS). Available at: <http://www.meps.ahrq.gov/mepsweb/>. Accessed 23 Mar 2011.
- Alexandrescu, R., O'Brien, S. J., & Lecky, F. E. (2009). A review of injury epidemiology in the UK and Europe: some methodological considerations in constructing rates. *BMC Public Health*, 9, 226.
- American College of Surgeons Committee on Trauma (2011). National Trauma Data Bank® (NTDB). Available at: <http://www.facs.org/trauma/ntdb/index.html>. Accessed 23 Mar 2011.
- Annet, J. L., Fingerhut, L. A., Gallagher, S. S., Centers for Disease Control and Prevention, et al. (2008). Strategies to improve external cause-of-injury coding in state-based hospital discharge and emergency department data systems: recommendations of the CDC Workgroup for Improvement of External Cause-of-Injury Coding. *MMWR Recommendations and Reports*, 57, 1–15.

- Baker, S. P., O'Neill, B., Haddon, W., et al. (1974). The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma*, *14*, 187–196.
- Barell, V., Aharonson-Daniel, L., Fingerhut, L. A., et al. (2002). An introduction to the Barell body region by nature of injury diagnosis matrix. *Injury Prevention*, *8*, 91–96.
- Bergen, G., Chen, L. H., Warner, M., et al. (2008). *Injury in the United States: 2007 Chartbook*. Hyattsville, MD: National Center for Health Statistics.
- Bronstein, A. C., Spyker, D. A., Cantilena, L. R., et al. (2009). 2008 Annual Report of the American Association of Poison Control Centers' National Poison Data System (NPDS): 26th Annual Report. *Clinical Toxicology*, *47*, 911–1084.
- Burd, R. S., & Madigan, D. (2009). The impact of injury coding schemes on predicting hospital mortality after pediatric injury. *Academic Emergency Medicine*, *16*, 639–645.
- Centers for Disease Control and Prevention. (1996). *Comprehensive plan for epidemiologic surveillance*. Atlanta, GA: Centers for Disease Control and Prevention.
- Centers for Disease Control and Prevention. (2004). Medical expenditures attributable to injuries—United States, 2000. *Morbidity and Mortality Weekly Report*, *53*, 1–4.
- Centers for Medicare and Medicaid Services (2010). Uniform Billing (UB-04) Implementation. Available at: <https://www.cms.gov/transmittals/downloads/R1104CP.pdf>. Accessed 23 Mar 2010.
- Centers for Medicare and Medicaid Services (2011). EHR Incentive Programs—Meaningful Use. Available at: https://www.cms.gov/EHRIncentivePrograms/30_Meaningful_Use.asp. Accessed 23 Mar 2011.
- Chen, L. H., Baker, S. P., Braver, E. R., et al. (2000). Carrying passengers as a risk factor for crashes fatal to 16- and 17-year-old drivers. *Journal of the American Medical Association*, *283*, 1578–1582.
- Chen, L. H., Warner, M., Fingerhut, L., et al. (2009). Injury episodes and circumstances: National Health Interview Survey, 1997–2007. *Vital Health Stat 10* (241). National Center for Health Statistics, Hyattsville, MD.
- Corso, P., Finkelstein, E., Miller, T., et al. (2006). Incidence and lifetime costs of injuries in the United States. *Injury Prevention*, *12*, 212–218.
- Cryer, C., Gulliver, P., Russell, D., et al. (2007). *A chartbook of the New Zealand Injury Prevention Strategy serious injury outcome indicators: 1994–2005*, p 134, Commissioned by New Zealand Injury Prevention Strategy Secretariat. ACC: Injury Prevention Research Unit, University of Otago, Dunedin, New Zealand.
- Cryer, C., & Langley, J. (2008). *Developing indicators of injury incidence that can be used to monitor global, regional and local trends*. Dunedin, New Zealand: Injury Prevention Research Unit, University of Otago.
- Cryer, C., Langley, J. D., Jarvis, S. N., et al. (2005). Injury outcome indicators: the development of a validation tool. *Injury Prevention*, *11*, 53–57.
- Cryer, C., Langley, J. D., Stephenson, S. C., et al. (2002). Measure for measure: The quest for valid indicators of non-fatal injury incidence. *Public Health*, *116*, 257–262.
- Davies, M., Connolly, A., & Horan, J. (2001). *State Injury Indicators Report*. National Center for Injury Prevention and Control, Atlanta, GA: Centers for Disease Control and Prevention.
- Du, W., Hayen, A., Finch, C., et al. (2008). Comparison of methods to correct the miscounting of multiple episodes of care when estimating the incidence of hospitalised injury in child motor vehicle passengers. *Accident Analysis and Prevention*, *40*, 1563–1568.
- Expert Group on Injury Severity Measurement (2011). Discussion document on injury severity measurement in administrative datasets. Available at: www.cdc.gov/nchs/data/injury/DicussionDocu.pdf. Accessed 23 Mar 2011.
- Farchi, S., Camilloni, L., Rossi, P. G., et al. (2007). Agreement between emergency room and discharge diagnoses in a population of injured inpatients: Determinants and mortality. *Journal of Trauma*, *62*, 1207–1214.
- Fingerhut, L. A. (2011). Recommended definition of initial injury visits to emergency departments for use with the NHAMCS-ED data. Available at: <http://www.cdc.gov/nchs/data/hestat/injury/injury.htm>. Accessed Mar 23.
- U.S. Fire Administration (2011). The National Fire Incident Reporting System (NFIRS). Available at: <http://nfirs.fema.gov/>. Accessed 23 Mar 2011.
- Flegal, K. M., & Pamuk, E. R. (2007). Interpreting trends estimated from national survey data. *Preventive Medicine*, *45*, 115–116.
- Gedeborg, R., Engquist, H., Berglund, L., & Michaelsson, K. (2008). Identification of incident injuries in hospital discharge registers. *Epidemiology*, *19*, 860–867.
- Gennarelli, T. A., & Wodzin, E. (2006). AIS 2005: A contemporary injury scale. *Injury, International Journal of Care Injured*, *37*, 1083–1091.
- German, R. R., Lee, L. M., Horan, J. M., et al. (2001). Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. *MMWR Recommendations and Reports* *50*, 1–35, quiz CE31–37.
- Hall, M., DeFrances, C., Williams, S., et al. (2010). *National Hospital Discharge Survey: 2007 summary, National Health Statistics Reports; No 29*. Hyattsville, MD: National Center for Health Statistics.
- Harel, Y., Overpeck, M. D., Jones, D. H., et al. (1994). The effects of recall on estimating annual nonfatal injury rates for children and adolescents. *American Journal of Public Health*, *84*, 599–605.

- Heinen, M., Hall, M., Boudreault, M., et al. (2005). *National trends in injury hospitalizations, 1979–2001*. Hyattsville, MD: National Center for Health Statistics.
- Heinen, M., McGee, K. S., & Warner, M. (2004). Injury questions on household surveys from around the world. *Injury Prevention, 10*, 327–329.
- Holder, Y., Peden, M., Krug, E., et al. (2001). *Injury surveillance guidelines*. Geneva, Switzerland: World Health Organization.
- Horan, J. M., & Mallonee, S. (2003). Injury surveillance. *Epidemiologic Reviews, 25*, 24–42.
- Hsiao, C.-J., Hing, E. S., Socey, T. C., et al. (2011). Electronic medical record/electronic health record systems of office-based physicians: United States, 2009 and Preliminary 2010 State Estimates. Available at: http://www.cdc.gov/nchs/data/hestat/emr_ehr_09/emr_ehr_09.htm. Accessed 23 Mar 2011.
- Hunt, P. R., Hackman, H., Berenholz, G., et al. (2007). Completeness and accuracy of international classification of disease (ICD) external cause of injury codes in emergency department electronic data. *Injury Prevention, 13*, 422–425.
- Injury Surveillance Workgroup. (2003). *Consensus recommendations for using hospital discharge data for injury surveillance*. Marietta, GA: State and territorial Injury Prevention Directors Association.
- Injury Surveillance Workgroup. (2007). *Consensus recommendations for injury surveillance in state health departments*. Marietta, GA: State and Territorial Injury Prevention Directors Association.
- Institute of Medicine. (2007). *Hospital-based emergency care: at the breaking point*. Washington, DC: National Academy Press.
- Jette, N., Quan, H., Hemmelgarn, B., et al. (2010). The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. *Medical Care, 48*, 1105–1110.
- Johnston, B. D. (2009). Surveillance: to what end? *Injury Prevention, 15*, 73–74.
- Korn, E., & Graubard, B. (1999). *Analysis of health surveys*. New York, NY: Wiley.
- Langley, J. D., Davie, G. S., & Simpson, J. C. (2007). Quality of hospital discharge data for injury prevention. *Injury Prevention, 13*, 42–44.
- Langley, J., Stephenson, S., & Cryer, C. (2003). Measuring road traffic safety performance: monitoring trends in nonfatal injury. *Traffic Injury Prevention, 4*, 291–296.
- Langley, J., Stephenson, S., Thorpe, C., et al. (2006). Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges. *Injury Prevention, 12*, 58–61.
- Last, J. M. (2001). *A dictionary of epidemiology*. New York: Oxford University Press.
- Lawrence, B. A., Miller, T. R., Weiss, H. B., et al. (2007). Issues in using state hospital discharge data in injury control research and surveillance. *Accident Analysis and Prevention, 39*, 319–325.
- LeMier, M., Cummings, P., & West, T. A. (2001). Accuracy of external cause of injury codes reported in Washington State hospital discharge records. *Injury Prevention, 7*, 334–338.
- Lyons, R. A., Brophy, S., Pockett, R., et al. (2005). Purpose, development and use of injury indicators. *International Journal of Injury Control and Safety Promotion, 12*, 207–211.
- MacIntyre, C. R., Ackland, M. J., Chandraraj, E. J., et al. (1997). Accuracy of ICD-9-CM codes in hospital morbidity data. Victoria: implications for public health research. *Australian and New Zealand Journal of Public Health, 21*, 477–482.
- MacKenzie, E. J., Hoyt, D. B., Sacra, J. C., et al. (2003). National inventory of hospital trauma centers. *Journal of the American Medical Association, 289*, 1515–1522.
- Mackenzie, E. J., Steinwachs, D. M., & Shankar, B. (1989). Classifying trauma severity based on hospital discharge diagnoses – validation of an ICD-9CM to AIS-85 conversion table. *Medical Care, 27*, 412–422.
- McCall, B. P., Horwitz, I. B., & Taylor, O. A. (2009). Occupational eye injury and risk reduction: Kentucky workers' compensation claim analysis 1994–2003. *Injury Prevention, 15*, 176–182.
- McClure, R., Turner, C., Peel, N., et al. (2005). Population-based interventions for the prevention of fall-related injuries in older people. *Cochrane Database of Systematic Reviews 1*.
- McGee, K., Sethi, D., Peden, M., & Habibula, S. (2004). Guidelines for conducting community surveys on injuries and violence. *Injury Control and Safety Promotion, 11*, 303–306.
- McKenzie, K., Enraght-Moony, E., Harding, L., et al. (2008). Coding external causes of injuries: Problems and solutions. *Accident Analysis and Prevention, 40*, 714–718.
- McKenzie, K., Enraght-Moony, E. L., Walker, S. M., et al. (2009). Accuracy of external cause-of-injury coding in hospital records. *Injury Prevention, 15*, 60–64.
- McKenzie, K., Harding, L. F., Walker, S. M., et al. (2006). The quality of cause-of-injury data: where hospital records fall down. *Australian New Zealand Journal of Public Health, 30*, 509–513.
- McKenzie, K., Scott, D. A., Campbell, M. A., et al. (2010). The use of narrative text for injury surveillance research: A systematic review. *Accident Analysis and Prevention, 42*, 354–363.
- Substance Abuse and Mental Health Services Administration (2011). National Survey on Drug Use and Health Available at: <http://oas.samhsa.gov/nhsda.htm>. Accessed 23 Mar 2011.
- Mikkelsen, G., & Aasly, J. (2003). Narrative electronic patient records as source of discharge diagnoses. *Computer Methods and Programs in Biomedicine, 71*, 261–268.

- Mitchell, R. J., Williamson, A. M., & O'Connor, R. (2009). The development of an evaluation framework for injury surveillance systems. *BMC Public Health*, 9, 260.
- Mock, C., Acheampong, F., Adjei, S., et al. (1999). The effect of recall on estimation of incidence rates for injury in Ghana. *International Journal of Epidemiology*, 28, 750–755.
- Moore, L., & Clark, D. E. (2008). The value of trauma registries. *Injury, International Journal of Care Injured*, 39, 686–695.
- National Center for Health Statistics. (2010). *Health, United States, 2009: With Special Feature on Medical Technology*. Hyattsville, MD: National Center for Health Statistics.
- National Center for Health Statistics (2011). Ambulatory Health Care Data-Questionnaires, Datasets, and Related Documentation. Available at: http://www.cdc.gov/nchs/ahcd/ahcd_questionnaires.htm. Accessed 23 Mar 2011.
- National Center for Health Statistics (2011). National Health Interview Survey: Questionnaires, Datasets, and Related Documentation, 1997 to the Present. Available at: http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm. Accessed 23 Mar 2011.
- National Center for Health Statistics (2011). International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Available at: <http://www.cdc.gov/nchs/icd/icd10cm.htm>. Accessed 23 Mar 2011.
- National Center for Health Statistics (2011). NHIS Linked Mortality Files. Available at: http://www.cdc.gov/nchs/data_access/data_linkage/mortality/nhis_linkage.htm. Accessed 23 Mar 2011.
- National Center for Health Statistics (2011). NCHS Data Linked to CMS Medicare Enrollment and Claims Files. Available at: http://www.cdc.gov/nchs/data_access/data_linkage/cms.htm. Accessed 23 Mar 2011.
- National Center for Health Statistics (2011) NCHS Data Linked to Social Security Benefit History Data. Available at: http://www.cdc.gov/nchs/data_access/data_linkage/ssa.htm. Accessed 23 Mar 2011.
- National Center for Health Statistics (2011). NCHS Research Data Center (RDC). Available at: <http://www.cdc.gov/rdc/>. Accessed 23 Mar 2011.
- National Center for Injury Prevention and Control. (2009). *Recommended actions to improve external-cause-of-injury coding in state-based hospital discharge and emergency department data systems*. Atlanta, GA: National Center for Injury Prevention and Control.
- National Center for Injury Prevention and Control (2011). Inventory of national injury data systems. Available at: <http://www.cdc.gov/Injury/wisqars/InventoryInjuryDataSys.html>. Accessed 23 Mar 2011.
- National Center for Injury Prevention and Control (2011). WISQARS (Web-based Injury Statistics Query and Reporting System). Available at: <http://www.cdc.gov/injury/wisqars/index.html>. Accessed 23 Mar 2011.
- National Highway Traffic Safety Administration (2000). Problems, Solutions and Recommendations for Implementing CODES (Crash Outcome Data Evaluation System), Washington DC.
- National Highway Traffic Safety Administration (2011). Trauma System Agenda for the Future Available at: <http://www.nhtsa.gov/PEOPLE/injury/ems/emstraumasystem03/>. Accessed 23 Mar 2011.
- National Highway Traffic Safety Administration (2011). The Crash Outcome Data Evaluation System (CODES) And Applications to Improve Traffic Safety Decision-Making. Available at: <http://www-nrd.nhtsa.dot.gov/Pubs/811181.pdf>. Accessed 23 Mar 2011.
- National Highway Traffic Safety Administration. (2010). *National automotive sampling system general estimates system analytical users manual 1988–2009*. Washington, DC: National Highway Traffic Safety Administration.
- Osler, T., Baker, S. P., & Long, W. (1997). A modification of the injury severity score that both improves accuracy and simplifies scoring. *Journal of Trauma*, 43, 922–925.
- Osler, T., Rutledge, R., Deis, J., et al. (1996). ICISS: An international classification of disease-9 based injury severity score. *Journal of Trauma*, 41, 380–387.
- Pless, B. (2008). Surveillance alone is not the answer. *Injury Prevention*, 14, 220–222.
- Rein, D. (2010). A snapshot of situational awareness: using the NC DETECT system to monitor the 2007 heat wave. In T. Z. X. Kass-Hout (Ed.), *Biosurveillance: a health protection priority*. Boca Raton, FL: CRC Press.
- Schappert, S., & Rechsteiner, E. (2008). Ambulatory Medical Care Utilization Estimates for 2006, in National Health Statistics Reports, National Center for Health Statistics, Hyattsville, MD.
- Schroeder, T., & Ault, K. (2001). *The NEISS sample: design and implementation, 1997 to present*. Washington, DC: US Consumer Product Safety Commission.
- Sethi, D., Habibula, S., McGee, K., et al. (2004). *Guidelines for conducting community surveys on injuries and violence*. Geneva, Switzerland: WHO.
- Smith, L. E., Greene, M. A., & Singh, H. A. (2002). Study of the effectiveness of the US safety standard for child resistant cigarette lighters. *Injury Prevention*, 8, 192–196.
- Stephenson, S., Henley, G., Harrison, J. E., et al. (2004). Diagnosis based injury severity scaling: investigation of a method using Australian and New Zealand hospitalizations. *Injury Prevention*, 10, 379–383.
- NEMSIS Technical Assistance Center (2011). National Emergency Medical Services Information System (NEMSIS). Available at: <http://www.nemsis.org/index.html>. Accessed 23 Mar 2011.

- US Department of Health and Human Services. (2000). *Healthy People 2010: Understanding and Improving Health* (2nd ed.). Washington, DC: US Department of Health and Human Services.
- US Department of Health and Human Services (2011). Healthy People 2020. Available at: <http://www.healthypeople.gov/2020/default.aspx>. Accessed 23 Mar 2011.
- US Department of Health and Human Services (2011). Health Indicators Warehouse. Available at: www.HealthIndicators.gov. Accessed 23 Mar 2011.
- Waehrer, G., Leigh, J. P., Cassady, D., et al. (2004). Costs of occupational injury and illness across states. *Journal of Occupational and Environmental Medicine*, 46, 1084–1095.
- Warner, M., Schenker, N., Heinen, M. A., et al. (2005). The effects of recall on reporting injury and poisoning episodes in the National Health Interview Survey. *Injury Prevention*, 11, 282–287.

Chapter 3

Injury Surveillance in Special Populations

R. Dawn Comstock

Introduction

Long before injury surveillance was implemented in general populations as a standard public health practice, it was adopted by military and occupational medicine to improve the effectiveness of soldiers (Manring et al. 2009; Retsas 2009; Salazar 1998) and the productivity of workers (Meigs 1948; Williams and Capel 1945). Thus, military personnel and workers were, in essence, the first special populations in which injury surveillance was conducted. In his discussion of the historical development of public health surveillance, Thacker (2000) notes the advancements made via the contributions of John Graunt, Johann Peter Frank, Lemuel Shattuck, William Farr, and others, which led to Langmuir's (1963) cultivation of the three tenets of a public health surveillance system: (a) the systematic collection of pertinent data, (b) the orderly consolidation and evaluation of these data, and (c) the prompt dissemination of results to those who need to know. It is worth noting that although early surveillance systems that primarily reported mortality data included injury, as public health surveillance advanced, the focus quickly turned to infectious disease. Although the three tenets elicited by Langmuir (1963) remain just as pertinent today as they were in 1963 and are just as applicable to injury as they are to infectious disease, it was not until the late 1980s and early 1990s that the full importance of the development/use of public health surveillance systems for injury was "rediscovered" (Graitcer 1987). The rapid technological advancements during that time period provided public health professionals and researchers with newfound capabilities to conduct public health surveillance, including dramatic advancements in injury surveillance. Continued technological advancements, particularly the microcomputer and the Internet, have similarly increased the ability to conduct injury surveillance in general population groups and, more specifically, in special populations.

Injury surveillance programs are invaluable for public health professionals, academic researchers, policy makers, and others interested in injury prevention. This is because development of effective injury prevention efforts requires a thorough knowledge of injury rates, patterns of injury, and

R.D. Comstock, PhD (✉)

Center for Injury Research and Policy, The Research Institute at Nationwide Children's Hospital,
700 Children's Drive, Columbus, OH 43205, USA

Department of Pediatrics, College of Medicine, The Ohio State University, 700 Children's Drive, Columbus,
OH 43205, USA

Division of Epidemiology, College of Public Health, The Ohio State University, Columbus, OH 43205, USA
e-mail: comstoc@pediatrics.ohio-state.edu

similar data provided by injury surveillance systems. Unfortunately, implementing and maintaining injury surveillance programs has commonly been believed to be difficult, time consuming, and costly. The many important uses of injury surveillance data should justify the effort and cost. Further, through innovative approaches including the use of advanced methodology and new technologies, injury surveillance has become easier and less expensive.

Injury surveillance programs generally fall into one of two categories – broad surveillance programs covering general population groups that frequently fail to provide data specific enough to address injury issues in special populations or surveillance programs applied specifically to special populations that may not provide the data necessary to allow comparisons to broader populations. Both categories frequently fall short of providing a full description of the epidemiology of injury in special populations. This is unfortunate as injury often poses a greater burden to special populations than the general population. Comparing injury patterns in special populations to general populations not only demonstrates such disparities but also provides insight into differences in patterns of injury and risk factors for injury. Such knowledge is required to drive the development of effective injury prevention efforts in special populations. This chapter discusses approaches to injury surveillance in special populations. Specifically, this chapter will address methodological issues in conducting injury surveillance in special populations with innovative technologies.

Injury Surveillance in Special Populations

Basic instructions on methodological approaches for conducting injury surveillance are available from several sources (Holder et al. 2001; Horan and Mallonee 2003). Like general populations, special populations can be studied using prospective or retrospective surveillance methodologies, with active or passive surveillance systems, and with surveillance systems capturing preexisting data from single or multiple sources or novel data captured specifically for the surveillance project. However, conducting surveillance in special populations frequently requires heightened sensitivity to the perception of the public health importance of the injury issue within the population, as this may drive the type of data to be collected, the method of data collection, the interpretation of collected data, as well as the distribution and implementation of findings. Whether researchers are working in the context of applied or academic epidemiology, when injury surveillance is being conducted in special populations, there should be added emphasis on striving for analytic rigor as well as public health consequence. As noted by Koo and Thacker (2000), this requires “technical competence blended with good judgment and awareness of context.”

What is a special population requiring innovative surveillance methods? In the context of injury surveillance, a special population is simply one that has received little prior attention from the public health or academic research communities or one for which little is known despite previous attention due to prior injury surveillance efforts that were unsuccessful, incomplete, or not applicable to the needs of the special population. Basically, if previous injury surveillance efforts have successfully provided the data required to fully describe the epidemiology of injury in a population, there probably is little need to adopt innovative methods to replicate previous work or to replace existing surveillance systems. That said, directors of long-standing, ongoing injury surveillance programs should constantly strive to improve their surveillance efforts by being aware of technological innovations that may either improve surveillance efforts or decrease the cost of surveillance efforts. Most importantly, directors of surveillance systems must work to identify the most effective ways to apply surveillance data to inform and assess injury prevention efforts.

An excellent example of utilizing emerging technologies and novel methodology to upgrade an existing surveillance effort is the National Violent Death Reporting System (NVDRS), which was developed to improve surveillance of violent death incidents using innovative methods to capture

and share data rapidly both from and between multiple sources (Paulozzi et al. 2004). Although violent deaths were already being captured by various vital statistics systems, public health professionals identified persons who had died of assaults or self-inflicted injury as a special population that, compared to the general population, had an excessive injury burden and for which existing surveillance systems did not provide adequate data. Thus, NVDRS was created to provide high-quality data to “open a new chapter in the use of empirical information to guide public policy around violence in the United States.” NVDRS has dramatically improved the ability to systematically collect pertinent data on violent injury deaths, to orderly consolidate and evaluate these data, and to rapidly disseminate these data so they can be used to drive prevention efforts and policy decisions – thus demonstrating the value of applying Langmuir’s approach to public health surveillance to an injury issue in a special population.

Defining the Population

Special populations include groups of individuals that are difficult to study using data from traditional, existing surveillance systems. Such groups can include individuals frequently identified as belonging to special populations from a broader public health perspective such as specific immigrant populations, small religious sects, individuals living in isolated low-resource settings, and individuals with language barriers. However, special populations in the context of injury surveillance needs may also include individuals from broader population groups who are at risk of injury due to specific age and activity combinations (e.g., young athletes and elderly pedestrians), gender and occupational combinations (e.g., female military personnel in combat zones and male nurses), or location and activity combinations (e.g., “backcountry” hikers and campers and aid workers in conflict areas). Special populations requiring innovative surveillance methods may also include individuals from the broader general public with relatively rare injury events (e.g., bear attack victims and lightning strike victims), individuals with newly identified injuries/injury syndromes (e.g., “Wii elbow” and “texting thumb”), as well as individuals participating in relatively uncommon activities (e.g., base jumping and washing external skyscraper windows). Given the wealth of data available to researchers as a result of modern public health advances, if a source of data cannot be identified for the target population, it is likely a special population that may require innovative surveillance methodologies.

Once identified as a special population, the population must still be clearly defined before surveillance efforts should begin. As in any sound epidemiologic study, clear inclusion and exclusion criteria must be applied to the special population to be included in any surveillance study. This can be challenging for special populations where defining the study population too broadly will likely result in an underestimation of the injury burden in the special population and a muting of the associations between risk or protective factors and the outcome of interest. Conversely, defining the study population too narrowly may result in an inability to accurately describe the epidemiology of injury in the special population due to a lack of generalizability of study results to the entire special population.

Whenever possible, all those at risk of the injury of interest should be under surveillance. Too often, surveillance projects in special populations have captured only injury incidence data, with no exposure data being captured for the special population. This restricts the ability to calculate injury rates, thus limiting the usefulness of the surveillance data. Having clearly defined target and study populations will improve the quality and applicability of the data captured by the surveillance system. Additionally, a clearly defined population will allow the researcher to focus personnel efforts to maximize resources and minimize costs.

Defining the Variables

As in any sound epidemiologic study, surveillance studies in special populations require clear definition of the variables to be captured. This includes the outcome of interest (e.g., the injury of interest, the clinical outcome of an injury) as well as demographic factors, potential risk factors, protective factors, confounders, and effect modifiers. Frequently, the best surveillance systems capture the minimum amount of data required to address the public health concern using the simplest format possible. This “elegant simplicity” approach tends to minimize error/bias while maximizing resources. This is particularly important in surveillance of special populations where economic and personnel resources are likely to be minimal, access to the population may be limited, sensitivity to the time burden of data reporting may be heightened, etc.

When determining which of the multitude of potential variables of interest will ultimately be included in the surveillance system’s data collection tools, consideration should be given to the goals following final interpretation of captured data. Surveillance studies in special populations often fail to include data captured from control groups due to lack of resources, feasibility, etc. Rather, surveillance data from special populations usually must be compared to previously captured data from general populations (e.g., comparing injured elderly pedestrians to all other injured pedestrians, comparing aid workers in conflict areas to individuals working in similar occupations – truck driver, physician, etc. – in peaceful settings) or previously studied somewhat similar special populations (e.g., comparing high school athletes to collegiate athletes, comparing tsunami victims to tornado victims). When determining which variables will be captured by the surveillance system, consideration should be given to whether or not data can be captured from a control population and, if not, what preexisting data may be available for comparisons of interest among potential control populations.

To accurately identify which variables should be captured by the surveillance system and the best method for data capture, the researcher must gain a deep familiarity with the special population. This includes obtaining a thorough understanding of their concerns, their culture, their common language (e.g., actual language, slang, activity-specific terminology, commonly used acronyms), their comfort level with various data collection technologies, etc. When at all possible, researchers should involve members of the special population as well as their community leaders and stakeholders in the development of and pilot testing of the surveillance system’s data collection tools. Conducting focus groups to discuss results of initial pilot tests of data collection tools can help identify which questions may be misinterpreted, which questions need additional answer options, which questions need to be added to capture key data currently missing, and which questions should be candidates for elimination if there is a need to cut time burden. For a surveillance system in a special population to be useful, the data captured must be as complete as possible, accurate, applicable to the public health concern, and acceptable to the special population.

Capturing the Population

Effectively and efficiently capturing special populations in surveillance studies can be particularly challenging. The first step after identifying a special population with an injury issue of public health importance is to determine if any existing sources of data can be used to study the problem or if novel data must be collected. Community leaders and stakeholders may be able to help identify useful preexisting data sources if they are included in a discussion of the goals of the surveillance project and data required to accomplish these goals. Available preexisting data sources may include medical records, school records, insurance records, etc., that researchers are comfortable using or they may include less familiar data sources such as church records, immigration documents, or oral histories.

Preexisting data sources, if robust enough, may eliminate the need to capture novel data from the population. However, the personnel costs of abstracting data, cleaning data, combining data sources, etc., may not be any lower than the cost of collecting novel data directly from the special population. Gaining access to the special population for direct collection of novel data requires an understanding of the culture of the special population as well as a strong relationship with community leaders and stakeholders.

Of special consideration should be the determination of the scope of data capture. Should the surveillance system capture data from a national sample of the population, a regional sample, or a local sample? Use of local samples is attractive because the researcher can establish a close relationship with the special population under study and can maintain that relationship throughout the surveillance project. This will allow the researcher to interact with the special population over time to maintain enthusiasm for the surveillance project, to quickly intervene if problems arise, and to provide feedback to the special population as data become available. However, data collected from a local sample may not be generalizable to the special population on the whole if the local sample is not representative of the broader sample. Capturing data from a larger, regional sample should increase generalizability but may reduce the completeness and accuracy of captured data if the strong relationship between researcher and special population cannot be maintained due to distance or size. Expanding the scope of the surveillance system to capture a national population should provide the most generalizable data but this will likely come at a cost to the closeness of the relationship between the researcher and the special population which, in turn, may affect the quality of collected data.

Regardless of the size of the population under surveillance, researchers can use several methods to improve the quality of captured data. Enlisting the support of community leaders and stakeholders and their assistance in engendering enthusiasm for the surveillance project from the special population in general and the data reporters specifically is important. Providing incentives for participating members of the special population and for data reporters should be strongly considered. Linking incentives to compliance with reporting methodology is likely to improve completeness of reporting and data quality. Conducting data audits throughout the surveillance project will also improve data quality. Providing feedback to the special population throughout the surveillance project and utilizing captured surveillance data to drive efforts to reduce their burden of injury is a must. Such efforts can be conducted face-to-face with local population samples. Modern communication technologies also allow such efforts to be effectively conducted in large, widely dispersed regional or national population samples with minimal research personnel time burden. Using modern computing and communication technologies can also make the cost of distributing results of surveillance efforts to regional or national samples comparable to, or even less than, distributing results to local samples.

An additional approach is to conduct surveillance in a relatively small but representative national sample, particularly if the sampling methodology enables calculation of national estimates based on data collected from the sample under surveillance. This combines the researcher's ability to establish a close relationship with the community stakeholders and leaders as well as data reporters and the ability to closely monitor the data being collected by the surveillance system with the advantages of capturing data that are generalizable to the national population at minimal economic and personnel time costs. However, the actual generalizability of the data captured from a small sample will depend heavily upon how representative the small study sample is to the broader special population as a whole. A thorough understanding of the special population is required in order to develop a sampling scheme capable of capturing a small but representative study sample.

Feasibility, Funds, and Framework

Virtually all epidemiologic methodologies are applicable to surveillance projects in special populations. However, conducting surveillance in special populations frequently requires innovative

methodologies due to feasibility, funding, and framework constraints. The most appropriate injury surveillance approach for any specific surveillance project in a special population will depend upon the feasibility of utilizing various methodologies, the funds available, and the conceptual framework within which the researcher and the special population are approaching the injury issue.

Feasibility is not usually a limiting factor in modern public health surveillance efforts that utilize proven methodologies and established technologies. However, researchers planning surveillance in a special population must thoroughly consider the feasibility of various methodological options when designing a surveillance system. For example, existing, long-standing injury surveillance systems and traditional data sources (e.g., medical records) rarely contain enough data on special populations to fully describe the epidemiology of injury among such populations. Thus, retrospective surveillance or data abstraction from existing records may not be feasible. If such methodologies can be used, they may pose a large personnel time burden and can be very expensive unless innovative computer data capture methodologies can be used. Similarly, while modern communication technology like cell phones or satellite phones can be the only way to conduct surveillance in some special populations dispersed over a wide geographic region, in other special populations (e.g., religious sects who do not use electricity and individuals living in low-income areas of developing countries), such technology may not be available or its use may not be feasible due to an inability to train and provide support for data reporters. Working closely with the community leaders and stakeholders of the special population during the early design and planning stages will help researchers gauge the relative feasibility of various methodologies. Another invaluable resource is other researchers who have conducted surveillance in somewhat similar special populations who can share their knowledge regarding potential feasibility concerns and methodologies that may be used to overcome such concerns.

Similarly, funding of modern, established public health injury surveillance systems is usually a point of concern for policy makers rather than the public health professionals or academic researchers who utilize the resulting data. This is because the maintenance of most established long-term injury surveillance systems is usually financed by government agencies. Funding concerns are an unfortunate driving force in methodological decisions during the development of most surveillance systems in special populations. Public health professionals, academic researchers, and policy makers are all aware of the value of surveillance data. Although surveillance data are widely used, there are few resources available to fund surveillance systems. Traditional funding agencies, such as the National Institute of Health and the National Science Foundation, have undervalued the scientific and public health impact of injury surveillance systems and, thus, have rarely provided funding for such efforts. The one federal agency that had traditionally provided funding for the implementation and maintenance of injury surveillance studies, the Centers for Disease Control and Prevention (CDC), has now primarily shifted its research agenda to focus on development, implementation, and evaluation of interventions. As a result, little funding for the establishment of injury surveillance systems in special populations is currently available from traditional funding sources. Researchers seeking funding for injury surveillance in special populations must frequently either disguise surveillance efforts within the type of specific hypothesis-driven research question that is currently more acceptable to federal funding review panels or they must rely upon nontraditional funding sources, which often provide only relatively small amounts of short-term funding. Such constraints are unfortunate since the true value of surveillance systems is their ability to capture large amounts of data over long periods of time to enable subgroup analyses and analyses of time trends. Current funding constraints make it more appealing to use innovative communications and computing technologies to reduce the cost of surveillance projects while expanding methodological options.

It is important to note that the framework within which the researcher and the special population approach an injury issue will also drive decisions regarding the most appropriate methodological

approach. Special populations may have sensitivity to “outsiders” entering their world and may have heightened concerns about the intended goals of surveillance efforts. If community leaders and stakeholders are not involved in the planning of the surveillance project, the special population may misinterpret researchers’ motivations and intentions. Rather than approaching a special population by telling them that they have an injury problem and what they must do to help the researcher resolve their problem, a dialog should be established to enable the researcher to understand the special population’s perception of the injury issue, their concerns regarding the injury issue and their desires regarding efforts to address the injury issue, their willingness to assist with addressing the issue via the public health approach, and their long-term goals and expectations. The long-term goal of any surveillance system should be to provide high-quality data that can be used to drive the development and implementation of evidence-based injury prevention efforts and to subsequently evaluate the effectiveness of prevention efforts by monitoring injury trends over time. Such outcomes can be accomplished only if the researcher maintains a full understanding of the framework of the injury issue within the special population.

Technology

In the mid 1980s, public health programs around the world began utilizing emerging communication and computer technologies to rapidly advance the field of surveillance. For the first time, public health entities implemented national computer-based surveillance systems that established a mechanism for numerous local entities to transmit information of public health importance into one centralized system using standardized record and data transmission protocols; this enabled rapid analyses of extremely large datasets and provided the ability to create reports of these analyses and distribute them broadly in near real time (Graitcer and Thacker 1986; Graitcer and Burton 1986). These early versions were precursors to the new generation of modern surveillance systems (Fan et al. 2010). New technologies present opportunities for centralized, automated, multifunctional detection and reporting systems for public health surveillance that are equally appropriate for large national systems applied to general populations or for small systems applied to special populations.

Technology has advanced so rapidly that a public health evolution has occurred, complete with accompanying changes in terminology. One example is “infodemiology” or “infoveillance,” defined as the science of evaluating distributions and determinants of information in an electronic medium with the aim of informing public health policy, which has been evaluated as a Web-based tool for a wide range of public health tasks including syndromic surveillance, evaluation of disparities in health information availability, and tracking the effectiveness of health marketing campaigns (Eysenbach 2006, 2009). Another example is “eHealth” or “personalized health applications,” defined broadly as a range of medical informatics applications for providing personalized Web-based interactions based on a health consumer’s specific characteristics (Pagliari et al. 2005; Bennett and Glasgow 2009; Chou et al. 2009; Fernandez-Luque et al. 2010). Additionally, emergency surveillance in disaster settings is now e-mail-based (CDC 2010), global databases capturing not only injury data but also information on treatment modalities and outcomes have been called for (Clough et al. 2010), and school nurses are being encouraged to develop “Twitter” surveys (Patillo 2010). Such advancements in the application of new computer technologies have affected every aspect of public health including surveillance, research, development and implementation of preventive interventions, and development and distribution of health information. Such changes are being hastened both by lowering costs for public health information systems in general and by an ever-opening market for free technology exchange (Yi et al. 2008).

Growing Options

A multitude of rapid technological advancements have afforded public health professionals and academic researchers with a growing number of innovative options for conducting injury surveillance in special populations. The technologies that have proven most useful for injury surveillance to date center around the Internet (including e-mail and social networks), wireless communications devices (cell phones, satellite phones, and pagers), and combinations of technologies. Inexpensive microcomputers and widespread availability of advanced programming applications have presented unprecedented computing power for screening, abstracting, collating, and analyzing incredibly large datasets from individual or multiple electronic sources. Additionally, advancements in word processing, computer graphics, and presentation tools have provided researchers with the ability to more quickly and more clearly communicate findings. The rapid increase in numbers of and access to scientific journals; the Internet complete with e-mail, blogs, and social networking tools; and the instantaneous connections of media throughout the world have established expanded audiences with which surveillance data are monitored and findings are communicated.

Access to the Internet has become nearly ubiquitous in developed countries during the past decade, ushering in a new era in public health in general as well as in surveillance projects specifically. The Internet has been used in many ways for surveillance, from conducting general Web searches for key terms linked to public health issues of interest, to soliciting responses to one-time surveys via e-mail, to using the Internet for the application of specifically designed data collection tools during long-term surveillance projects. For example, while there are many reports of general Internet searches being used for surveillance of infectious disease outbreaks (Chew and Eysenbach 2010; Corley et al. 2010), Internet searches have also been demonstrated to be useful for passive surveillance of injury issues (McCarthy 2010). At the other extreme, the Internet has been proven to be an effective and economical method for conducting active surveillance by applying the same survey to large representative samples of special populations multiple times over weeks, months, or even years during prospective surveillance studies to monitor injury rates and exposures to risk and protective factors over time (Comstock et al. 2006; Bain et al. 2010). Internet-based questionnaires have been proven to be reliable and valid for capturing exposure and outcome data (De Vera et al. 2010). In several studies of special populations, Web-based questionnaires have been shown to be cost- and time-efficient as well as capable of capturing more complete and more accurate data compared to paper questionnaires (Kypri et al. 2004; Bech and Kristensen 2009; Russell et al. 2010). The use of Internet-based questionnaires has become so commonplace that van Gelder et al. titled their recent discussion of the advantages and disadvantages of these tools “Web-based Questionnaires: The Future of Epidemiology?” (van Gelder et al. 2010). The popularity of social networks has provided another novel, incredibly fast, and inexpensive method for researchers to simultaneously identify and survey populations of interest as demonstrated by a study of students who misused prescription medications; results from study samples captured via social networks were consistent with results from traditional surveys (Lord et al. 2011). Social networking sites have also been evaluated for their potential utility in distributing and monitoring public health education messages (Ahmed et al. 2010).

Similarly, the dramatic decrease in costs of wireless communication that resulted in a widespread proliferation of communication devices, particularly cell phones, has also changed the public health landscape. Cell phones have become so prevalent in developed countries that the usefulness of land-line random digit dialing (RDD), long a staple of epidemiologic studies, has been questioned and, as fewer individuals maintain landlines, inclusion of cellular telephone numbers in RDD studies has been recommended (Voigt et al. 2011). Mobile phones have also enabled telemedical interaction between patients and health-care professionals, thus providing a novel mechanism for clinicians and researchers to monitor patients' compliance with treatment/management plans (Kollmann et al. 2007). Cell phones can also provide a cost-effective mechanism for researchers and clinicians to

assess long-term outcomes via follow-up surveys or even by providing a mechanism for individuals to take photos of recovering injuries and transmit them to researchers (Walker et al. 2011).

Combining several modern technologies is another innovative methodology available for public health activities in general as well as surveillance studies specifically. For example, research demonstrated that patient outcomes could be improved when a Web-based diary approach for self-management of asthma was replaced by a multiple-technology approach including collecting data via cell phone, delivering health messages via cell phone, and using a traditional Web page for data display and system customization (Anhoj and Moldrup 2004). In another example, in rural Kenya, a clinic's existing medical record database was linked to data captured by a handheld global positioning system creating digital maps of injury spatial distribution using geography information systems software to demonstrate the value of combining these technological tools for injury surveillance, epidemiologic research, and injury prevention efforts (Odero et al. 2007). Combinations of technologies can drive innovations within advanced health-care settings as well. For example, research has demonstrated that an automated prospective surveillance screening tool was effective in continuously monitoring multiple information sources (laboratory, radiographic, demographic, surgical, etc.) in a large health-care system to improve recognition of patients with acute lung injury who could benefit from protective ventilation (Koenig et al. 2011). In another example, researchers have concluded that expanding the use of vehicle-to-satellite communication technologies for real-time motor vehicle crash surveillance and linking such a surveillance system to traditional emergency medical systems could dramatically improve emergency response times, particularly in rural areas (Brodsky 1993).

There is no doubt that technology will continue to advance rapidly. The challenge to public health professionals and academic researchers is to monitor technological advancements, to be aware of new technologies that may have public health surveillance applications, and to embrace such change and the novel methodological approaches they make possible.

Positives and Negatives

As with all epidemiologic methodologies, those utilizing advanced technologies have both positives and negatives. In order to optimize the positives of specific methodologies, public health professionals and researchers must also recognize the negatives.

Improved data quality coupled with decreased personnel and economic costs is among the most important positives associated with utilizing advanced technologies such as Internet-based data collection tools, tools capable of automated data collection from preexisting records, and cell phones. Such technologies allow large quantities of data to be collected in short periods of time by small numbers of researchers while simultaneously ensuring captured data is of the highest possible quality. For example, once study subjects are identified, a single researcher can utilize an Internet-based data collection tool to conduct injury surveillance in 10 or 10,000 study subjects for the same cost. Improvements in data quality can be made by reducing time burden for reporters and by reducing the opportunity for data reporters or researchers to make errors. For example, automatic validation checks can be incorporated with real-time prompts to alert data reporters to missing, incomplete, or illogical responses as they are entering data, thus resulting in collection of more complete and more accurate data. Similarly, data reporter compliance can be improved by combining automatic compliance checks with e-mail or cell phone reminders and by the automatic application of response-based skip patterns in Internet-based data collection tools which reduce the time burden for data reporters while reducing the number of missing responses. Such capabilities allow researchers using Internet-based surveillance tools to incorporate a greater number of questions while minimizing reporter fatigue. Additionally, because electronically captured data can be automatically transformed into analyzable formats, errors associated with secondary data entry and data coding are eliminated.

Because cell phones have become nearly ubiquitous in developed countries, many cell phone users now have devices with Internet connection capabilities, and most cell phone users carry their phones with them and answer them throughout the day, researchers who utilize cell phones in surveillance efforts have unprecedented access to study populations. Populations in developed countries have become so comfortable with the Internet and cell phones that researchers using these technologies for surveillance have found people, particularly young adults and adolescents, are more willing to complete surveys online, via e-mail, or via cell phone than in person, via surface mail, or via landline telephone. Additionally, this technology generation is often more comfortable reporting dangerous behaviors or answering sensitive questions via technology than face-to-face. These are simply a few examples of the positives associated with using modern technologies for injury surveillance in special populations.

Most of the negatives associated with using advanced technologies center around the potential disconnects between the researchers and study subjects who probably will never interact in person. One of the most concerning realities of the relationship between researcher and study subject being, in most cases, a virtual one, is that it is impossible for researchers to monitor with absolute certainty exactly who is reporting data to the surveillance system. Additionally, it can be difficult to establish the representativeness of a geographically dispersed study sample captured via the Internet. For example, while integrating landline and cellular telephone samples is being encouraged as a way to improve representativeness of study samples, actually achieving representative study samples through such integration can be difficult because cellular phone numbers are assigned to individuals whereas landline phone numbers are assigned to households (Voigt et al. 2011). Thus, researchers may unwittingly enroll multiple study subjects from a single household. Additionally, given cell phone users' practice of maintaining their phone number after moving to a new geographic region, researchers using area codes to identify regional samples will likely enroll subjects no longer living in an area unless they screen for current residence prior to enrollment. Another challenge is engendering and maintaining study subjects' enthusiasm for participation in a surveillance project when their only interaction with the research team may be e-mails or cell phone calls. For example, response rates in studies using truly novel technology, like having cell phone users transmit photos of injuries to researchers, have been reported to be low, indicating the need to more fully investigate methods to motivate and retain study subjects (Walker et al. 2011). An additional concern for researchers using technologies like the Internet and cell phones for surveillance is the study population's access to such devices. Internet coverage, cell phone coverage, battery life and mobile recharging options for mobile devices, etc., are all concerns for researchers using such technologies for surveillance, particularly in special populations in rural areas of developing countries. Another negative of advanced technology is that such tools are so easy to use and so inexpensive that they can be used inappropriately by untrained or inattentive individuals. While technological advancements offer exciting opportunities for new surveillance methodologies, public health practitioners and academic researchers must remember that surveillance projects utilizing modern technologies still need to be sound epidemiologically and must still follow ethical standards (Bull et al. 2011).

Elegant Simplicity

Although the rapid advancement of technology has provided a plethora of novel methodological options, researchers undertaking surveillance in special populations should be encouraged to remember the mantra of elegant simplicity. Often, big impacts can be made with the simplest solutions while complex methodologies can introduce multiple potential opportunities for error. For example, public health professionals worldwide recognize that reliable cause of death data, essential to the development of national and international disease and injury prevention policies, are available for less than

30% of the deaths that occur annually worldwide. To address this deficiency, the Global Burden of Disease Study combined multiple available data sources, making corrections for identifiable miscoding, to estimate both worldwide and regional cause of death patterns by age–sex group and region (Murray and Lopez 1997). Such simple, inexpensive surveillance can provide a foundation for a more informed debate on public health priorities while freeing valuable resources for the development and implementation of prevention efforts. Similarly, while syndromic surveillance systems designed for early detection of outbreaks are typically highly complex, technology-driven, automated tools in developed countries, low-technology applications of syndromic surveillance have been proven to be feasible and effective tools in developing countries (May et al. 2009). Recognizing the methodological resources required to accomplish the goals of a surveillance project in a special population and the cost–benefit ratio of utilizing any more complex methodologies than the minimum required is challenging for researchers eager to take advantage of rapidly advancing technologies.

Statistical Aspects

Computer technology advancements have not only provided new and varied means of capturing injury surveillance data, a secondary byproduct is advancements in statistical methodologies available for analysis of surveillance data. Basic activities such as data cleaning and evaluation of data distributions can now easily be automated. Additionally, surveillance data quality checks can be automated if surveillance systems capture data from electronic sources. In addition to such simple tasks, the power of modern computer technology available to most public health professionals and researchers has allowed for a diversity of surveillance systems ranging from those that capture massive amounts of data to those that study unique populations. The diversity of statistical methodologies available for analysis of surveillance data has similarly grown.

Growing Options

The advances in statistical methodologies have focused primarily on addressing the need to analyze increasingly large datasets in novel ways and on addressing power issues and data distribution issues in special populations. Some of the advancements in analysis of injury surveillance data have resulted from the simple application of statistical methodologies commonly applied in other areas of public health research. For example, multivariate regression analyses and correlation analyses have become commonplace in injury surveillance studies in special populations. Time series analysis has enabled more meaningful evaluation of injury surveillance data. Statistical methods initially developed for epidemiologic investigation of disease have proved effective as they have more frequently been applied in injury epidemiology research. For example, Lorenz-curve analyses were used to calculate cause of death patterns for both disease and injury in the Global Burden of Disease Study (Murray and Lopez 1997). Additionally, researchers studying injuries in occupational cohorts identified the need for the development of innovative statistical techniques to account for recurrent injuries to workers over time and the temporary removal of workers from the occupational sample while recuperating from injury or during times of illness (Wassell et al. 1999). They found that subject-specific random effects and multiple event times could be addressed by applying frailty models that characterizes the dependence of recurrent events over time and proportional hazards regression models could be used to estimate the effects of covariates for subjects with discontinuous intervals of risk. Problems achieving statistical power in studies of special populations due to such issues as population heterogeneity and small sample sizes can lead to difficulties in identifying risk

factors and demonstrating efficacy of interventions. However, such problems sometimes can be addressed through appropriate statistical methodologies. For example, researchers investigating the application of both conventional and innovative methods for the analysis of randomized controlled trials in traumatic brain injury (TBI) populations have demonstrated that statistical power can be considerably increased by applying covariate adjustments as well as by conducting ordinal analyses such as proportional odds and sliding dichotomy (Maas and Lingsma 2008). Other approaches include innovative combinations of methodologies. Researchers found that combining disease mapping and regression methods was relatively efficient for analyses in special populations such as individuals with iatrogenic injury (MacNab et al. 2006). In this case, Bayesian statistical methodologies made it possible to study associations between injury and risk factors at the aggregate level while accounting for unmeasured confounding and spatial relationships. More specifically, a unified Bayesian hierarchical spatial modeling framework (the joint use of empirical Bayes and full Bayesian inferential techniques) enabled simultaneous examinations of potential associations between iatrogenic injury and regional characteristics, age effects, residual variation, and spatial autocorrelation. Such combined approaches can draw on the strengths of each method while minimizing the weaknesses of each.

Positives and Negatives

Both the positives and negatives of applying advanced statistical methodologies to the analysis of injury surveillance data in special populations lie in matching the most appropriate methodology to the public health problem. While researchers must be encouraged to utilize advanced statistical methodologies to fully recognize the value of the data captured by surveillance systems, they must refrain from applying exotic methodologies merely because they can, as doing so can lead to confusion or distrust among special populations and can even lead to misinterpretation of or misapplication of surveillance data for special populations.

Elegant Simplicity

The importance of utilizing advanced statistical methodologies pales in comparison to the importance of capturing the most applicable, complete, and useful data during surveillance of special populations. Advanced statistical methodologies falter when they are too complex to explain to the special population under study. Only a thorough understanding of injury patterns and risk factors can drive the development of effective preventive interventions. If members of a special population do not understand or do not “trust” the data driving an intervention, they may not be willing to adopt the intervention.

Special Considerations

Forging and Maintaining Strong Ties with Community Stakeholders

When conducting surveillance in a special population, it is important to identify community leaders and stakeholders as access to the population may depend upon their approval. Additionally, while these individuals may or may not be under surveillance themselves, they can provide crucial insight

into the culture of the population which will assist in the development of the surveillance methodology. For example, such individuals may offer insight into which data collection technologies would be best accepted by the population, which individuals might make the best data reporters, etc. Special populations may have some reservations about participating in surveillance studies due to a lack of experience with, a lack of understanding of, or a lack of comfort with public health epidemiology. Even those special populations with an eagerness to participate in surveillance efforts will likely require a thorough explanation of the purpose of the surveillance study, their role in the study, what will be expected of them as participants, and the possible outcomes that may result from interpretation of the data collected during the surveillance study. To maintain enthusiasm for the surveillance project and to soothe any lingering concerns held by the special population, researchers should maintain communication with the special population's community leaders and stakeholders throughout the surveillance project, providing updates as they are available and responding to any problems that may arise promptly. Thus, clear communication between those conducting injury surveillance in special populations and the special population's community leaders and stakeholders is paramount.

Gain Knowledge of the Special Population

The success of injury surveillance projects in special populations is dependent upon the researchers' knowledge of the special population. Simply knowing which variables should be captured by a surveillance system requires a thorough understanding of the special population. For example, researcher evaluating laser radiation exposures found that databases compiled by existing laser incident surveillance systems did not provide sufficient information to enable a thorough evaluation of laser exposure incidents or tracking of trends over time (Clark et al. 2006). Using the Delphi technique, expert panels of health and safety professionals experienced with laser systems and medical evaluation of laser injuries were surveyed and the knowledge gained was used to develop a novel surveillance system that captures 100 data fields identifying the most valuable items for injury and injury trend analysis. By gaining a better understanding of the needs of the special population, researchers were able to dramatically improve surveillance methodology just by improving the data fields captured. Similarly, while injuries have a substantial effect on the health and quality of life in both developed and developing countries, it is important to understand that although injury surveillance is needed to inform policy makers and direct public health efforts worldwide, knowledge of regional differences should drive decisions regarding the most appropriate surveillance methodology. For example, the scarcity of resources in developing countries means there is limited preexisting data available and few injury surveillance systems currently in place (Graitcer 1992). However, researchers with knowledge of the effect of financial constraints on injury surveillance in developing countries have been able to develop innovative injury surveillance methods using easy to use, low-cost Social Web and GeoWeb technologies (Cinnamon 2010). Establishing close relationships with community stakeholders and leaders should entail two-way communication with the researcher, so they can learn as much as possible about the special population in order to best serve their needs in addressing the injury issue under study.

Understand and Acknowledge Culture

Even the most methodologically sound surveillance project may fail if it is not acceptable to the special population of interest. Thus, the most appropriate injury surveillance approach for any specific surveillance project in a special population will depend upon a full understanding of the culture of the special population and the conceptual framework within which the researcher and the special

population are approaching the injury issue. For example, a study of farm injuries among Old Order Anabaptist communities concluded that injury patterns in the community reflected the fact that their agricultural practices remain largely nonmotorized, instead depending primarily upon mules and horses (Jones and Field 2002). As the researchers concluded, however, it would not be appropriate to apply recommendations for injury prevention measures based on the current body of knowledge in agricultural safety to this special population because Old Order Anabaptist choices concerning farm safety issues are directly related to their socio-religious beliefs. This is an excellent reminder that, to be effective, injury prevention efforts resulting from injury surveillance must be sensitive to the culture of the special population.

Public Health Importance

Public health professionals and academic researchers must never forget that the goal of injury surveillance in special populations is to collect the data necessary to drive development of effective injury prevention efforts. Special populations may be particularly unfamiliar with public health epidemiology and thus may be leery of being “used” by researchers. Researchers must combat this by providing the special population under study with tangible and timely products of the surveillance efforts. These can range from simple summary reports interpreting analysis of the surveillance data to the implementation of prevention efforts developed in response to knowledge gained via surveillance data. The key is to provide the special population with evidence that their participation in the surveillance project was meaningful. For example, researchers conducting epidemiologic surveillance in Peace Corps volunteers working in developing countries recognized that although the surveillance system was established to provide the data needed to plan, implement, and evaluate health programs and to monitor health trends in that special population, it could also provide a model for surveillance in other groups of temporary and permanent residents of developing countries (Bernard et al. 1989). Thus, this surveillance system not only directly benefited Peace Corps volunteers but also benefited the very populations Peace Corps volunteers work to help. Demonstrating that surveillance efforts will provide tangible and timely benefit to the special population under study is not only the right thing to do but it will also improve the relationship with the special population and thus, the potential for the success of the surveillance effort.

Injury surveillance studies in special populations should never be mere academic exercises whose impacts reach no further than an article in a peer-review journal. Langmuir’s third tenet of public health surveillance was “the prompt dissemination of results to those who need to know.” This emphasizes the expectation that surveillance efforts should not only advance the body of scientific knowledge but should also directly benefit the population under surveillance.

Conclusion

Innovation is a common characteristic of successful injury surveillance projects. However, as noted in a study of adverse events in trauma surgery, even after the epidemiology of injury in a special population is fully understood, there may be a further need for innovation in the development of prevention efforts (Clarke et al. 2008). Thus, public health professionals and academic researchers must recognize that innovation in surveillance methodology is only the first step; effective application of surveillance data to drive positive change in the special populations under surveillance is the real goal. As Thacker (2000) so eloquently stated, “The critical challenge in public health surveillance today, however, continues to be the assurance of its usefulness.”

References

- Ahmed, O. H., Sullivan, S. J., Schneiders, A. G., & McCrory, P. (2010). iSupport: Do social networking sites have a role to play in concussion awareness? *Disability and Rehabilitation*, *32*(22), 1877–1883.
- Anhoj, J., & Moldrup, C. (2004). Feasibility of collecting diary data from asthma patients through mobile phones and SMS (short message service): Response rate analysis and focus group evaluation from a pilot study. *Journal of Medical Internet Research*, *6*(4), e42.
- Bain, T. M., Frierson, G. M., Trudelle-Jackson, E., & Morrow, J. R. (2010). Internet reporting of weekly physical activity behaviors: The WIN Study. *Journal of Physical Activity & Health*, *7*(4), 527–532.
- Bech, M., & Kristensen, M. B. (2009). Differential response rates in postal and web-based surveys among older respondents. *Survey Research Methods*, *3*(1), 1–6.
- Bennett, G. G., & Glasgow, R. E. (2009). The delivery of public health interventions via the Internet: Actualizing their potential. *Annual Review of Public Health*, *30*, 273–292.
- Bernard, K. W., Graitcer, P. L., van der Vlugt, T., Moran, J. S., & Pulley, K. M. (1989). Epidemiological surveillance in Peace Corps Volunteers: A model for monitoring health in temporary residents of developing countries. *International Journal of Epidemiology*, *18*(1), 220–226.
- Brodsky, H. (1993). The call for help after an injury road accident. *Accident; Analysis and Prevention*, *25*(2), 123–130.
- Bull, S. S., Breslin, L. T., Wright, E. E., Black, S. R., Levine, D., & Santelli, J.S. (2011). Case Study: An ethics case study of HIV prevention research on Facebook: The Just/Us Study. *Journal of Pediatric Psychology*, *36*(10), 1082–1092.
- Centers for Disease Control and Prevention. (2010). Launching a National Surveillance System after an earthquake – Haiti, 2010. *MMWR*, *59*(30), 933–938.
- Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One*, *5*(11), e14118.
- Chou, W. Y., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the United States: Implications for health communication. *Journal of Medical Internet Research*, *11*(4), e48.
- Cinnamon, J. & Schuurman, N. (2010). Injury surveillance in low-resource settings using Geospatial and Social Web technologies. *International Journal of Health Geographics*, *9*, 25.
- Clark, K. R., Neal, T. A., & Johnson, T. E. (2006). Creation of an innovative laser incident reporting form for improved trend analysis using Delphi technique. *Military Medicine*, *171*(9), 894–899.
- Clarke, D. L., Gouveia, J., Thomson, S. R., & Muckart, D. J. (2008). Applying modern error theory to the problem of missed injuries in trauma. *World Journal of Surgery*, *32*(6), 1176–1182.
- Clough, J. F., Zirkle, L. G., & Schmitt, R. J. (2010). The role of SIGN in the development of a global orthopaedic trauma database. *Clinical Orthopaedics and Related Research*, *468*(10), 2592–2597.
- Comstock, R. D., Knox, C., Yard, E., & Gilchrist, J. (2006). Sports-related injuries among high school athletes – United States, 2005–06 school year. *MMWR*, *55*(38), 1037–1040.
- Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Using Web and social media for influenza surveillance. *Advances in Experimental Medicine and Biology*, *680*, 559–564.
- De Vera, M. A., Ratzlaff, C., Doerfling, P., & Kopec, J. (2010). Reliability and validity of an internet-based questionnaire measuring lifetime physical activity. *American Journal of Epidemiology*, *172*(10), 1190–1198.
- Eysenbach, G. (2006). Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. *AMIA Annual Symposium Proceedings, 2006*, 244–248.
- Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research*, *11*(1), e11.
- Fan, S., Blair, C., Brown, A., Gabos, S., Honish, L., Hughes, T., Jaipaul, J., Johnson, M., Lo, E., Lubchenko, A., Mashinter, L., Meurer, D. P., Nardelli, V., Predy, G., Shewchuk, L., Sosin, D., Wicentowich, B., & Talbot, J. (2010). A multi-function public health surveillance system and the lessons learned in its development: The Alberta Real Time Syndromic Surveillance Net. *Canadian Journal of Public Health*, *101*(6), 454–458.
- Fernandez-Luque, L., Karlsen, R., Krogstad, T., Burkow, T. M., & Vognild, L. K. (2010). Personalized health applications in the Web 2.0: The emergence of a new approach. *Conference Proceedings: IEEE Engineering in Medicine and Biology Society, 2010*, 1053–1056.
- Graitcer, P. L. (1987). The development of state and local injury surveillance systems. *Journal of Safety Research*, *18*(4), 191–198.
- Graitcer, P. L. (1992). Injury surveillance in developing countries. *MMWR*, *41*(1), 15–20.
- Graitcer, P. L., & Burton, A. H. (1986). The Epidemiologic Surveillance Project: Report of the pilot phase. *American Journal of Preventive Medicine*, *76*, 1289–1292.
- Graitcer, P. L., & Thacker, S. B. (1986). The French connection. *AJPH*, *76*(11), 1285–1286.
- Holder, Y., Peden, M., Krug, E., Lund, J., Gururaj, G., & Kobusingye, O. (2001). *Injury surveillance guidelines*. Geneva, Switzerland in conjunction with the Centers for Disease Control and Prevention, Atlanta, GA: World Health Organization.

- Horan, J. M., & Mallonee, S. (2003). Injury surveillance. *Epidemiologic Reviews*, 25, 24–42.
- Jones, P. J., & Field, W. E. (2002). Farm safety issues in Old Order Anabaptist communities: Unique aspects and innovative intervention strategies. *Journal of Agricultural Safety and Health*, 8(1), 67–81.
- Koenig, H. C., Finkel, B. B., Khalsa, S. S., Lanken, P. N., Prasad, M., Urbani, R., & Fuchs, B. D. (2011). Performance of an automated electronic acute lung injury screening system in intensive care unit patients. *Critical Care Medicine*, 39(1), 98–104.
- Kollmann, A., Riedi, M., Kastner, P., Schreier, G., & Ludvik, B. (2007). Feasibility of a mobile phone-based data service for functional insulin treatment of type 1 diabetes mellitus patients. *Journal of Medical Internet Research*, 9(5), e36.
- Koo, D., & Thacker, S. B. (2010). In Snow's footsteps: Commentary on shoe-leather and applied epidemiology. *American Journal of Epidemiology*, 172(6), 737–739.
- Kypri, K., Gallagher, S. J., & Cashll-Smith, M. L. (2004). An internet-based survey method for college student drinking research. *Drug and Alcohol Dependence*, 76(1), 45–53.
- Langmuir, A. D. (1963). Surveillance of communicable diseases of national importance. *The New England Journal of Medicine*, 268, 182–192.
- Lord, S., Brevard, J., & Budman, S. (2011). Connecting to young adults: An online social network survey of beliefs and attitudes associated with prescription opioid misuse among college students. *Substance Use & Misuse*, 46(1), 66–76.
- Maas, A. I., & Lingsma, H. F. (2008). New approaches to increase statistical power in TBI trials: Insights from the IMPACT study. *Acta Neurochirurgica Supplementum*, 101, 119–124.
- MacNab, Y. C., Kmetc, A., Gustafson, P., & Sheps, S. (2006). An innovative application of Bayesian disease mapping methods to patient safety research: A Canadian adverse medical event study. *Statistics in Medicine*, 25(23), 3960–3980.
- Manring, M. M., Hawk, A., Calhoun, J. H., & Andersen, R. C. (2009). Treatment of war wounds: A historical review. *Clinical Orthopaedics and Related Research*, 467(8), 2168–2219.
- May, L., Chretien, J. P., & Pavlin, J. A. (2009). Beyond traditional surveillance: Applying syndromic surveillance to developing settings – opportunities and challenges. *BMC Public Health*, 16(9), 242.
- McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of Affective Disorders*, 122(3), 277–279.
- Meigs, J. W. (1948). Illness and injury rates in small industrial plants; a study in factor epidemiology. *Occupational Medicine*, 5(1), 11–23.
- Murray, C. J., & Lopez, A. D. (1997). Mortality by cause for eight regions of the world: Global Burden of Disease Study. *The Lancet*, 349(9061), 1269–1276.
- Odero, W., Rotich, J., Yiannoutsos, C. T., Ouna, T., & Tierney, W. M. (2007). Innovative approaches to application of information technology in disease surveillance and prevention in Western Kenya. *Journal of Biomedical Informatics*, 40(4), 390–397.
- Pagliari, C., Sloan, D., Gregor, P., Sullivan, F., Detmer, D., Kahan, J. P., Oortwijn, W., & MacGillivray, S. (2005). What is eHealth (4): A scoping exercise to map the field. *Journal of Medical Internet Research*, 7(1), e9.
- Patillo, R. (2010). Are you using Twitter for your next survey? *Nurse Educator*, 35(5), 207.
- Paulozzi, L. J., Mercy, J., Frazier, L., & Annest, J. L. (2004). CDC's National Violent Death Reporting System: Background and methodology. *Injury Prevention*, 10(1), 47–52.
- Retsas, S. (2009). Alexander's (356–323 BC) expeditionary Medical Corps 334–323 BC. *Journal of Medical Biography*, 17(3), 165–169.
- Russell, C. W., Boggs, D. A., Palmer, J. R., & Rosenberg, L. (2010). Use of a web-based questionnaire in the Black Women's Health Study. *American Journal of Epidemiology*, 172(11), 1286–1291.
- Salazar, C. F. (1998). Medical care for the wounded in armies of ancient Greece (article in German). *Sudhoffs Archiv*, 82(1), 92–97.
- Thacker, S. B. (2000). Historical development. In S. M. Teutsch & R. E. Churchill (Eds.), *Principles and practice of public health surveillance* (2nd ed.). New York: Oxford University Press.
- van Gelder, M. M., Bretveld, R. W., & Roeleveld, N. (2010). Web-based questionnaires: The future in epidemiology? *American Journal of Epidemiology*, 172(11), 1292–1298.
- Voigt, L. F., Schwartz, S. M., Doody, D. R., Lee, S. C., & Li, C. I. (2011). Feasibility of including cellular telephone numbers in random digit dialing for epidemiologic case-control studies. *American Journal of Epidemiology*, 173(1), 118–126.
- Walker, T. W., O'Conner, N., Byrne, S., McCann, P. J., & Kerin, M. J. (2011). Electronic follow-up of facial lacerations in the emergency department. *Journal of Telemedicine and Telecare*, 17(3), 133–136.
- Wassell, J. T., Wojciechowski, W. C., & Landen, D. D. (1999). Recurrent injury event-time analysis. *Statistics in Medicine*, 18(23), 3355–3363.
- Williams, R. E., & Capel, E. H. (1945). The incidence of sepsis in industrial wounds. *British Journal of Industrial Medicine*, 2, 217–220.
- Yi, Q., Hoskins, R. E., Hillringhouse, E. A., Sorensen, S. S., Oberle, M. W., Fuller, S. S., & Wallace, J. C. (2008). Integrating open-source technologies to build low-cost information systems for improved access to public health data. *International Journal of Health Geographics*, 7, 29.

Chapter 4

Surveillance of Traumatic Brain Injury

Jean A. Langlois Orman, Anbesaw W. Selassie, Christopher L. Perdue,
David J. Thurman, and Jess F. Kraus*

Traumatic Brain Injury (TBI) Surveillance in Civilian Populations

Clinical Case Definitions

Clinical case definitions describe the criteria for diagnosing TBI and provide an important background for evaluating epidemiologic case definitions. Two clinical indicators, the *occurrence* of impairment of consciousness [also referred to as alteration of consciousness (AOC), including loss of consciousness (LOC)] and post-traumatic amnesia (PTA), are the indicators most commonly used to assess acute brain injury severity and thus figure prominently in TBI clinical case definitions. The Glasgow Coma Scale (GCS) is the most widely used tool for assessing impaired consciousness (Teasdale and Jennett 1974) (Table 4.1).

Disclaimer

*The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the Department of the Army, the Department of Defense, or the Centers for Disease Control and Prevention.

J.A.L. Orman, ScD, MPH (✉)
Statistics and Epidemiology, US Army Institute of Surgical Research,
3698 Chambers Pass Bldg 3611, ATTN MCMR-SRR, Fort Sam Houston, TX 78234-6315, USA
e-mail: jean.a.orman@amedd.army.mil

A.W. Selassie, DrPH
Department of Biostatistics, Bioinformatics and Epidemiology,
Medical University of South Carolina, 135 Cannon Street, Charleston, SC 29425, USA
e-mail: selassie@musc.edu

C.L. Perdue, MD, MPH
Armed Forces Health Surveillance Center, 11800 Tech Road, Suite 220, Silver Spring, MD 20904, USA
e-mail: christopher.perdue@amedd.army.mil

D.J. Thurman, MD, MPH
National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease
Control and Prevention, 4770 Buford Highway, Mailstop K-51, Atlanta, GA 30341, USA
e-mail: dxt9@cdc.gov

J.F. Kraus, MPH, PhD
Department of Epidemiology, University of California at Los Angeles,
Los Angeles, CA, USA
e-mail: jkraus3637@roadrunner.com

Table 4.1 Glasgow Coma Scale

Type of response		Score
Eye opening	Spontaneous	4
	To speech	3
	To pain	2
	None	1
Motor	Obeys commands	6
	Localizes pain	5
	Withdrawal	4
	Abnormal flexion	3
	Extension	2
	No response	1
	Verbal	Oriented
	Confused	4
	Inappropriate	3
	Incomprehensible	2
	No response	1
Total ^a		

Source: adapted from (Teasdale and Jennett 1974)

^aTotal is the sum of the highest score from each category (range 3–15) (maximum = 15); higher score = less severe injury

Table 4.2 Severity of brain injury stratification

Criteria	Mild/concussion	Moderate	Severe
Structural imaging	Normal ^a	Normal or abnormal	Normal or abnormal
Abbreviated injury scale (AIS) anatomical/structural injury	1–2	3	4–6
Loss of consciousness (LOC)	0–30 min	>30 min and <24 h	>24 h
Alteration of consciousness/mental state (AOC)	A moment up to 24 h	>24 h; severity based on other criteria	
Post-traumatic amnesia (PTA)	≤1 day	>1 and <7 days	>7 days
Glasgow Coma Scale (best available score in first 24 h) ^b	13–15	9–12	3–8

Source: adapted from VA/DoD (Clinical Practice Guideline 2009)

^aNote that minor abnormalities possibly not related to the brain injury may be present on structural imaging in the absence of LOC, AOC, and PTA

^bSome studies report the best available GCS score within the first 6 h or some other time period

PTA, also referred to as anterograde amnesia, is defined as a period of hours, weeks, days, or months after the injury when the person exhibits a loss of day-to-day memory. TBI can be categorized as mild, moderate, or severe based on the *length* of impaired consciousness, LOC, or PTA. Criteria for determining acute severity are summarized in Table 4.2. Acute injury severity is best determined at the time of the injury (VA/DoD 2009).

Another commonly used method of assessing TBI severity is the Abbreviated Injury Scale (AIS) (AAAM 1990). This measure relies on anatomic descriptors of the injury sustained and the immediate consequences such as LOC and degree of cerebral hemorrhage. The most appropriate method of scoring AIS is manual assignment of the seven-digit codes by trained coders. Trauma centers in the USA use the AIS to grade the severity of injuries in their trauma registries. Unlike physiological measures of severity such as GCS that are best performed within minutes after TBI, AIS can be assigned after the patient has been stabilized. The AIS score for the head only is used to describe the severity of TBI (see Table 4.2).

In 1995, the US Centers for Disease Control and Prevention (CDC) published *Guidelines for Surveillance of Central Nervous System Injury* (Thurman et al. 1995a), one of the first systematic

efforts to develop a standard TBI case definition. They defined TBI as craniocerebral trauma, specifically, “an occurrence of injury to the head (arising from blunt or penetrating trauma or from acceleration/deceleration forces) that is associated with any of these symptoms attributable to the injury: decreased level of consciousness, amnesia, other neurologic or neuropsychological abnormalities, skull fracture, diagnosed intracranial lesions, or death.” Additional considerations in defining and diagnosing TBI based on more recent research have been summarized in Saatman et al. (2008) and Menon et al. (2010).

Because of increased recognition of concussion or mild TBI as a specific clinical entity, separate definitions have been developed to diagnose this subgroup of persons with TBI. Although the terms concussion and mild TBI have been used interchangeably, “concussion” is preferred because it refers to a specific injury event that may or may not be associated with persisting symptoms. Therefore, although both of these terms are used in the literature cited here, the term “concussion/mTBI” is used in the remainder of this chapter.

In the USA, the most widely accepted clinical criteria for concussion/mTBI are those proposed by the American College of Rehabilitation Medicine (ACRM 1993) as follows:

A traumatically induced physiological disruption of brain function, as manifested by *at least one of* the following:

- Any loss of consciousness
- Any loss of memory for events immediately before or after the accident
- Any alteration in mental state at the time of the accident (injury) (e.g., feeling dazed, disoriented, or confused); focal neurological deficit(s) that may or may not be transient

But where the severity of the injury does not exceed the following:

- Loss of consciousness of approximately 30 minutes or less
- After 30 minutes, an initial Glasgow Coma Scale score of 13–15
- Post-traumatic amnesia (PTA) not greater than 24 hours

Criteria for concussion/mTBI used by other groups include the CDC (National Center for Injury Prevention and Control 2003) and the World Health Organization (WHO) (Carroll et al. 2004) definitions. In summary, most experts agree that the common criteria for concussion/mTBI include an initial GCS score of 13–15 or only a brief LOC, brief PTA, and normal structural findings of neuroimaging studies [e.g., head computed tomography (CT)]. (VA/DoD 2009) (Table 4.2).

Case Definitions for Administrative Data Systems

The standard TBI case definition developed by the CDC is among the most widely used for surveillance in which cases are identified using International Classification of Diseases (ICD) diagnosis codes (Marr and Coronado 2004) (Table 4.3). This definition has some limitations. First, although included in the definition as an indicator of TBI, skull fracture by itself is not necessarily a brain injury per se.¹ Second, to avoid underestimating TBIs, the code 959.01, “head injury, unspecified,” is included because its introduction to ICD-9-CM (Department of Health

¹ However, a strong relationship between cranial and intracranial injury has long been recognized, with skull fracture taken as an indicator that the brain has been exposed to injurious forces. For that reason, the term “craniocerebral trauma” is still retained as a synonym for TBI (Thurman et al. 1995a; Ropper and Samuels 2009). It should be noted also that current accepted indications for radiologic imaging studies of head trauma patients are directed principally to those who already meet clinical criteria for TBI or concussion/mTBI (Jagoda et al. 2008). Therefore, the likelihood of diagnosing skull fractures in the absence of clinical TBI or mTBI appears low and probably of small effect in epidemiologic estimates of TBI incidence in general populations.

Table 4.3 CDC TBI case definition for use with data systems

TBI morbidity (ICD-9-CM codes)	
800.0–801.9	Fracture of the vault or base of the skull
803.0–804.9	Other and unqualified and multiple fractures of the skull
850.0–854.1	Intracranial injury, including concussion, contusion, laceration, and hemorrhage
950.1–950.2	Injury to the optic chiasm, optic pathways, and visual cortex
959.01	Head injury, unspecified (beginning 10/1/97)
995.55	Shaken infant syndrome
TBI mortality (ICD-10 codes)	
S01.0–S01.9	Open wound of the head
S02.0, S02.1, S02.3, S02.7–S02.9	Fracture of skull and facial bones
S04.0	Injury to optic nerve and pathways
S06.0–S06.9	Intracranial injury
S07.0, S07.1, S07.8, S07.9	Crushing injury of head
S09.7–S09.9	Other and unspecified injuries of head
T01.0	Open wounds involving head with neck
T02.0	Fractures involving head with neck
T04.0	Crushing injuries involving head with neck
T06.0	Injuries of brain and cranial nerve with injuries of nerves and spinal cord at neck level
T90.1, T90.2, T90.4, T90.5, T90.8, T90.9	Sequelae of injuries of head
	Note: according to the CDC, these codes should be considered provisional until sensitivity and predictive value are evaluated

Source: (Marr and Coronado 2004)

and Human Services 1989) in the 1997 annual update resulted in a rise in its use and a corresponding drop in the use of the code 854, “intracranial injury of other and unspecified nature” (Faul et al. 2010). Some of the cases included using this definition may be head injuries (e.g., injuries to the scalp), but not brain injuries, and thus may not meet the clinical criteria for TBI. In the USA, ICD-10 codes (WHO 2007) are used for identifying TBI-related deaths, and ICD-9-CM codes (Department of Health and Human Services 1989) for hospitalizations, emergency department (ED) visits, and outpatient visits, until such time as ICD-10-CM is implemented. In anticipation of the change to ICD-10-CM, the CDC has also released a proposed surveillance case definition using the new codes (Table 4.4).

In an effort to facilitate surveillance of concussion/mTBIs, the CDC developed a proposed ICD-9-CM code-based definition for mild TBI designed to be used with data for persons treated in health-care facilities (National Center for Injury Prevention and Control 2003) (Table 4.5). Bazarian et al. (2006) conducted a prospective cohort study of patients presenting to an ED and compared real-time clinical assessment of mild TBI with the ICD-9-CM codes for this definition assigned after ED or hospital discharge. They found that the sensitivity and specificity of these codes for identifying concussion/mTBIs were 45.9 and 97.8%, respectively, suggesting that estimates based on these codes should be interpreted with caution.

Of note, CDC periodically updates the TBI surveillance case definitions; thus, a more recent version may be in use.

Administrative Data Sources

Quantitative data for population-based assessment of injuries, including TBI, are available from several sources in most high-income countries, including the USA. Many of the data sets that are easy to obtain were designed for other administrative purposes, for example, hospital billing, and thus

Table 4.4 Proposed CDC ICD-10-CM case definition for traumatic brain injury

S01.0	Open wound of scalp	S07.0	Crushing injury of face
S01.1	Open wound of eyelid and periocular area ^a	S07.1	Crushing injury of skull
		S07.8	Crushing injury of other parts of head ^a
S01.2	Open wound of nose ^a	S07.9	Crushing injury of head, part unspecified ^a
S01.3	Open wound of ear ^a		
S01.4	Open wound of cheek and temporomandibular area ^a	S09.7	Multiple injuries of head
		S09.8	Other specified injuries of head
S01.5	Open wound of lip and oral cavity ^a	S09.9	Unspecified injury of head
S01.7	Multiple open wounds of head		
S01.8	Open wound of other parts of head	T01.0	Open wounds involving head with neck
S01.9	Open wound of head, part unspecified	T02.0	Fractures involving head with neck ^a
		T04.0	Crushing injuries involving head with neck ^a
S02.0	Fracture of vault of skull	T06.0	Injuries of brain and cranial nerves with injuries of nerves and spinal cord at neck level
S02.1	Fracture of base of skull		
S02.3	Fracture of orbital floor ^a		
S02.7	Multiple fractures involving skull and facial bones	T90.1	Sequelae of open wound of head
S02.8	Fracture of other skull and facial bones	T90.2	Sequelae of fracture of skull and facial bones
S02.9	Fracture of skull and facial bones, part unspecified	T90.4	Sequelae of injury of eye and orbit ^a
		T90.5	Sequelae of intracranial injury
S04.0	Injury of optic nerves and pathways	T90.8	Sequelae of other specified injuries of head
		T90.9	Sequelae of unspecified injury of head
S06.0	Concussion		
S06.1	Traumatic cerebral edema		
S06.2	Diffuse brain injury		
S06.3	Focal brain injury		
S06.4	Epidural hemorrhage (traumatic extradural hemorrhage)		
S06.5	Traumatic subdural hemorrhage		
S06.6	Traumatic subarachnoid hemorrhage		
S06.7	Intracranial injury with prolonged coma		
S06.8	Other intracranial injuries		
S06.9	Intracranial injury, unspecified		

Source: (Marr and Coronado 2004)

^aThe CDC recommends including these codes on a provisional basis until sensitivity and positive predictive value are evaluated

Table 4.5 Administrative concussion/mTBI data definition for surveillance or research (ICD-9-CM)

ICD-9-CM first four digits	ICD-9-CM fifth digit
800.0, 800.5, 801.0, 801.5, 803.0, 803.5, 804.0, 804.5, 850.0, 850.1, 850.5 or 850.9	0, 1, 2, 6, 9, or missing
854.0	0, 1, 2, 6, 9, or missing
959.0 ^a	1

Source: (National Center for Injury Prevention and Control 2003)

^aThe current inclusion of code 959.01 (i.e., head injury, unspecified) in this definition is provisional. Although a recent clarification in the definition of this code is intended to exclude concussions, there is evidence that nosologists have been using it to code TBIs. Accordingly, this code may be removed from the recommended definition of mild TBI when there is evidence that in common practice, nosologists no longer assign this code for TBI

have limited information concerning the causes and clinical characteristics of TBI cases. Sometimes linkage with other data sources, for example, with data abstracted separately from medical records, can be used to enhance the information they contain. Because they are among the most useful for epidemiologic research, population-based data sources are the primary focus of this section. Unless otherwise specified, TBI cases are identified from these data sources using ICD codes.

Mortality

In the USA, *National Vital Statistics System (NVSS)* mortality data [also referred to as *Multiple Cause of Death Data (MCDD)*] consist of death certificate data from all US states and territories and are collected by the National Center for Health Statistics (NCHS) (NCHS 2011). Similar mortality data are collected in other high-income and most middle- and low-income countries based on death certificates that are generally consistent with the WHO standards (WHO 1979). The compiled data are coded according to the International Classification of Diseases (WHO 2011). Because TBI, if present on the death certificate, is listed in Part I in the sequence of conditions leading to death and not as the underlying cause (which is always the external cause code, or E code), deaths involving TBI are most accurately reported as *TBI-related* deaths. An important limitation in using MCDD to identify TBI-related deaths is the fact that the conditions listed in the sequence leading to death, such as TBI, are manually coded from the death certificates. The reliability of these codes is therefore dependent upon the accuracy and completeness of the information listed, which may vary depending on who completes the certificate. In the USA, death certificates can be completed either by coroners (publicly elected officials) or medical examiners (forensic pathologists). Death certificates completed by medical examiners have a high level of accuracy (Hanzlick and Combs 1998). An example of a study that used NVSS data is Adekoya et al. (2002) in which trends in TBI-related death rates in the USA were reported.

Morbidity

Hospital Discharge Data

The National Hospital Discharge Survey (NHDS), another annual survey conducted by NCHS (NCHS 2011), includes patient discharges from a nationally representative sample of nonfederal hospitals. The NHDS provides information on principal discharge diagnosis and up to six secondary diagnoses, demographics, length of stay, and payer information. In 2010, additional secondary discharge diagnoses were added, allowing for up to fourteen. For complete ascertainment of TBI cases, it is important to search for the diagnosis in both the primary and secondary diagnosis fields. Beginning in 2011, the NHDS will be incorporated into the National Hospital Care Survey which will include all Uniform Billing form (UB-04) data on inpatient discharges from sampled hospitals. Examples of the use of NHDS data are two CDC reports (Langlois et al. 2004; Faul et al. 2010) in which NHDS data were combined with mortality and ED data to calculate estimates of the incidence of TBI in the USA.

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (H-CUP) sponsored by the Agency for Healthcare Research and Quality (AHRQ) is a nationally representative cluster sample of discharges from nonfederal, short-term general and other specialty hospitals, excluding hospital units of institutions (AHRQ 2011a). When compared with TBI hospitalization rates for the USA calculated using the NHDS, the rates calculated using the NIS tend to be somewhat lower. The NIS data set was used to calculate TBI-related hospital admission rates in an AHRQ report (Russo and Steiner 2007).

State-based hospital discharge data (HDD) are available in some states that create hospital discharge data sets from their hospital care claims data. These standardized data are coded according to the Uniform Billing form (UB-92) promulgated in 1992 by the US Health Care Financing Administration [now the Center for Medicare and Medicaid Services (CMS)]. The Uniform Billing form has been updated to UB-04 as of 2007 (CMS 2010). Among states that require all hospitals within their jurisdiction to report these data, HDD sets can be used to calculate reliable estimates of the number of TBI-related hospitalizations. Using state HDD collected as part of CDC's statewide TBI surveillance initiative, some reports have presented individual state data (Hubbard 2010) or combined data from several states (Eisele et al. 2006; Langlois et al. 2003). State-based HDD for many states are also represented in the HCUP State Inpatient Databases (SID) (AHRQ 2011b). According to the AHRQ, combined SID data for all available states encompass about 90% of all US community hospital discharges. SID data have been used to compare TBI hospitalization rates across states with differing helmet laws (Weiss et al. 2010; Coben et al. 2007).

Emergency Department Data

The *National Hospital Ambulatory Medical Care Survey (NHAMCS)*, also from NCHS, includes a sample of visits to a nationally representative sample of emergency and outpatient departments of nonfederal, noninstitutional (e.g., excluding prison hospitals) general and short-stay hospitals (NCHS 2011). Beginning in 2013, NHAMCS will be incorporated into the National Hospital Care Survey. This new survey will have the potential to link emergency and outpatient department visits with hospital discharge data. Schootman and Fuortes (2000) used NHAMCS data in their study of ambulatory care for TBI in the USA. Some states maintain and analyze their own aggregate state-wide ED visit data sets, for example, South Carolina (Saunders et al. 2009).

The *National Electronic Injury Surveillance System-All Injury Program (NEISS-AIP)* is an expansion of the Consumer Product Safety Commission's (CPSC) National Electronic Injury Surveillance System (NEISS) used to monitor consumer-product-related injuries (CDC 2001). NEISS-AIP includes nonfatal injuries and poisonings treated in US hospital EDs, including those that are not associated with consumer products. The NEISS-AIP uses a subsample of the EDs included in NEISS for its data collection. The NEISS-AIP coding system does not use ICD codes but rather has a fixed number of categories relevant to consumer-product-related injuries for the primary part of the body affected and for the principal diagnosis. Some limitations in TBI case ascertainment using NEISS have been reported (Xiang et al. 2007). Bakhos et al. (2010) used NEISS and NEISS-AIP data to study ED visits for concussion in young child athletes, and the CDC (2007) used NEISS-AIP to investigate nonfatal TBIs from sports and recreation activities in the US population.

Ambulatory Medical Care

The *NCHS Ambulatory Medical Care Survey (NAMCS)*, another annual survey, provides information on ambulatory medical care provided by nonfederally employed office-based physicians (NCHS 2011). It is based on a sample of visits to a national probability sample of office-based physicians. According to the 2007 survey estimate, there were 106.5 million office visits due to injury (Hsiao et al. 2010). The data includes 24 items with up to three ICD-9-CM diagnoses and offer the opportunity to estimate the proportion of TBIs treated in an outpatient setting. Schootman and Fuortes (2000) included NAMCS data in their study of rates of TBI-related ambulatory care in the USA.

Data from statewide trauma registries can also be used to study serious injury, but they vary considerably in composition and content (Mann et al. 2006) and typically are not representative. The National Trauma Databank (NTDB) represents the largest aggregation of US trauma registry data, and the data from the research data sets (RDS) can be used for studies that do not require population-based estimates (American College of Surgeons 2011a). Data from more recent years are more complete due to the implementation of the NTDB National Trauma Data Standard beginning in 2007.

The NTDB National Sample Program (NSP) is a national probability sample of data from Level I and II trauma centers selected from the NTDB (American College of Surgeons 2011b). It was developed to overcome limitations in the ability to draw inferences about the incidence and outcomes of injured patients at the national level inherent in the NTDB because of biases associated with voluntary reporting (Goble et al. 2009). Thus, the NSP can be used to provide nationally representative baseline estimates of trauma care for clinical outcomes research and injury surveillance. The NSP data were used by the National Highway Traffic Safety Administration to investigate the incidence rates of incapacitating injuries including TBI among children in motor vehicle traffic crashes (National Highway Traffic Safety Administration 2010).

Motor-Vehicle-Related Fatalities

The Fatality Analysis Reporting System (FARS) contains data on all vehicle crashes that occur on a public roadway and involve a fatality within 30 days after the crash (National Highway Traffic Safety Administration 2011) and is an important source of information on TBI-related deaths associated with this cause. Beginning in 1988, the General Estimates System (GES) was added to FARS. GES is a nationally representative sample of police-reported motor vehicle crashes of all types, from minor to fatal, which allows estimation of nonfatal, crash-related TBIs in the USA. FARS has been used to investigate the proportion of bicyclist fatalities for which head injury was a contributing factor (Nicaaj et al. 2009).

Sports

Because they are not routinely coded in the administrative data sets used for surveillance, sports and recreation activities are frequently underestimated as a cause of TBI, especially concussion/mTBI. For this reason, there has been increased interest in using other sports-related injury data collection systems for injury surveillance. Two examples are the NCAA Injury Surveillance System (ISS), a free internet-based athletic training record that allows monitoring of college level athletic participation, injuries, and treatments for all NCAA varsity sports (Dick et al. 2007; Hootman et al. 2007), and High School RIO™, the Internet-based data collection tool used in the National High School Sports-Related Injury Surveillance Study, a surveillance study of injuries in a national sample of US high school athletes (Center for Injury Research and Policy 2011). Examples of studies using these data sets are Gessel et al. (2007) and Frommer et al. (2011). Rates of TBI resulting from sports activities have also been derived from NEISS-AIP (Thurman et al. 1998; CDC 2007).

Use of Administrative Data Sets in Other Countries

Most of the previous examples illustrating the use of administrative data sources to assess TBI occurrence in populations are drawn from the USA. However, it should be noted that comparable resources exist and have been used to describe the epidemiology of TBI in other high-income

(Hyder et al. 2007; Tagliaferri et al. 2006) and some middle- and low-income countries (Hyder et al. 2007; Puvanachandra and Hyder 2009). Indeed, among countries with universal health-care systems with public insurance, medical records may be linked across all medical care venues—hospital, ED, and even outpatient sites. This may facilitate more comprehensive assessments of the spectrum of mild, moderate, and severe TBI occurrence (Colantonio et al. 2010). Linking such records for individual patients also enables the correction of duplicate reports that can arise when patients are treated at more than one site or at different times for the same injury. The WHO Collaborating Centres for Injuries have provided general guidelines for conducting TBI surveillance in high-income as well as middle- and low-income countries (Thurman et al. 1995b).

Quality of Data Sources

The incompleteness of some important data elements is a major problem in hospital discharge and ED data systems and trauma registries. This is in part due to limitations in the quality of clinical information that health-care providers record in the medical record, which adversely affect the accuracy of ICD coding. Glasgow Coma Scale scores, for example, may not be recorded in as many as 40% of the hospital medical records of patients with TBI (Thurman et al. 2006).

Alcohol use among TBI patients can complicate diagnosis in the ED by depressing the level of consciousness, resulting in inaccuracy in the initial assessment of TBI severity. In one study, this effect reportedly was independent of the severity of the injury (Jagger et al. 1984). Findings from more recent studies, however, suggested that alcohol intoxication generally did not result in a clinically relevant reduction in GCS in trauma patients with TBI (Stuke et al. 2007) except in those with the most severe injuries (Sperry et al. 2006) and those with very high blood alcohol levels (200 mg/dl or higher) who also had intracranial abnormalities detected on CT scan (Lange et al. 2010). Inaccurate assessment of individuals with TBI, especially concussion/mTBI, in the ED can contribute to missed diagnoses (Powell et al. 2008) and underestimates of the incidence of medically treated TBI.

Because most administrative data sets do not include measures of TBI severity such as the GCS, ICD code–based injury severity measures are often applied to these data sets. Examples are ICDMAP-90 software, which assigns Abbreviated Injury Scale 1990 (AIS) scores of the head based on TBI-related ICD-9-CM codes (MacKenzie et al. 1989). Alternatively, the Barell matrix (Clark and Ahmad 2006) categorizes TBIs into Type I (most severe), II, or III (least severe) (see Table 4.6). A limitation of these approaches is that the ICD-9-CM code 959.01—“head injury unspecified”—is not included; thus, cases with this code are not automatically assigned a level of severity. Some researchers using ICDMAP-90 or the Barell matrix make the assumption that all 959.01 cases are in the mild range of AIS scores for TBI or represent Type III cases in the Barell matrix, or simply modify the matrix to include an “unspecified severity” category.

Representativeness of the data source is an important concern in TBI surveillance using administrative data sets. Representativeness means that either (a) the data source accurately captures *all* of the events of interest (e.g., the NVSS from the US National Center for Health Statistics) or (b) the data source *samples* the events, that is, TBIs, in a systematic manner so that the sample reflects the referent population (e.g., HDD from the US National Center for Health Statistics). Methods for detecting and assessing the magnitude of the bias are discussed elsewhere (Klaucke 1992). The use of hospital discharge data for TBI surveillance without including Emergency Department data can result in a lack of representativeness. For example, analysis of TBI surveillance data from Emergency Departments in South Carolina revealed that black females and the uninsured were less likely to be admitted to hospital, even after adjustment for TBI severity and preexisting conditions (Selassie et al. 2004).

Table 4.6 Barel matrix for TBI

ICD-9-CM codes	Description
Type 1 TBIs (most severe)	
800, 801, 803, 804 (0.03–0.05, 0.1–0.4, 0.53–0.55, 0.6–0.9)	Recorded evidence of intracranial injury or moderate/prolonged (≥ 1 h), LOC, or injuries to optic nerve pathways
850 (0.2–0.4)	
851–854	
950 (0.1–0.3)	
995.55	
Type 2 TBIs	
800, 801, 803, 804 (0.00, 0.02, 0.06, 0.09, 0.50, 0.52, 0.56, 0.59)	No recorded evidence of intracranial injury and LOC <1 h or of unknown duration or unspecified
850 (0.0, 0.1, 0.5, 0.9)	
Type 3 TBIs (least severe)	
800, 801, 803, 804 (0.01, 0.51)	No recorded evidence of intracranial injury and no LOC

Source: (Clark and Ahmad 2006)

Similarly, the validity of TBI surveillance data is also a concern and should be evaluated. Methods for evaluating TBI surveillance data sets are described in the CDC's Central Nervous System Injury Surveillance Data Submission Standards – 2002 (Marr and Coronado 2004). They include calculating the predictive value positive (PVP) and the sensitivity of the ICD codes used for surveillance. These measures require identification of a confirmatory diagnostic measure such as information from neurological evaluations that could be extracted from medical chart review or neuroimaging data, for example, computed tomography (CT). These methods are described in detail by Fletcher et al. (1988) and Fleiss et al. (2003).

Epidemiologic Measures in TBI Surveillance and Research

In this section, key measures used in previous studies are defined, selected measurement tools are described, and some relevant publications using these measures are summarized, focusing primarily on population-based studies.

Incidence and Related Measures

Incidence refers to the number of new TBI events that occur in a specific population or geographic region within a specified period of time. In population-based studies of TBI, incidence is typically calculated using data from administrative data sets. Incidence represents the number of people who *had* a TBI event whether or not they experienced related symptoms or problems after the acute phase of the injury. It is important to note that these numbers include people who experienced a TBI but may have fully recovered.

Faul et al. (2010) estimated the incidence of TBI in the USA by analyzing combined data from the National Center for Health Statistics (NCHS) regarding TBI (1) deaths (NVSS), (2) hospital discharges (NHDS), and (3) ED visits (NHAMCS) using the CDC case definition (Marr and Coronado 2004) (Table 4.3). Denominator data were obtained from the US Census. Using this approach, Faul et al. (2010) reported an estimated average annual incidence of TBI in the USA of 1.7 million per year (579.0 per 100,000 per year, age-adjusted to the 2000 US standard population).

An important limitation of the study is its failure to include non-fatal cases that only received medical attention in outpatient care settings. In addition, because the NHDS and NHAMCS data are based on hospitalizations and visits to EDs, not on individual persons, there may be some duplication of cases treated for the same injury; however, the estimated effects were small (Faul et al. 2010; Langlois et al. 2004) For details of the limitations of studies combining these three data sets, see the methods sections from these reports.

The incidence of TBI in the USA occurring in the year 2000 was calculated using different data sets (Finkelstein et al. 2006). As in Faul et al. (2010), they used NVSS for mortality. However, unlike the Faul et al. study, Finkelstein et al. estimated the incidence of nonfatal injuries that resulted in medical treatment without hospitalization or ED treatment from the 1999 Medical Expenditure Panel Survey (MEPS), a survey of the civilian, noninstitutionalized population (AHRQ 2011c). Because the MEPS sample size for nonfatal hospitalized and ED-treated injuries is small, they estimated the incidence of these injuries using the 2000 Healthcare Cost and Utilization Project-Nationwide Inpatient Sample (HCUP-NIS) for counts of hospitalized injuries. They estimated the incidence of injuries treated in the ED from the 2001 National Electronic Injury Surveillance System – All Injury Program (NEISS-AIP) (note: 2001 is the first complete year of NEISS data collection). For the denominator of the incidence rates, they used population counts from the 1999 MEPS. Using these data, they estimated that more than 1.3 million TBIs occurred in the USA in 2000 (486/100,000 per year).

Recurrent TBI, also known as repetitive TBI, refers to the occurrence of multiple incident TBI or concussion/mTBI events to the same person. Recurrent TBI, including concussion/mTBI, is important because it is associated with prolonged recovery (Guskiewicz et al. 2003) and increased risk of a catastrophic outcome such as second impact syndrome (CDC 1997). Previous head injury (including TBI) has also been shown to be a risk factor for subsequent head injury in children (Swaine et al. 2007) and for repeat concussion in collegiate athletes (Guskiewicz et al. 2003). In studies using administrative databases, recurrent TBI is ascertained by identifying other TBI event(s) for each case that are unrelated to the first (i.e., that are not readmissions or transfers) using unique patient identifiers.

In one of the first population-based studies of recurrent TBI, Annegers et al. (1980) reviewed medical record data for a 10-year period and reported that 7.1% of males and 3.0% of females experienced a second head injury. In a more recent study, Saunders et al. (2009) used statewide hospital discharge and ED records and reported that 7% of those hospitalized with a TBI had a least one recurrent TBI during the follow-up period. As mentioned above, studies that include only injury events resulting in medical attention underestimate the true incidence rate because they exclude less severe TBIs.

Trends in TBI rates, that is, increases or decreases in the incidence rates of TBI over time, are of interest because they may reflect important changes in health care practices or the effects of prevention. Using the National Hospital Discharge Survey, Thurman and Guerrero (1999) reported a 51% decline in US hospitalization for TBI, especially mild TBI, during the period from 1980 through 1995. Similar findings in Canada during the decade 1992–2002 have been reported by Colantonio et al. (2009). Bowman et al. (2008), using the HCUP Nationwide Inpatient Sample (NIS), reported that the estimated annual incidence rate of US pediatric hospitalizations associated with TBI decreased from 1991 to 2005.

Lifetime Prevalence of a History of TBI

Lifetime prevalence of TBI refers to the number or percent of individuals who have “ever” experienced a TBI whether or not they continue to have persistent symptoms or related disability. McKinlay et al. (2008) reported a lifetime prevalence of TBI of 30% in a birth cohort followed from

ages 0 through 25 years. Lifetime prevalence is an important indicator of the impact of TBI because preceding TBI has been shown in studies of birth cohorts to be associated with negative effects on psychosocial development (McKinlay et al. 2008) and later psychiatric morbidity (Timonen et al. 2002). It is also considered to be an important comorbid condition with implications for treatment, for example, in persons with substance abuse problems (Olson-Madden et al. 2010; Walker et al. 2007; Corrigan and Deutschle 2008).

Because prospective studies are not always possible, retrospective methods for determining a person's self-reported lifetime history of TBI have also been developed (Cantor et al. 2004; Corrigan and Bogner 2007). The Ohio State University Traumatic Brain Injury (TBI) Identification Method (OSU TBI-ID) is a standardized procedure for eliciting lifetime history of TBI via a structured interview (Corrigan and Bogner 2007). The instrument is based on CDC case definitions (Marr and Coronado 2004) (Table 4.3). The OSU TBI-ID was designed to use self- or proxy-reports to elicit summary indices reflecting TBIs occurring over a person's lifetime (see figure for the short version; a long version can be requested from the authors). Preliminary support for the reliability and validity of the measure has been published (Corrigan and Bogner 2007; Bogner and Corrigan 2009) (Fig. 4.1). According to the authors, the OSU TBI-ID can be adapted for specific populations and situations, primarily by modifying the "probe" questions (the first five questions in the short version). Because it is essential to spend time helping a respondent recall injuries and events that may have resulted in a TBI, the authors recommend that the OSU TBI-ID be administered via interview (telephone or face-to-face). Professionals with a background in TBI typically grasp the tool quickly, as do novice interviewers who have had some basic training about TBI. Using the OSU TBI-ID, Olson-Madden et al. (2010) found that 55% of a sample of veterans seeking outpatient substance abuse treatment had a history of previous TBI.

Outcomes

Long-Term Adverse Health Outcomes

Of particular concern after TBI are adverse outcomes that affect health and the ability to function in society. Unique population-based studies involving surveillance of longer-term TBI outcomes (up to 3 years postinjury) were supported by the CDC. In both the Colorado Traumatic Brain Injury Registry and Follow-up System (Brooks et al. 1997) and the South Carolina TBI Registry (Pickelsimer et al. 2006), representative samples of persons hospitalized with TBI were identified from statewide hospital discharge data surveillance systems and interviewed by telephone to obtain information about TBI-related outcomes including service needs (Corrigan et al. 2004; Pickelsimer et al. 2007), problems with psychosocial health (McCarthy et al. 2006), and alcohol use (Horner et al. 2005). Limitations of these studies included the exclusion of patients with less severe injuries seen in EDs, outpatient clinics, and those not receiving care.

Disability

Incidence of TBI-related disability refers to the number of people in a defined geographic region within a specified time period who have experienced a TBI and have long-term or lifelong disability. Methods for estimating the incidence of TBI-related disability involve the development and validation of a predictive model and application of the predictors from that model to a population-based data set. Selassie et al. (2008) developed a predictive model using logistic

Ohio State University TBI Identification Method—Short Form*
(Version 10/19/10-Lifetime: to be used when querying about lifetime history of TBI)

I am going to ask you about injuries to your head or neck that you may have had anytime in your life. *Interviewer instruction :* Record cause and any details provided spontaneously in the box at the bottom of the page. You do not need to ask further about loss of consciousness or other details during this step.

1. In your lifetime, have you ever been hospitalized or treated in an emergency room following an injury to your head or neck? Think about any childhood injuries you remember or were told about.
 Yes—Record cause in table below
 No
2. In your lifetime, have you ever injured your head or neck in a car accident or from crashing some other moving vehicle like a bicycle, motorcycle or ATV?
 Yes—Record cause in table below
 No
3. In your lifetime, have you ever injured your head or neck in a fall or from being hit by something (for example, falling from a bike or horse, rollerblading, falling on ice, being hit by a rock)? Have you ever injured your head or neck playingsports or on the playground?
 Yes—Record cause in table below
 No
4. In your lifetime, have you ever injured your head or neck in a fight, from being hit by someone, or from being shaken violently? Have you ever been shot in the head?
 Yes—Record cause in table below
 No
5. In your lifetime, have you ever been nearby when an explosion or a blast occurred? If you served in the military, think about any combat-or training-related incidents.
 Yes—Record cause in table below
 No
6. If all above are “no” then proceed to question 7. If answered “yes” to *any* of the questions above, ask the following for each injury: **Were you knocked out or did you lose consciousness (LOC)? If yes, how long? If no, were you dazed or did you have a gap in your memory from the injury? How old were you?**

Cause	Loss of consciousness (LOC)/knocked out				Dazed/Mem Gap		Age
	No LOC	< 30 min	30 min-24 hrs	> 24 hrs.	Yes	No	

If more injuries with LOC : How many more? ___ Longest knocked out? ___ How many ≥ 30 mins.? ___
 Youngest age? ___

7. Have you ever lost consciousness from a drug overdose or being choked? ___# overdose ___#
 choked

SCORING

- _____ # TBI-LOC (number of TBI’s with loss of consciousness from #6)
- _____ # TBI-LOC ≥ 30 (number of TBI’s with loss of consciousness ≥ 30 minutes from #6)
- _____ age at first TBI-LOC (youngest age from #6)

Fig. 4.1 Ohio State University TBI Identification Method – Short Form*. (Version 10/19/10-Lifetime: to be used when querying about lifetime history of TBI)

- _____ **TBI-LOC before age 15** (if youngest age from #6 < 15 then =1, if ≥ 15 then = 0)
- _____ **Worst Injury (1-5):**
- If responses to #1-5 are “no” classify as 1 “**improbable TBI**”.
 - If in response to #6 reports never having LOC, being dazed or having memory lapses classify as 1 “**improbable TBI**”.
 - If in response to #6 reports being dazed or having a memory lapse classify as 2 “**possible TBI**”.
 - If in response to #6 loss of consciousness (LOC) does not exceed 30 minutes for any injury classify as 3 “**mild TBI**”.
 - If in response to #6 LOC for any one injury is between 30 minutes and 24 hours classify as 4 “**moderate TBI**”.
 - If in response to #6 LOC for any one injury exceeds 24 hours classify as 5 “**severe TBI**”.
- _____ **# anoxic injuries** (sum of incidents reported in #7)

*adapted with permission from the Ohio State University TBI Identification Method (Corrigan, J.D., Bogner, J.A. (2007). Initial reliability and validity of the OSU TBI Identification Method. *J Head Trauma Rehabil*, 22(6):318-329,

© reserved 2007, The Ohio Valley Center for Brain Injury Prevention and Rehabilitation

Fig. 4.1 (continued)

regression and data on post-TBI disability from a population-based sample of persons hospitalized with TBI from the South Carolina TBI Follow-up Registry (Pickelsimer et al. 2006). The regression coefficients were then applied to the 2003 HCUP NIS data to estimate the annual incidence of long-term disability in the USA following TBI hospitalization. In that study, an estimated 43.3% of hospitalized TBI survivors in the USA in 2003 experienced a TBI with related long-term disability (Selassie et al. 2008). These figures are likely underestimates because they are based on hospitalizations only and exclude TBIs treated in other settings or for which treatment was not sought.

Prevalence of TBI-related disability refers to the number of people in a defined geographic region, such as the USA, who have ever experienced a TBI *and* are living with symptoms or problems related to the TBI. This excludes people who had a TBI and recovered from it. Zaloshnja et al. (2008) estimated the number of people who experienced long-term disability from TBI each year in the past 70 years by applying estimates from a previous study of the incidence of TBI-related disability (Selassie et al. 2008) to data from the National Hospital Discharge Survey from 1979 to 2004. Then, after accounting for the mortality among TBI survivors, the authors estimated their life expectancy and calculated how many were expected to be alive in 2005. Applying this method, the estimated number of persons in the USA living with disability related to a TBI hospitalization was 3.2 million.

Estimates of the incidence and prevalence of TBI-related disability using these methods are limited by the omission of cases of less severe TBI. These studies used hospital discharge data only and thus do not include persons treated and released from Emergency Departments or who received no medical care. This is in part because data for TBI incidence and for mortality over

an extended period of time, for example, 70 years, are needed and are not readily available for persons treated in these health-care settings. Thus, available data only allow for meaningful estimates of the risk of disability after moderate and severe TBI. Another limitation is that there is no universally agreed-upon definition of TBI-related disability. The definition used by Selassie et al. (2008) was based on the findings from their study and included three domains: general health, mental and emotional health, and cognitive symptoms. Finally, it is important to consider the potential contribution of comorbid conditions to long-term disability. Selassie et al. (2008) found that preexisting comorbidity as assessed from the ICD-9-CM codes found in the hospital discharge records was strongly associated with disability, and thus, they adjusted for it in their model.

Late Mortality

Late mortality refers to TBI-related death occurring after the acute phase of recovery is over. In most previous population-based studies, late mortality has been assessed after discharge from acute care hospitalization (Selassie et al. 2005; Ventura et al. 2010). Information about late mortality is of interest because of the potential for serious injury such as TBI to adversely affect overall health and thus contribute to reduced life expectancy (Shavelle et al. 2006). Ventura et al. (2010) found that patients with TBI carried about 2.5 times the risk of death compared with the general population. As in the studies of disability described above, these late mortality findings are not generalizable to persons with less severe TBI who were not hospitalized, and the causal link between the TBI event and death can only be inferred.

Economic Cost

The economic burden of traumatic brain injury was investigated as part of a large and comprehensive study of the incidence and economic burden of injuries in the USA (Finkelstein et al. 2006). The authors combined several data sets to estimate the incidence of fatal and nonfatal injuries in the year 2000. They calculated unit medical and productivity costs, multiplied these costs by the corresponding incidence estimates, and reported the estimated lifetime costs of injuries occurring in 2000, with the estimated lifetime costs of TBI in their study totaling more than \$60 billion. Orman et al. (2011) reported more detailed estimates of the lifetime costs of TBI. Unlike the previous estimates, the latter included lost quality of life. They found that, in 2009 dollars, the estimated total lifetime comprehensive costs of fatal, hospitalized, and nonhospitalized TBI among civilians that were medically treated in the year 2000 totaled more than \$221 billion, including \$14.6 billion for medical costs, \$69.2 billion for work loss costs, and \$137 billion for the value of lost quality of life. Notably, the nonhospitalized TBI category included cases presenting for ED, office-based, or hospital outpatient visits. These cost estimates are limited by the fact that they do not adequately account for the costs of extended rehabilitation, services, and supports, such as informal caregiving, that are needed by those with long-term or lifelong TBI-related disability nor the value of lost quality of life or productivity losses for informal caregivers, including parents. Conversely, these estimates represent only TBIs associated with medical treatment. It is likely that the per person costs associated with most concussion/mTBIs are substantially less than the estimates resulting from this study methodology.

TBI Surveillance in Military Personnel and Veterans

Clinical Case Definition

The Department of Veterans Affairs/Department of Defense (VA/DoD 2009) TBI case definition was developed with input from both military and civilian TBI experts. Because it addresses issues specific to TBI among service members and veterans and differs slightly from previous definitions developed for civilian populations, the VA/DoD definition is summarized here:

- TBI is defined as a traumatically induced structural injury and/or physiological disruption of brain function as a result of an external force that is indicated by new onset of at least one of the following clinical signs, immediately following the event:
 - Any period of loss of or a decreased level of consciousness
 - Any loss of memory for events immediately before or after the injury
 - Any alteration in mental state at the time of the injury [confusion, disorientation, slowed thinking, etc., also known as alteration of consciousness (AOC)]
 - Neurological deficits (weakness, loss of balance, change in vision, praxis, paresis/plegia, sensory loss, aphasia, etc.) that may or may not be transient
 - Intracranial lesion
- External forces may include any of the following events: the head being struck by an object, the head striking an object, the brain undergoing an acceleration/deceleration movement without direct external trauma to the head, a foreign body penetrating the brain, forces generated from events such as a blast or explosion, or other force yet to be defined.

It is important to note that the above criteria define the “event” of a TBI. Not all individuals exposed to an external force will sustain a traumatic brain injury, but any person who has a history of such an event with manifestations of any of the above signs and symptoms, most often occurring immediately or within a short time after the event, can be said to have had a TBI. (VA/DoD 2009)

When evaluating the VA/DoD clinical case definition, it is important to keep in mind that diagnosing TBI among service members, especially those injured in combat, presents some unique challenges compared with the civilian setting. Although the diagnosis of moderate and severe TBI among service members is relatively straightforward even in a theater of war because the clinical signs and symptoms, abnormalities seen on neuroimaging, and the resulting functional deficits typically are readily apparent, the accurate identification of concussion/mild TBIs can be problematic. The reasons include the fact that (a) the often high pace of combat operations, referred to as OPTEMPO, and constraints on access to health care clinics in theater decrease the likelihood that an injured service member will be evaluated by a qualified provider soon after the injury event while concussion/mTBI signs and symptoms are observable; (b) there are limited diagnostic tools with known sensitivity and specificity that can be administered in the combat environment; (c) diagnoses based on self-report of exposure to an injury event are adversely affected by problems with recall, especially when the period of AOC or LOC is brief; and (d) concussion/mTBI symptoms overlap with those of other conditions such as acute stress reaction/post-traumatic stress disorder (Iverson et al. 2009; Hoge et al. 2008; Schneiderman et al. 2008; Marx et al. 2009; Pietrzak et al. 2009; Cooper et al. 2010; Kennedy et al. 2010; Polusny et al. 2011).

It is important to note that the case definition for concussion/mTBI summarized above was designed to be applied in the *acute* injury period. Thus, it lacks essential criteria for assessment of concussion/mTBI history, including the lack of specific symptoms, time course, and functional impairment. (Hoge et al. 2009). As a result, when it is used to assess concussion/mTBI weeks or months after the injury based on self-report, such as in some health screening programs, including the DoD’s postdeployment health assessment (PDHA Form 2796) and postdeployment health reassessment

(PDHRA Form 2900), subjective attribution of non-mTBI related symptoms to concussion/mTBI may occur (Hoge et al. 2009; Iverson et al. 2009). Misattribution of nonspecific symptoms, for example, headache, which may be due to other causes and not related to the injury event, can result in an overestimate of the true number of cases of concussion/mTBI. Estimates of the occurrence of TBI, including concussion/mTBI, based on results of screening have been reported (Hoge et al. 2008; Tanielian and Jaycox 2008; Terrio et al. 2009).

Enhanced surveillance for concussion/mTBI among deployed service members may be possible using the Blast Exposure and Concussion Incident Report (BECIR) (U.S. Medicine 2011). Under current Department of Defense guidelines for BECIR, every service member who is exposed to a potential concussion/mTBI, for example, who is within a specified distance of an explosion or blast, must be screened for common concussion/mTBI-related signs and symptoms, and the results must be recorded in the military's operational information system. Although originally designed to facilitate identification and clinical management of service members who sustain concussion/mTBI during deployment, the BECIR data may be useful in improving estimates of the incidence of combat-related concussion/mTBI.

DoD's Standard TBI Surveillance Case Definition for Administrative Health Care Data

A collaborative effort among experts from the Departments of Defense and Veterans Affairs and the civilian sector resulted in a standard case definition for surveillance of TBI among military personnel (AFHSC 2008, 2009, 2011a) (Table 4.7). The Armed Forces Health Surveillance Center (AFHSC) reports published prior to October 2008 used an older surveillance case definition (AFHSC 2008). Both the new and old DoD case definitions are similar, but not directly comparable, to that recommended by the CDC (Marr and Coronado 2004). Unlike the CDC definition, the DoD definition includes a range of V-codes and DoD-specific "extender codes" used within the DoD health system to capture information about self-reported history of injury (Tricare 2009). [These "extender codes" appear as an underscore followed by a number or letter directly after the V-code (see Table 4.7)]. Thus, the DoD definition allows inclusion of potential prevalent cases of TBI. An adapted version of the Barell Index for use with the DoD/VA standard surveillance case definition has been published (Wojcik et al. 2010a). Of note, the AFHSC definition is updated periodically, and a more recent version may currently be in use.

DoD Surveillance Methods

Two primary sources routinely report surveillance data for TBI among service members. The first source, the DoD TBI Numbers Web site, reports the numbers of service members with TBI diagnosed by a medical provider (DoD 2011). Cases are ascertained from electronic records of service members diagnosed anywhere in the world where the standard Department of Defense electronic health-care record, the Armed Forces Health Longitudinal Tracking Application (AHLTA), is used (DHIMS 2011). Second, population-based estimates of the numbers of service members and Veterans who sustain a TBI at any level of severity are routinely reported as a "deployment-related condition of special surveillance interest" by the AFHSC in their monthly publication, the *Medical Surveillance Monthly Report* (MSMR), available on line at the AFHSC Web site.

In a special report also in MSMR, the AFHSC published a detailed description of their surveillance methods and the challenges in calculating the incidence of TBI among service members using

Table 4.7 Department of Defense standard TBI surveillance case definition

The following ICD9 codes are included in the case definition^{a, b}:

ICD-9-CM codes

310.2 (postconcussion syndrome)

800.0x–800.9x (fracture of vault of skull)

801.0x–801.9x (fracture of base of skull)

803.0x–803.9x (other and unqualified skull fractures)

804.0x–804.9x (multiple fractures involving skull or face with other bones)

850.x (concussion)

851.0x–851.9x (cerebral laceration and contusion)

852.0x–852.5x (subarachnoid, subdural, and extradural hemorrhage, following injury)

853.0x–853.1x (other and unspecified intracranial hemorrhage following injury)

854.0x–854.1x (intracranial injury of other and unspecified nature)

907.0 (late effect of intracranial injury *without* skull or facial fracture)

950.1–950.3 (injury to optic chiasm/pathways or visual cortex)

959.01 (head injury, unspecified)

(Personal history of TBI)

V15.52 (no extenders); V15.52_0 thru V15.52_9; V15.52_A thru V15.52_F (currently only codes in use)

V15.5_1 thru V15.5_9; V15.5_A thru V15.5_F

V15.59_1 thru V15.59_9; V15.59_A thru V15.59_F

Source: (Armed Forces Health Surveillance Center AFHSC 2011a, b)

^aICD-9-CM code 995.55 (shaken infant syndrome) is included in the standard DoD TBI case definition in an effort to be consistent with the CDC. This code is not used by AFHSC as it is not relevant to military surveillance objectives

^bCase definition and ICD-9-CM codes are based on “TBI: Appendix F-G dated 5/1/10 and Appendix 7 dated 2/26/10: from *Military Health System Coding Guidance: Professional Services and Specialty Coding Guidelines* (Version 3.2) by the Unified Biostatistical Utility working group”

administrative health-care data (AFHSC 2009). Special considerations in reporting TBI surveillance data for service members include the classification of injury severity. Specifically, in addition to mild, moderate and severe, penetrating injuries are considered to have different prognostic significance and thus are categorized separately. With regard to external cause and setting, war-related TBIs are often associated with mechanisms not specified in routine civilian surveillance reports. These include explosions or blasts (Bell et al. 2009; Ling and Ecklund 2011) and high-caliber gunshot wounds (Bell et al. 2009). Whether the injury occurred in a battle vs. nonbattle setting is also of interest (AFHSC 2007; Wojcik et al. 2010b) but has typically been very difficult to differentiate reliably. External cause categories reported by AFHSC (2007) include falls, athletics/sports, assault, and accidental weapon-related. Although of considerable interest due to the ongoing conflicts in Iraq and Afghanistan, in one report, estimates of battle casualty-related TBIs accounted for a very small proportion of all TBI-related hospitalizations both prewar (0.3%) and during the wars (3.2%) (Orman et al. 2011).

Trends in TBI-related health-care encounters are also of interest. AFHSC (2011b) reported a trend toward increasing numbers of TBI-related ED visits among active duty US Armed Forces from 2001 to 2010, excluding visits for military personnel in civilian facilities and deployed settings. The potential effects of a wide range of changes since 2001, the onset of the conflicts in Afghanistan and Iraq, should be considered when interpreting these findings. Such changes include changes in TBI-related diagnostic procedures and guidelines, diagnostic coding practices, and awareness and concern among service members, commanders and supervisors, family members, and primary care and other health-care providers, which may have contributed to the higher rates (AFHSC 2011b).

Surveillance data for TBIs among service members based on health-care encounters have some limitations. As for civilians, the number of service members who receive medical care but for

whom the TBI is not diagnosed, or who sustain a TBI but do not seek care, is not known. Also, external cause information is incomplete and was missing/invalid for 25% of prewar TBI-related hospitalizations and 38% of those occurring postwar (AFHSC 2007). Finally, because denominator data, that is, the total number of deployed service members at risk of TBI, are not routinely available, deployment-specific TBI rates typically are not calculated but have been estimated in two studies (Ivins 2010; Wojcik et al. 2010b). This limits interpretation and comparison with data from other sources, such as from civilian data surveillance systems. Calculation of rates is needed to increase the usefulness of military TBI surveillance for guiding prevention efforts.

Combat-Related Trauma

As for TBI among civilians, trauma registries can be a useful source of data for studying serious traumatic brain injury among military personnel. Developed in 2004 at the United States Army Institute of Surgical Research (USAISR), The Joint Theater Trauma Registry (JTTR) is a standardized, retrospective data collection system for all echelons of combat casualty care that is similar in design to civilian trauma registries. The JTTR was the first organized effort by the US military to collect data on trauma occurring during an active military conflict (Glenn et al. 2008) and was designed to inform advances in medical care aimed at improving the outcome of soldiers wounded on the battlefield (Eastridge et al. 2006, 2009). Although not currently used for surveillance of combat-related TBI, the JTTR includes a range of data that would be useful for TBI surveillance, such as demographics, injury cause, mechanism and type, intentionality, ICD-9-CM diagnosis codes, external cause of injury codes (E codes), medical procedure codes (V-Codes), Abbreviated Injury Scale scores (AIS), Injury Severity Scores, and Glasgow Coma Scale scores. Because the JTTR includes detailed information about the medical care received, the data could be used for studies of trends in the types of TBI treatments used at various times and their association with changes in outcomes such as mortality. To date, few studies specifically focused on TBI have been conducted using JTTR data; however, DuBose et al. (2011) showed the potential for using JTTR to identify severe cases of combat-related TBI in their study of the relationship between neurosurgical interventions and outcomes.

Disability

For military personnel, disability is routinely defined as the inability to return to duty. Within the US Army, ability to return to duty is determined by the Army Physical Evaluation Board (PEB), an administrative body made up of medical personnel and Army officers who are responsible for determining if an ill or injured soldier is able to perform his or her job in the Army, that is, whether they are “fit for duty” (Cross et al. 2011). A condition that is judged to contribute to a soldier’s inability to return to duty is referred to as an “unfitting condition.” Studies conducted at the USAISR were among the first to quantify the disability associated with the wars in Afghanistan (OEF) and Iraq (OIF) by reviewing the PEB database. Cross et al. found that TBI was the eighth most frequent unfitting condition among soldiers injured between October 2001 and January 2005 identified from the JTTR. More recently, Patzkowski et al. (2011) queried the full PEB database and reported that for the first 3 months of 2009, TBI comprised 8% of the unfitting conditions for Army soldiers and ranked sixth, following back pain, osteoarthritis, PTSD, foot and ankle conditions, and psychiatric conditions. Similar studies for the other armed services would provide a more complete picture of the impact of TBI on return to duty for the entire US military force.

Future Directions in TBI Surveillance

Technological advancements are likely to lead to improvements in TBI diagnosis and related increases in the accuracy of case ascertainment for research and surveillance, especially for concussion/mTBI. Some examples include the following:

Neuroimaging. Accurate diagnosis of concussion/mTBI remains challenging due to the limitations of sign- and symptom-based diagnosis. However, recent studies suggest that structural abnormalities identified using more advanced neuroimaging techniques such as diffusion tensor imaging (DTI) might serve as quantitative biomarkers for concussion/mTBI (Niogi et al. 2008a, b; Wilde et al. 2008; Benzinger et al. 2009; MacDonald et al. 2011). Improvements in TBI diagnosis based on neuropathology will lead to an improved classification system for all levels of TBI severity not only for clinical research (Saatman et al. 2008) but also for epidemiologic studies.

Serum Biomarkers. Levels of certain biomarkers in blood measured after traumatic brain injury (TBI) may prove to be useful diagnostic and prognostic tools in addition to clinical indices for detection of blast-induced neurotrauma (Svetlov et al. 2009). If such biomarkers were found to be reliable for detecting concussion/mTBI, they would provide a more objective measure than symptom reporting. Promising candidates include S100B and GFAP (Vos et al. 2010).

Helmet Sensors. Electronic sensors have been placed in both football helmets (McCaffrey et al. 2007) and the helmets of service members (Army Technology 2011) to detect impacts from physical contact or blast/explosions. Data from these devices can be used as indicators of the impact to the brain of exposure to external forces and provide alerts to the possibility of sufficient impact to cause a concussion. Although not diagnostic, these sensors can be used to monitor the need to assess for symptoms of possible concussion. They can also be used to monitor the cumulative effect of multiple impacts that may be associated with recurrent concussions.

References

- Adekoya, N., Thurman, D. J., White, D. D., et al. (2002). Surveillance for traumatic brain injury deaths – United States, 1989–1998. *Morbidity and Mortality Weekly Report Surveillance Summaries*, 51(10), 1–14.
- Agency for Healthcare Research and Quality (AHRQ) (2011a). Overview of HCUP. <http://www.hcup-us.ahrq.gov/overview.jsp>. Accessed 13 Apr 2011.
- Agency for Healthcare Research and Quality (AHRQ) (2011b). Overview of the state inpatient databases (SID). <http://www.hcup-us.ahrq.gov/sidoverview.jsp>. Accessed 13 Apr 2011.
- Agency for Healthcare Research and Quality (AHRQ) (2011c). Medical Expenditure Panel Survey (MEPS). <http://www.meps.ahrq.gov/mepsweb/>. Accessed 15 June 2011.
- American College of Surgeons (2011a). National Trauma Data Bank Research Data Sets. <http://facs.org/trauma/ntdb/ntdbapp.html>. Accessed 13 Apr 2011
- American College of Surgeons (2011b). National Trauma Data Bank National Sample Program. <http://facs.org/trauma/ntdb/nsp.html>. Accessed 13 Apr 2011
- American Congress of Rehabilitation Medicine (ACRM). (1993). Definition of mild traumatic brain injury. *Journal of Head Trauma Rehabilitation*, 8, 86–87.
- Annegers, J. F., Grabow, J. D., Kurland, L. T., et al. (1980). The incidence, causes and secular trends of head trauma in Olmsted County, Minnesota, 1935–1974. *Neurology*, 30, 912–919.
- Armed Forces Health Surveillance Center (AFHSC) (2007). Traumatic brain injury among members of active components, U.S. Armed Forces, 1997–2006. *Medical Surveillance Monthly Report*, 14(5), 2–6.
- Armed Forces Health Surveillance Center (AFHSC) (2008). New surveillance case definitions for traumatic brain injury. *Medical Surveillance Monthly Report*, 15(8), 24.
- Armed Forces Health Surveillance Center (AFHSC). (2009). Deriving case counts from medical encounter data: considerations when interpreting health surveillance reports. *Medical Surveillance Monthly Report*, 16(12), 2–8.
- Armed Forces Health Surveillance Center (AFHSC) (2011a). Surveillance case definitions. http://www.afhsc.mil/viewDocument?file=CaseDefs/Web_13_NEUROLOGY_MAR11.pdf. Accessed 3 June 2011.

- Armed Forces Health Surveillance Center (AFHSC) (2011b). Surveillance Snapshot: Emergency department visits for traumatic brain injury. *I8(5)*, 15.
- Army Technology (2011). Net Resources International. <http://www.army-technology.com/news/news4449.html>. Accessed 15 June 2011.
- Association for the Advancement of Automotive Medicine (AAAM). (1990). *The Abbreviated Injury Scale (1990 Revision)*. Des Plaines, IL: Association for the Advancement of Automotive Medicine.
- Bakhos, L. L., Lockhart, G. R., Myers, R., & Linakis, J. G. (2010). Emergency department visits for concussion in young child athletes. *Pediatrics*, *126*, e550–e556.
- Bazarian, J. J., Veazie, P., Mookerjee, S., & Lerner, B. (2006). Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes. *Academic Emergency Medicine*, *13*, 31–38.
- Bell, R. S., Vo, A. H., Neal, C. J., et al. (2009). Military traumatic brain and spinal column injury: a 5-year study of the impact of blast and other military grade weaponry on the central nervous system. *Journal of Trauma*, *66(4 Suppl)*, S104–S111.
- Benzinger, T. L., Brody, D., Cardin, S., et al. (2009). Blast-related brain injury: imaging for clinical and research applications: report of the 2008 St Louis Workshop. *Journal of Neurotrauma*, *26*, 2127–2144.
- Bogner, J., & Corrigan, J. D. (2009). Reliability and predictive validity of the Ohio State University TBI identification method with prisoners. *Journal of Head Trauma Rehabilitation*, *24*, 279–291.
- Bowman, S. M., Bird, T. M., Aitken, M. E., et al. (2008). Trends in hospitalizations associated with pediatric traumatic brain injuries. *Pediatrics*, *122*, 988–993.
- Brooks, C. A., Gabella, B., Hoffman, R., et al. (1997). Traumatic brain injury: designing and implementing a population-based follow-up system. *Archives of Physical Medicine and Rehabilitation*, *78(8 Suppl 4)*, S26–30.
- Cantor, J. B., Gordon, W. A., Schwartz, M. E., et al. (2004). Child and parent responses to a brain injury screening questionnaire. *Archives of Physical Medicine and Rehabilitation*, *85(4 Suppl 2)*, S54–60.
- Carroll, L. J., Cassidy, J. D., Holm, L., et al. (2004). Methodological issues and research recommendations for mild traumatic brain injury: the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *Journal of Rehabilitation Medicine*, (43 Suppl), 113–125
- Center for Injury Research and Policy (2011). High School Rio™. <http://injuryresearch.net/highschoolrio.aspx>. Accessed 13 Apr 2011.
- Centers for Disease Control and Prevention (CDC). (1997). Sports-related recurrent brain injuries – United States. *Morbidity and Mortality Weekly Report Surveillance*, *46*, 224–227.
- Centers for Disease Control and Prevention (CDC). (2001). National estimates of nonfatal injuries treated in hospital emergency departments –United States, 2000. *Morbidity and Mortality Weekly Report Surveillance*, *50*, 340–346.
- Centers for Disease Control and Prevention (CDC) (2007). Nonfatal traumatic brain injuries from sports and recreation activities – United States, 2001–2005. *Morbidity and Mortality Weekly Report Surveillance*, *56*, 733–737
- Centers for Medicare and Medicaid Services (CMS) (2010). UB-04 Overview. http://www.cms.gov/MLNProducts/downloads/ub04_fact_sheet.pdf – 2010-09-07. Accessed 13 Apr 2011.
- Clark, D. E., & Ahmad, S. (2006). Estimating injury severity using the Barell matrix. *Injury Prevention*, *12*, 111–116.
- Coben, J. H., Steiner, C. A., & Miller, T. R. (2007). Characteristics of motorcycle-related hospitalizations: comparing states with different helmet laws. *Accident Analysis and Prevention*, *39*, 190–196.
- Colantonio, A., Croxford, R., Farooq, S., et al. (2009). Trends in hospitalization associated with traumatic brain injury in a publicly insured population, 1992–2002. *Journal of Trauma*, *66*, 179–83. theater.
- Colantonio, A., Saverino, C., Zagorski, B., et al. (2010). Hospitalizations and emergency department visits for TBI in Ontario. *Canadian Journal of Neurological Sciences*, *37*, 783–90.
- Cooper, D. B., Kennedy, J. E., Cullen, M. A., et al. (2010). Association between combat stress and post-concussive symptom reporting in OIF/OEF service members with mild traumatic brain injuries. *Brain Injury*, *25*, 1–7.
- Corrigan, J. D., & Bogner, J. (2007). Initial reliability and validity of the Ohio State University TBI Identification method. *Journal of Head Trauma Rehabilitation*, *22*, 318–329.
- Corrigan, J. D., & Deuschle, J. J. (2008). The presence and impact of traumatic brain injury among clients in treatment for co-occurring mental illness and substance abuse. *Brain Injury*, *22*, 223–31.
- Corrigan, J. D., Whiteneck, G., & Mellick, D. (2004). Perceived needs following traumatic brain injury. *Journal of Head Trauma Rehabilitation*, *19*, 205–216.
- Cross, J. D., Ficke, J. R., Hsu, J. R., et al. (2011). Battlefield orthopaedic injuries cause the majority of long-term disabilities. *Journal of the American Academy of Orthopaedic Surgeons*, *19(Suppl 1)*, S1–S7.
- Defense Health Information Management System (DHIMS) (2011). About AHLTA. <http://dhims.health.mil/userSupport/ahlta/about.aspx>. Accessed 11 Apr 2011
- Department of Defense (DOD) (2011). Traumatic brain injury numbers. 17 Feb 2011. <http://www.dvbic.org/TBI-Numbers.aspx>. Accessed 16 Mar 2011.
- Department of Health and Human Services (1989). International Classification of Diseases: 9th Revision, Clinical Modification, 3rd ed. (ICD-9-CM). Washington (DC): Department of Health and Human Services (US).

- Department of Veterans Affairs, Department of Defense (VA/DoD) (2009) VA/DoD Clinical Practice Guideline for Management of Concussion/Mild Traumatic Brain Injury (mTBI), Version 1.0. http://www.healthquality.va.gov/mtbi/concussion_mtbi_full_1_0.pdf. Accessed 16 Mar 2011.
- Dick, R., Hertel, J., Agel, J., et al. (2007). Descriptive epidemiology of collegiate men's basketball injuries: National Collegiate Athletic Association Injury Surveillance System, 1988–1989 through 2003–2004. *Journal of Athletic Training*, *42*, 194–201.
- DuBose, J., Barmparas, G., Inaba, K., et al. (2011). Isolated severe traumatic brain injuries sustained during combat operations: demographics, mortality outcomes, and lessons to be learned from contrasts to civilian counterparts. *Journal of Trauma*, *70*, 11–18.
- Eastridge, B. J., Costanzo, G., Jenkins, D., et al. (2009). Impact of Joint Theater Trauma System initiatives on battlefield injury outcomes. *American Journal of Surgery*, *198*, 852–857.
- Eastridge, B. J., Jenkins, D., Flaherty, S., et al. (2006). Trauma system development in a theater of war: experiences from Operation Iraqi Freedom and Operation Enduring Freedom. *Journal of Trauma*, *61*, 1366–1373.
- Eisele, J. A., Kegler, S. R., Trent, R. B., et al. (2006). Nonfatal traumatic brain injury-related hospitalization in very young children – 15 states, 1999. *Journal of Head Trauma Rehabilitation*, *6*, 537–543.
- Faul, M., Su, L., Wald, M. M., et al. (2010) Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations, and Deaths 2002–2006. Atlanta (GA): Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. http://www.cdc.gov/traumaticbraininjury/pdf/blue_book.pdf. Accessed 19 Mar 2010.
- Finkelstein, E. A., Corso, P. S., Miller, T. R., et al. (2006). *Incidence and economic burden of injuries in the United States*. Oxford: New York.
- Fleiss, J., Levin, B., & Paik, M. (2003). *An introduction to applied probability: the evaluation of a screening test. In statistical methods for rates and proportions* (3rd ed., pp. 1–16). New York: Wiley.
- Fletcher, R., Fletcher, S., & Wagner, E. (1988). *Diagnosis: bias in establishing sensitivity and specificity. In clinical epidemiology: the essentials* (2nd ed., pp. 51–54). Baltimore: Williams and Wilkins.
- Frommer, L. J., Gurka, K. K., Cross, K. M., et al. (2011). Sex differences in concussion symptoms of high school athletes. *Journal of Athletic Training*, *46*, 76–84.
- Gessel, L. M., Fields, S. K., Collins, C. L., et al. (2007). Concussions among United States high school and collegiate athletes. *Journal of Athletic Training*, *42*, 495–503.
- Glenn, M. A., Martin, K. D., Monzon, D., et al. (2008). Implementation of a combat casualty trauma registry. *Journal of Trauma Nursing*, *15*, 181–184.
- Goble, S., Neal, M., Clark, D. E., et al. (2009). Creating a nationally representative sample of patients from trauma centers. *Journal of Trauma*, *67*, 637–644.
- Guskiewicz, K. M., McCrea, M., Marshall, S. W., et al. (2003). Cumulative effects associated with recurrent concussion in collegiate football players: the NCAA Concussion Study. *Journal of the American Medical Association*, *290*, 2549–2555.
- Hanzlick, R., & Combs, D. (1998). Medical examiner and coroner systems: history and trends. *Journal of the American Medical Association*, *279*, 870–874.
- Hoge, C. W., Goldberg, H. M., & Castro, C. A. (2009). Care of war veterans with mild traumatic brain injury – flawed perspectives. *The New England Journal of Medicine*, *360*, 1588–1591.
- Hoge, C. W., McGurk, D., Thomas, J. L., et al. (2008). Mild traumatic brain injury in U.S. soldiers returning from Iraq. *The New England Journal of Medicine*, *358*, 453–463.
- Hootman, J. M., Dick, R., & Agel, J. (2007). Epidemiology of collegiate injuries for 15 sports: summary and recommendations for injury prevention initiatives. *Journal of Athletic Training*, *42*, 311–319.
- Horner, M. D., Ferguson, P. L., Selassie, A. W., et al. (2005). Patterns of alcohol use 1 year after traumatic brain injury: a population-based epidemiologic study. *Journal of the International Neuropsychological Society*, *11*, 322–330.
- Hsiao, C. J., Cherry, D. K., Beatty, P. C., et al. (2010) National Ambulatory Medical Care Survey: 2007 Summary. National Health Statistics Reports; no 27 (pp. 1–32) Hyattsville, MD: National Center for Health Statistics
- Hubbard, G. (2010). *New Mexico injury indicators report*. Santa Fe: New Mexico Department of Health.
- Hyder, A. A., Wunderlich, C. A., Puvanachandra, P., et al. (2007). The impact of traumatic brain injuries: a global perspective. *NeuroRehabilitation*, *22*, 341–353.
- Iverson, G. L., Langlois, J. A., McCrea, M. A., et al. (2009). Challenges associated with post-deployment screening for mild traumatic brain injury in military personnel. *Clinical Neuropsychology*, *23*, 1299–1314.
- Ivins, B. J. (2010). Hospitalization associated with traumatic brain injury in the active duty US Army: 2000–2006. *NeuroRehabilitation*, *26*, 199–212.
- Jagger, J., Fife, D., Venberg, K., et al. (1984). Effect of alcohol intoxication on the diagnosis and apparent severity of brain injury. *Neurosurgery*, *15*, 303–306.
- Jagoda, A. S., Bazarian, J. J., Bruns, J. J., Jr., et al. (2008). Clinical policy: neuroimaging and decision making in adult mild traumatic brain injury in the acute setting. *Annals of Emergency Medicine*, *52*, 714–748.

- Klaucke, D. N. (1992). Evaluating a public health surveillance system. In W. Halperin & E. Baker Jr. (Eds.), *Public health surveillance* (1st ed., pp. 26–41). New York: Van Nostrand Reinhold.
- Kennedy, J. E., Leal, F. O., Lewis, J. D., et al. (2010). Posttraumatic stress symptoms in OIF/OEF service members with blast-related and no-blast-related mild TBI. *NeuroRehabilitation*, *26*, 223–231.
- Lange, R. T., Iverson, G. L., Burbacher, J. R., et al. (2010). Effect of blood alcohol level on Glasgow Coma Scale scores following traumatic brain injury. *Brain Injury*, *24*, 819–827.
- Langlois, J. A., Kegler, S. R., Butler, K. E., et al. (2003). Traumatic brain injury-related hospital discharges. Results from a 14-state surveillance system, 1997. *Morbidity and Mortality Weekly Report Surveillance Summaries*, *52*(4), 1–20.
- Langlois, J. A., Rutland-Brown, W., & Thomas, K. E. (2004). *Traumatic brain injury in the United States: emergency department visits, hospitalizations, and deaths*. Atlanta (GA): Centers for Disease Control and Prevention, National Center for Injury Prevention and Control.
- Ling, G. S., & Ecklund, J. M. (2011). Traumatic brain injury in modern war. *Current Opinion in Anaesthesiology*, *24*, 124–130.
- MacDonald, C. L., Johnson, A. M., Cooper, D., et al. (2011). Detection of blast-related traumatic brain injury in US military personnel. *The New England Journal of Medicine*, *364*, 2091–2100.
- MacKenzie, E. J., Steinwachs, D. M., & Shankar, B. (1989). Classifying trauma severity based on hospital discharge diagnoses. Validation of an ICD-9-CM to AIS-85 conversion table. *Medical Care*, *27*, 412–422.
- Mann, N. C., Guice, K., Cassidy, L., et al. (2006). Are statewide trauma registries comparable? reaching for a national trauma dataset. *Academic Emergency Medicine*, *13*, 946–953.
- Marr, A. L., & Coronado, V. G., (Eds.), (2004). *Central Nervous System Injury Surveillance Data Submission Standards – 2002*. Atlanta, GA: US Dept Health and Human Services, Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. <https://tbitac.norc.org/download/cdc-data-submission.pdf>. Accessed 1 Nov 2011.
- Marx, B. P., Brailey, K., Proctor, S. P., et al. (2009). Association of time since deployment, combat intensity, and post-traumatic stress symptoms with neuropsychological outcomes following Iraq War deployment. *Archives of General Psychiatry*, *66*, 996–1004.
- McCaffrey, M. A., Mihalik, J. P., Crowell, D. H., et al. (2007). Measurement of head impacts in collegiate football players: clinical measures of concussion after high- and low-magnitude impacts. *Neurosurgery*, *61*, 1236–1243.
- McCarthy, M. L., Dikmen, S. S., Langlois, J. A., et al. (2006). Self-reported psychosocial health among adults with traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *87*, 953–961.
- McKinlay, A., Grace, R. C., Horwood, L. J., et al. (2008). Prevalence of traumatic brain injury among children, adolescents and young adults: prospective evidence from a birth cohort. *Brain Injury*, *22*, 175–181.
- U.S. Medicine (2011). Technology makes for efficient application of new mTBI policy. <http://www.usmedicine.com/articles/technology-makes-for-efficient-application-of-new-mtbi-policy.html>. Accessed 15 June 2011
- Menon, D. K., Schwab, K., Wright, D. W., et al. (2010). Position statement: definition of traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *91*, 1637–1640.
- National Center for Health Statistics (NCHS) (2011). Surveys and data collection systems. <http://www.cdc.gov/nchs/surveys.htm>. Accessed 12 Apr 2011.
- National Center for Injury Prevention and Control (2003). Report to Congress on Mild Traumatic Brain Injury in the United States: Steps to Prevent a Serious Public Health Problem. Atlanta, GA: Centers for Disease Control and Prevention. <http://www.cdc.gov/ncipc/pub-res/mtbi/report.htm>. Accessed 16 Mar 2011.
- National Highway Traffic Safety Administration (2010). Children injured in motor vehicle traffic crashes. <http://www-nrd.nhtsa.dot.gov/Pubs/811325.pdf>. Accessed 17 Mar 2011.
- National Highway Traffic Safety Administration (2011). Fatality Analysis Reporting System. <http://www.nhtsa.gov/FARS>. Accessed 13 Apr 2011.
- Nicaj, L., Stayton, C., Mandel-Ricci, J., et al. (2009). Bicyclist fatalities in New York City: 1996–2005. *Traffic Injury Prevention*, *10*, 157–161.
- Niogi, S. N., Mukherjee, P., Ghajar, J., et al. (2008a). Structural dissociation of attentional control and memory in adults with and without mild traumatic brain injury. *Brain*, *131*, 3209–3221.
- Niogi, S. N., Mukherjee, P., Ghajar, J., et al. (2008b). Extent of microstructural white matter injury in postconcussive syndrome correlates with impaired cognitive reaction time: a 3T diffusion tensor imaging study of mild traumatic brain injury. *American Journal of Neuroradiology*, *29*, 967–973.
- Olson-Madden, J. H., Brenner, L., Harwood, J. E., et al. (2010). Traumatic brain injury and psychiatric diagnoses in veterans seeking outpatient substance abuse treatment. *Journal of Head Trauma Rehabilitation*, *25*, 470–479.
- Orman, J. A. L., Kraus, J. F., Zaloshnja, E., et al. (2011). Epidemiology. In J. M. Silver, T. W. McAllister, & S. C. Yudofsky (Eds.), *Textbook of traumatic brain injury* (2nd ed.). Washington, DC: American Psychiatric Publishing.
- Patzkowski, J. C., Cross, J. D., Ficke, J. R., et al. (2011). The changing face of army disability: the Operation Enduring Freedom and Operation Iraqi Freedom effect. *Journal of the American Academy of Orthopaedic Surgeons*. (in press).

- Pickelsimer, E. E., Selassie, A. W., Gu, J. K., et al. (2006). A population-based outcomes study of persons hospitalized with traumatic brain injury: operations of the South Carolina traumatic brain injury follow-up registry. *Journal of Head Trauma Rehabilitation, 21*, 491–504.
- Pickelsimer, E. E., Selassie, A. W., Sample, P. L., et al. (2007). Unmet service needs of persons with traumatic brain injury. *Journal of Head Trauma Rehabilitation, 22*, 1–13.
- Pietrzak, R. H., Johnson, D. C., Goldstein, M. B., et al. (2009). Posttraumatic stress disorder mediates the relationship between mild traumatic brain injury and health and psychosocial functioning in veterans of Operations Enduring Freedom and Iraqi Freedom. *Journal of Nervous and Mental Disease, 197*, 748–753.
- Polusny, M. A., Kehle, S. M., Nelson, N. W., et al. (2011). Longitudinal effects of mild traumatic brain injury and posttraumatic stress disorder comorbidity on postdeployment outcomes in National Guard soldiers deployed to Iraq. *Archives of General Psychiatry, 68*, 79–89.
- Powell, J. M., Ferraro, J. V., Dikmen, S. S., et al. (2008). Accuracy of mild traumatic brain injury diagnosis. *Archives of Physical Medicine and Rehabilitation, 89*, 1550–1555.
- Puvanachandra, P., & Hyder, A. A. (2009). The burden of traumatic brain injury in Asia: a call for research. *Pakistan Journal of Neurological Sciences, 4*, 27–32.
- Ropper, A. H., Adams, R. D., Samuels, M. A. & Victor, M., (2009). Craniocerebral trauma. In A. H. Ropper, M. A. Samuels (Eds.), *Adams and Victor's principles of neurology*, Ninth Edition, New York: McGraw-Hill.
- Russo, C. A., & Steiner, C., (2007). Statistical brief #27: Hospital admissions for traumatic brain injuries, 2004. Healthcare Cost and Utilization Project (HCUP), March. Agency for Healthcare Research and Quality, Rockville, MD. <http://www.hcup-us.ahrq.gov/reports/statbriefs/sb27.pdf>. Accessed 15 June 2011.
- Saatman, K. E., Duhaim, A. C., Bullock, R., et al. (2008). Classification of traumatic brain injury for targeted therapies. *Journal of Neurotrauma, 25*, 719–738.
- Saunders, L. L., Selassie, A. W., Hill, E. G., et al. (2009). A population-based study of repetitive traumatic brain injury among persons with traumatic brain injury. *Brain Injury, 23*, 866–872.
- Schneiderman, A. I., Braver, E. R., & Kang, H. K. (2008). Understanding sequelae of injury mechanisms and mild traumatic brain injury incurred during the conflicts in Iraq and Afghanistan: persistent postconcussive symptoms and posttraumatic stress disorder. *American Journal of Epidemiology, 167*, 1446–1452.
- Schootman, M., & Fuortes, L. J. (2000). Ambulatory care for traumatic brain injuries in the U.S., 1995–1997. *Brain Injury, 14*, 373–381.
- Selassie, A. W., McCarthy, M. L., Ferguson, P. L., et al. (2005). Risk of post-hospitalization mortality among persons with traumatic brain injury, South Carolina 1999–2001. *Journal of Head Trauma Rehabilitation, 20*, 257–269.
- Selassie, A. W., Pickelsimer, E. E., Frazier, L., Jr., et al. (2004). The effect of insurance status, race, and gender on emergency department disposition of persons with traumatic brain injury. *American Journal of Emergency Medicine, 22*, 465–473.
- Selassie, A. W., Zaloshnja, E., Langlois, J. A., et al. (2008). Incidence of long-term disability following traumatic brain injury hospitalization, United States, 2003. *Journal of Head Trauma Rehabilitation, 23*, 123–131.
- Shavelle, R. M., Strauss, D. J., Day, S. M., et al. (2006). Life expectancy. In N. D. Zasler, D. I. Katz, & R. D. Zafonte (Eds.), *Brain injury medicine: principles and practice* (pp. 247–61). New York: Demos.
- Sperry, J. L., Gentilello, L. M., Minei, J. P., et al. (2006). Waiting for the patient to “sober up”: effect of alcohol intoxication on Glasgow Coma Scale score of brain injured patients. *Journal of Trauma, 61*, 1305–1311.
- Stuke, L., Diaz-Arrastia, R., Gentilello, L. M., et al. (2007). Effect of alcohol on Glasgow Coma Scale in head-injured patients. *Annals of Surgery, 245*, 651–655.
- Svetlov, S. I., Larner, S. F., Kirk, D. R., et al. (2009). Biomarkers of blast-induced neurotrauma: profiling molecular and cellular mechanisms. *Journal of Neurotrauma, 26*, 913–921.
- Swaine, B. R., Tremblay, C., Platt, R. W., et al. (2007). Previous head injury is a risk factor for subsequent head injury in children: a longitudinal cohort study. *Pediatrics, 119*, 749–758.
- Tagliaferri, F., Compagnone, C., Korsi, M., et al. (2006). A systematic review of brain injury epidemiology in Europe. *Acta Neurochirurgica (Wien), 148*, 255–268.
- Tanielian, T., & Jaycox, L. H. (Eds.). (2008). *Invisible wounds of war: psychological and cognitive injuries, their consequences, and services to assist recovery*. Santa Monica: Rand Corporation.
- Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness. a practical scale. *Lancet, 2*, 281–284.
- Terrio, H., Brenner, L. A., Ivins, B. J., et al. (2009). Traumatic brain injury screening: preliminary findings in a US Army Brigade Combat Team. *Journal of Head Trauma Rehabilitation, 24*, 14–23.
- Thurman, D. J., Branche, C. M., & Sniezek, J. E. (1998). The epidemiology of sports-related traumatic brain injuries in the United States: recent developments. *Journal of Head Trauma Rehabilitation, 13*, 1–8.
- Thurman, D. J., Coronado, V., Selassie, A., (2006). In N. D. Zasler, D. I. Katz, R. D. Zafonte (Eds.), *Brain injury medicine: principles and practice*. (pp. 45–55) New York: Demos.
- Thurman, D. J., & Guerrero, J. (1999). Trends in hospitalization associated with traumatic brain injury. *Journal of the American Medical Association, 282*, 989–991.

- Thurman, D. J., Kraus, J. F., Romer, C. J. (1995b). Standards for Surveillance of Neurotrauma. World Health Organization Safety Promotion and Injury Control. http://www.who.int/violence_injury_prevention/publications/surveillance/neurotrauma/en/index.html. Accessed 15 June 2011.
- Thurman, D. J., Sniezek, J. E., Johnson, D., et al. (1995). *Guidelines for surveillance of central nervous system injury*. Atlanta: Centers for Disease Control and Prevention.
- Timonen, M., Miettunen, J., Hakko, H., et al. (2002). The association of preceding traumatic brain injury with mental disorders, alcoholism and criminality: the Northern Finland 1966 Birth Cohort Study. *Journal of Psychiatric Research*, *113*, 217–226.
- Tricare (Department of Defense) (2009). Appendix G: Special Guidance on Traumatic Brain Injury Coding. http://www.tricare.mil/ocfo/_docs/APPENDIX_G_2009_06_11.doc. Accessed 12 Apr 2011.
- Ventura, T., Harrison-Felix, C., Carlson, N., et al. (2010). Mortality after discharge from acute care hospitalization with traumatic brain injury: a population-based study. *Archives of Physical Medicine and Rehabilitation*, *91*, 20–29.
- Vos, P. E., Jacobs, B., Andriessen, T. M., et al. (2010). GFAP and S100B are biomarkers of traumatic brain injury: an observational cohort study. *Neurology*, *16*, 1786–1793.
- Walker, R., Cole, J. E., Logan, T. K., et al. (2007). Screening substance abuse treatment clients for traumatic brain injury: prevalence and characteristics. *Journal of Head Trauma Rehabilitation*, *22*, 360–367.
- Weiss, H., Agimi, Y., & Steiner, C. (2010). Youth motorcycle-related brain injury by state helmet law type: United States, 2005–2007. *Pediatrics*, *126*, 1149–1155.
- Wilde, E. A., McCauley, S. R., Hunger, J. V., et al. (2008). Diffusion tensor imaging of acute mild traumatic brain injury in adolescents. *Neurology*, *70*, 948–955.
- Wojcik, B. E., Stein, C. R., & Bagg, K. (2010). Traumatic brain injury hospitalizations of US Army soldiers deployed to Afghanistan and Iraq. *American Journal of Preventive Medicine*, *38*(1S), S108–S116.
- Wojcik, B. E., Stein, C. R., Orosco, J., et al. (2010). Creation of an expanded Baresell matrix to identify traumatic brain injuries of U.S. military members. *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, *7*, 157–166.
- World Health Organization (WHO) (1979). Medical Certification of Cause of Death. <http://whqlibdoc.who.int/publications/9241560622.pdf>. Accessed 12 Apr 2011
- World Health Organization (WHO) (2007). International Classification of Diseases, 10th Revision (ICD-10). <http://www.who.int/classifications/icd/en/> Accessed 12 Apr 2011.
- World Health Organization (WHO) (2011). International Classification of Diseases (ICD). <http://www.who.int/classifications/icd/en/> as in the United States. Accessed 4 Apr 2011.
- Xiang, H., Sinclair, S. A., Yu, S., et al. (2007). Case ascertainment in pediatric traumatic brain injury: challenges in using the NEISS. *Brain Injury*, *21*, 293–299.
- Zaloshnja, E., Miller, T., Langlois, J. A., et al. (2008). Prevalence of long-term disability from traumatic brain injury in the civilian population of the United States, 2005. *Journal of Head Trauma Rehabilitation*, *23*, 394–400.

Part II

Injury Causation

Chapter 5

Forensic Pathology

Ling Li

Introduction

There are two general types of medicolegal death investigation systems in the USA: the coroner system and medical examiner system.

The Coroner System

The historical development of the coroner system can be traced back to feudal England. The coroners were formalized into law in the twelfth century under King Richard I (Richard the Lion-hearted). The King dispatched coroners to a death scene to protect the crown's interest and collect duties (*coroner* is derived from Anglo-Norman *cornouner*, the “keepers of the crown's pleas”) (Platt 1993; Hanzlick 2003). There was little development of the coroner system in England until the middle of the nineteenth century. In 1877, a law was enacted requiring an inquest to be conducted whenever the coroner had reasonable cause to suspect violent or unnatural death or when the cause of death was unknown (Platt 1993). The modern coroner in England is usually a lawyer but may also be a doctor. Some of the coroners may also have legal qualifications. The coroner is employed by local government but functions under the Coroner's Acts and Roles laid down by Parliament. His basic function is to investigate all deaths that cannot be satisfactorily certified by physicians in the usual way.

The early American colonists, originating from England, brought the coroner system into the colonies in the early 1600s. Currently, coroners in the USA are usually elected officials (rather than appointed) in their jurisdictions and usually are not required to have any medical qualifications. Coroners must rely on pathologists (coroner's pathologists) to assist in death investigations and to conduct postmortem examinations. The coroner makes rulings as to cause and manner of death in cases that fall under the coroner law.

L. Li, MD (✉)

Office of the Chief Medical Examiner, State of Maryland, 111 Penn Street, Baltimore, MD 21201, USA
e-mail: ling001@aol.com

The Medical Examiner System

The first move toward reliance on a medical examiner took place in 1860 with the passage of Maryland legislation requiring the presence of a physician at the death inquest. In 1868, the legislature authorized the governor to appoint a physician as sole coroner for the city of Baltimore. In 1877 in Massachusetts, the Commonwealth adopted a statewide system designating a physician known as a medical examiner to determine the cause and manner of death (Platt 1993; Hanzlick 2003; DiMiao and DiMiao 2001). In 1915, New York City adopted a law eliminating the coroner's office and creating a medical examiner system. It was not until 1918 that New York City formed the first true medical examiner's office (The Office of the Chief Medical Examiner of the City of New York 1967). In 1939, the state of Maryland established the first formal statewide medical examiner system that covered all but one county of the state, which came under the system 2 years later. Medical examiners are usually appointed and are, with few exemptions, required to be licensed physicians and often pathologists or forensic pathologists.

Perhaps Death Investigation systems?

Death investigation systems are usually established on a statewide, regional or district, or county level. Each system is administrated by a medical examiner or coroner or someone such as a sheriff or justice of the peace acting in that capacity under provision of state law (Hanzlick 2006). As of 2003, 11 states have coroner-only systems, wherein each county in the state is served by a coroner; 22 states have medical examiner systems, most of which are statewide and are administered by state agencies; and 18 states have mixed systems: some counties are served by coroners, others by medical examiners (Hanzlick 2003).

Approximately 20% of the 2.4 million deaths in the USA each year are investigated by medical examiners and coroners, accounting for approximately 450,000 medicolegal death investigations annually (Hanzlick 2003).

The categories of medicolegal cases include the following:

1. Violent deaths, i.e., homicide, suicide, and accident
2. Sudden unexpected deaths
3. Deaths without physician attendance
4. Deaths under suspicious circumstances, i.e., those that may be due to violence
5. Deaths in police custody
6. Deaths related to therapeutic misadventure, i.e., medical malpractice

The objectives of medicolegal death investigation are as follows:

1. To determine the cause and manner of death
2. To determine the primary, secondary, and contributory factors in the cause of death when trauma and disease are present simultaneously
3. To make identification of the decedent, if unknown
4. To estimate the time of death and injury
5. To interpret how the injury occurred and the nature of the weapon used
6. To collect evidence from the bodies that may be used in criminal law cases
7. To provide medicolegal documents and expert testimony in criminal and civil law cases if the case goes to trial

Table 5.1 Example of cause of death

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. Acute Upper Gastrointestinal Bleeding Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Ruptured esophageal varices Due to (or as a consequence of)
	c. Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	d. Cirrhosis of Liver
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	Chronic alcoholism, Hepatitis B

Forensic Pathology

Forensic pathology is a branch of medicine that deals with the study of the cause and manner of death by examination of a dead body during the medicolegal investigation of criminal law and civil law cases in some jurisdictions.

Cause of Death

The cause of death is any disease or injury that is responsible for producing a physiological derangement in the body that results in the death of the individual. A competent cause of death is etiologically specific. “But for” this or that particular underlying event, the individual would not have died (Godwin 2005). There are primary (underlying or proximate) cause of death, immediate cause(s) of death, and intermediate cause(s) of death. The primary (underlying or proximate) cause of death is the disease or injury that initiated events resulting in death and without which death would not have occurred. The immediate cause(s) of death is (are) final complications and sequelae of the primary cause or last event resulting in death. Intermediate causes of death are diseases or conditions that contribute to death and are a result of the primary cause. Table 5.1 shows an example of immediate, intermediate, and primary cause of death (“Part I”). Other significant conditions are coexisting or preexisting disease(s)/condition(s) that contributed to death but did not result in the underlying cause (“Part II”).

Manner of Death

The manner of death is a description of the circumstances surrounding death and explains how the cause of death came about. In general, there are five manners of death: natural, accident, suicide, homicide, or undetermined (or “could not be determined”). There are basic, general “rules” for classification of manner of death by the medical examiners and coroners (National Association of Medical Examiners 2002):

- Natural deaths are caused solely or nearly totally by disease and/or the aging process.
- Accidental deaths are defined as those that are caused by unintentional injury or poisoning (when there is little or no evidence that the injury or poisoning occurred with intent to harm or cause death).
- Suicide results from an intentional, self-inflicted injury or poisoning (an act committed to do self-harm or cause the death of one’s self).

Table 5.2 Example of a natural death

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. Acute Pulmonary Thromboembolism Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Leg Deep Vein Thrombosis Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	c. Recent Gastric Bypass Surgery Due to (or as a consequence of)
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	d. Obesity
	Emphysema
	Manner of Death
× Natural <input type="checkbox"/> Accident <input type="checkbox"/> Suicide <input type="checkbox"/> Homicide <input type="checkbox"/> Undetermined	

Table 5.3 Example of an accidental death

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. Acute Pulmonary Thromboembolism Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Right Leg Deep Vein Thrombosis Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	c. Due to (or as a consequence of)
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	d. Fracture of Right Ankle while Playing Football
	Manner of Death
<input type="checkbox"/> Natural × Accident <input type="checkbox"/> Suicide <input type="checkbox"/> Homicide <input type="checkbox"/> Undetermined	

- Homicide results from injury or poisoning due to an act committed by another person to do harm, or cause fear or death.
- Undetermined or “could not be determined” is a classification used when there is insufficient information pointing to one manner of death that is more compelling than one or more other compelling manners of death, or, in some instances, when the cause of death is unknown.

The following are several examples of causes of death and manner of death certification.

Case 1 was a 45-year-old female who suddenly had shortness of breath and collapsed at home. She was pronounced dead on arrival at the hospital. She was obese and had undergone a gastric bypass surgery 6 days before her collapse. She also had a history of emphysema and had been smoking for more than 20 years. There was no history of injury. Autopsy showed that she weighed 300 lbs. Examination of the lungs showed a saddle occlusive pulmonary thromboembolus. Dissection of her legs revealed deep vein thrombosis (Table 5.2).

Case 2 was a 30-year-old male who was found dead in bed. Two weeks prior to his death, he fractured his right ankle while playing baseball. He had surgical repair of his right ankle and was wearing a hard cast. He had no other medical history. Autopsy examination revealed that he died of massive pulmonary thromboemboli due to right leg deep vein thrombosis. Since the fracture of the right ankle was the underlying cause of the pulmonary thromboemboli, the manner of death was ruled an accident (Table 5.3).

Table 5.4 Example of a suicide

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. Asphyxia Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	c. Hanging Due to (or as a consequence of)
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	Depression
	Manner of Death
<input type="checkbox"/> Natural <input type="checkbox"/> Accident <input checked="" type="checkbox"/> Suicide <input type="checkbox"/> Homicide <input type="checkbox"/> Undetermined	

Table 5.5 Example of a homicide

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. Pneumonia Complicated by Sepsis Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Quadriplegia Due to (or as a consequence of)
	c. Cervical Vertebral Fracture with Spinal Cord Transection Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	d. Gunshot Wound of Neck
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	
	Manner of Death
<input type="checkbox"/> Natural <input type="checkbox"/> Accident <input type="checkbox"/> Suicide <input checked="" type="checkbox"/> Homicide <input type="checkbox"/> Undetermined	

Case 3 was a 19-year-old male college student who was reportedly found unresponsive on the floor of his bedroom by his father. Resuscitation was performed at the scene but was unsuccessful. According to his father, he did not have any medical history. Autopsy examination revealed that there was a faint ligature mark around his neck. The ligature mark extended upward across both sides of the neck and became indistinct behind his right ear. There were bilateral conjunctival petechial hemorrhages noted. Postmortem toxicology analysis was positive for amitriptyline (an antidepressant medication). Further investigation revealed that he had been depressed since his girlfriend broke up with him 1 year ago. Later, his father stated that he found him hanging from his bunk bed with a bed sheet around his neck. The father cut the sheet and cleaned up the scene before the medical personnel arrived. He died of asphyxia due to hanging (Table 5.4).

Case 4 was a 59-year-old male who had been robbed and shot 20 years ago. He became quadriplegic due to a fracture of the second cervical vertebra and transection of the underlying cervical spinal cord. He was bedridden and had been in a nursing home ever since the shooting. He developed multiple episodes of pneumonia and urinary tract infection during the course of his care. He died of sepsis and pneumonia. His immediate cause of death was an infectious disease. However, the underlying cause that initiated the events resulting in his death was a gunshot wound to the neck. Although the shooting occurred 20 years earlier, the manner of death is still homicide (Table 5.5).

Table 5.6 Example of an undetermined cause of death

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. No Anatomic Cause of Death Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Due to (or as a consequence of)
	c. Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	d.
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	Manner of Death
<input type="checkbox"/> Natural <input type="checkbox"/> Accident <input type="checkbox"/> Suicide <input type="checkbox"/> Homicide <input checked="" type="checkbox"/> Undetermined	

Table 5.7 Example of another undetermined cause of death

	Cause of Death
Part I	Disease, injury, or complications that directly caused the death
Immediate Cause (Final disease or condition resulting in death)	a. Oxycodone and Morphine Intoxication Due to (or as a consequence of)
Intermediate Cause (diseases or conditions that contribute to death and are a result of the primary cause)	b. Due to (or as a consequence of)
	c. Due to (or as a consequence of)
Primary (underlying) Cause (disease or injury that initiated events resulting in death)	d.
Part II. Other significant conditions contributing to death but not resulting in the underlying cause in Part I	Manner of Death
<input type="checkbox"/> Natural <input type="checkbox"/> Accident <input type="checkbox"/> Suicide <input type="checkbox"/> Homicide <input checked="" type="checkbox"/> Undetermined	

Case 5 was a body found in the woods by a jogger. The severely decomposed and partially skeletonized body was that of a male clad in blue jeans with brand name “Back DAD,” striped boxer shorts, white socks, and white “Reebok” running shoes. Physical characteristics of the remains suggested that this was a middle-aged male in his late 30s to early 40s. His head was largely skeletonized with a small segment of dried, parchment-like soft tissue adhering to the left side of the calvarium. The neck was completely skeletonized. The hyoid bone and thyroid cartilage were missing. The rest of the body was partially skeletonized with severe decomposition of the attached soft tissues. There was no evidence of trauma on the remains. Further police investigation revealed that the physical characteristics of the skeletal remains matched characteristics of a missing person. He was identified based on a general description and a dental comparison as a 38-year-old African American male who had been missing for more than 7 months. Postmortem examination failed to reveal an anatomic cause of death. The advanced decomposition and skeletonization precluded relevant postmortem toxicological analysis. Therefore, the manner of death is certified as Undetermined (Table 5.6 and 5.7).

Case 6 was a 34-year-old woman who was found unresponsive in bed by her husband. She had a history of prescription drug abuse and was on pain medication because of back pain. According to her husband, she was also depressed and had attempted suicide by overdose 3 months prior to her death. Postmortem examination revealed no evidence of trauma or significant natural diseases.

Toxicology analysis revealed 1.2 mg/L oxycodone, 0.4 mg/L citalopram, and 160 mg/L morphine in the blood. She died of combined oxycodone and morphine intoxication. The manner of death was classified as undetermined because it cannot be ascertained if this is a case of suicide overdose or an accident in which she inadvertently took too much of her medication.

In summary, it is important to recognize that autopsy alone rarely divulges the manner of death (Godwin 2005). Determination of the manner of death depends upon the known facts concerning the circumstances of death by investigation and in conjunction with the findings at autopsy, including toxicological analyses.

Postmortem Toxicological Analysis

Did a drug or chemical substance play any role in the death under investigation? This question must be raised in every medicolegal death investigation. Reaching a correct conclusion requires collaboration of forensic pathology and forensic toxicology. Forensic toxicology evaluates the role of drugs and/or chemicals as a determinant or contributory factor in the cause and manner of death. Death caused by poisoning cannot be certain without toxicological analysis that demonstrates the presence of the poison in the deceased's tissues or body fluids. Autopsy findings in poisoning deaths are usually nonspecific, and the diagnosis is usually reached by toxicological analysis determined by circumstances elucidated during death investigation. Many times, the history suggests a particular drug or chemical substance may be involved, and the laboratory is requested to determine the presence or absence of that drug or chemical. Sometimes, a forensic pathologist may require toxicological analysis of certain prescribed medications, such as drugs to control seizures in the case of sudden death of a patient with a history of epilepsy. Most deaths from seizures occur without anatomic findings. Negative or low concentration of the antiseizure medications may explain the cause of sudden unexpected death in epilepsy. In other cases, where death is not due to poisoning, the forensic toxicologists are often able to provide valuable evidence concerning the circumstances surrounding a death. The presence of a high concentration of alcohol in the blood or tissues may be used to explain the erratic driving behavior of the victim of an automobile accident.

A poisoning death usually is first suspected because of information from scene investigation and the decedent's history. In cases where the history is suggestive of poisoning death, the following steps must be taken: (a) a thorough scene investigation and review of clinical history, (b) a complete autopsy examination, and (c) postmortem toxicological analysis.

At the scene, investigators should perform a systematic gathering of evidence, including (a) identification of the victim – age, gender, occupation, social class; (b) documentation of the environment and surroundings of the decedent's body; (c) collection of physical and biological evidence, such as drug paraphernalia (syringes and spoon cookers), empty medication bottles, open household products, and suspicious liquids and powders, suicide note, and body fluids (vomitus, urine, feces, or blood); (d) interview of witnesses, family members, and friends in regard to the decedent's recent activities, medical, social, and psychological problems; and (e) obtaining a clinical history from the decedent's doctors or from the hospital if the decedent had sought medical care.

The major functions of the autopsy are to exclude other obvious causes of death and to collect the appropriate specimens for toxicological analysis. It must be emphasized that with the possible exception of corrosive poisons, the autopsy findings are rarely conclusive. The majority of drug-related deaths show no specific findings at autopsy.

Toxicology specimens collected at the scene and at autopsy are the most important physical evidence in the investigation of any suspected poisoning deaths. The items and specimens collected must be protected by a chain of custody documentation to maintain medical and legal probity. Each specimen must be labeled with a unique identification number, the victim's name, the date and time

of the collection, as well as the source of the specimen. At autopsy in a suspected case of poisoning, samples to be collected and sent for toxicological analysis include blood, urine, bile, vitreous humor, stomach contents, liver, and kidneys. Lung tissue is useful when volatile substances are suspected. If metal poisoning is suspected, bone, nails, and hair are useful for detecting chronic poisoning. Muscle is of a great value in decomposed bodies.

Common Types of Injuries Associated with Deaths

Blunt Force Injury

Blunt force injury refers to a type of physical trauma inflicted either by forceful impacts of blunt objects, such as rods, hammers, baseball bats, fists, and the like to a body part, or by forceful contact of part of or the entire body against an unyielding surface, e.g., during a car accident when occupants are thrown forward against the steering wheel, dashboard, or the back of the seats, or from falls in which the head or trunk strikes the floor or pavement. The major types of blunt force injuries include abrasions, contusions, lacerations, and skeletal fractures.

Abrasion

An abrasion is a scraping and removal of the superficial layers (epidermis and dermis) of the skin. Abrasions usually are caused by the frictional force of scraping along a rough surface, as when a pedestrian is dragged over the pavement (Fig. 5.1) or in a fall. Abrasions can also be caused by localized force rubbing against the skin, e.g., in the case of hanging (Fig. 5.2a, b) or strangulation. A scratch is a special type of abrasion that is inflicted with a relatively sharp and pointed object.



Fig. 5.1 Pedestrian who was struck by a motor vehicle and received scraping (“brush burn”) abrasions from scraping along the pavement



Fig. 5.2 Hanging from construction scaffolding with a rope around the neck (a). Note ligature abrasion furrow over the front of the neck and extended upward to the back of the neck (b)



Fig. 5.3 Note multiple contusions with focal linear abrasions on the back of both forearms and hands of a 17-year-old woman who was killed by her boyfriend and died of multiple blunt force injuries

Contusion

A contusion (bruise) is an area of hemorrhage into the dermis, subcutaneous tissues, deep soft tissues, or internal organs, e.g., the brain, heart, lungs, or the liver due to rupture of blood vessels caused by impact with a blunt object. The hemorrhage may be limited and merely diffuse into the deep soft tissues (Fig. 5.3), or it may be massive with a large collection of blood (hematoma) in the area of the contusion. Contusions of the internal organs are usually caused by severe blunt force impact to the body, e.g., in motor vehicle accidents.

Fig. 5.4 A 17-year-old woman was found lying on the living room floor of her residence with multiple blunt injuries to her body. Note multiple linear and curved lacerations of the scalp



Lacerations

A laceration is a tear of skin, mucosa, visceral surfaces, or parenchyma as a result of crushing or stretching of tissues by the impact of blunt force. In general, a laceration possesses the following characteristics: (a) linear, stellate, curved, or angled; (b) ragged and irregular margins of the wound; and (c) multiple threads of nerves, small blood vessels, and connective tissues bridging the gap between opposing sides of the wound. Lacerations are usually seen in the skin over bony areas such as the scalp covering the skull (Fig. 5.4), the skin of the eyebrow, or the skin covering the cheek, chin, elbow, and knee. Owing to the anatomical structure and location, large blood vessels and internal organs can be lacerated if excessive blunt force is applied.

Skeletal Fractures

A skeletal fracture is a break in a bone. A fracture usually results from traumatic injury to bones when the application of force is sufficient to cause disruption of the continuity of bone tissues. The common locations and types of fractures encountered by the practicing forensic pathologist include (a) linear skull fractures, which usually occur when the head strikes a flat surface, such as a fall on the floor or when the head is thrown against a wall, resulting in a fractured skull with fracture lines radiating from the point of impact; (b) depressed skull fractures, commonly caused by localized forceful impact with a fairly small but heavy object, such as a hammer or a rock, or by a fall on a sharp corner of a piece of furniture; (c) basal skull fractures, usually caused by impact on either side of the head or as a result of impact on the face, forehead, or chin. Depending upon the direction and location of the impacting force, the fractures can be longitudinal (front-to-back), transverse (side-to-side), or ring shape; (d) rib fractures, commonly seen in transportation fatalities and in cases of child abuse; and (e) fractures of extremities and spinal vertebrae, which usually occur as a result of a fall or crash.

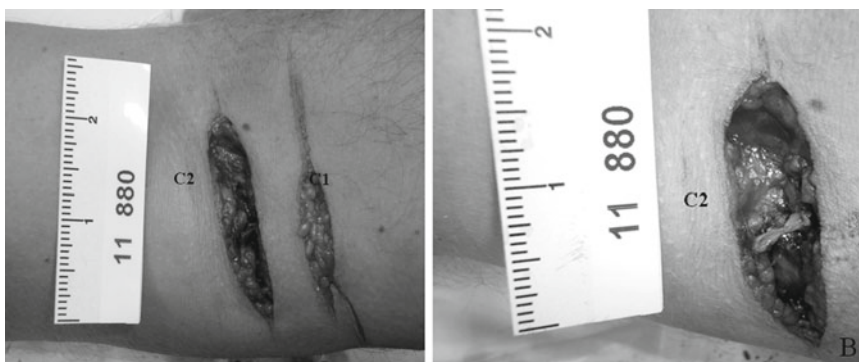


Fig. 5.5 A 45-year-old man was found unresponsive halfway down a hill lying in the snow. The snow surrounding him was saturated with blood. A blood-stained single-edged knife was recovered in the snow near his body. His vehicle was parked about half a mile up the hill. At autopsy, there were two cutting wounds noted on the anterior aspect of his left forearm (a). Note the cutting wound (C1) was superficial and cut through the skin and superficial layer of fatty tissues only. There were “hesitation marks” extending from each end of both cutting wounds. Note the cutting wound (C2) showing the partial severing of an artery (b)

Deaths resulting from blunt force injuries occur in a variety of situations. Blunt force trauma is the most common cause of accidental death involving motor vehicle collisions, pedestrians being struck by vehicles, airplane crashes, falling from heights, and boating incidents (Batalis 2010). Although firearms are by far the most common means of homicide in the USA, blunt force trauma, especially blunt force head injury, is the most common cause of death in child-abuse-related homicide (Collins and Nichols 1999; Lee and Lathrop 2010). Suicide by self-inflicted blunt force injuries is rare (Hunsaker and Thorne 2002). The common causes of blunt force injuries in cases of suicide include jumping from heights and suicide by trains.

Sharp Force Injury

Sharp force injury is a type of wound caused by pointed and sharp-edged instruments such as knives, daggers, glass, and razor blades. A distinctive characteristic of sharp force injury is a relatively well-defined traumatic separation of injured tissues with the absence of threads of nerves, small blood vessels, and connective tissues bridging the gap between opposing sides of the wound. There are three specific types of sharp force injuries: incised wounds, stab wounds, and chop wounds.

Incised Wounds

An incised wound (or cut) occurs when a pointed and sharp-edged instrument is drawn along the surface of the skin with sufficient pressure, producing a wound whose length on the skin is greater than its depth in the underlying tissues. The incised wounds can be linear, curved, or angled and have relatively sharply delineated edges.

Incised wounds can be suicidal, homicidal, and accidental. Suicidal incised wounds are often inflicted on the upper extremities, such as the wrist and antecubital fossa, followed by the neck and chest (Karger et al. 2000; Fukube et al. 2008). In self-inflicted incised wounds, one will usually note the presence of “hesitation cuts” or “hesitation marks.” These “hesitation marks” are a group of superficial, roughly parallel incised wounds, typically present on the palmar aspect of the wrists, adjacent to or overlying the fatal incised wound in suicide victims (Fig. 5.5a, b). Homicide by isolated incised



Fig. 5.6 Stab wound by single-edged knife. Note a sharp inferior end and blunt superior end

wounds is uncommon and usually associated with stab wounds (Brunel et al. 2010). Incised wounds of accidental origin that lead to death are rare and occur when an individual falls on glass materials or is struck by a flying fragment of glass or some other sharp-edged projectile in the neck, trunk, or head where there is a blood vessel large enough to give rise to rapidly fatal bleeding (DiMiao and DiMiao 2001; Demirci et al. 2008; Mason and Purdue 2000; Karger et al. 2001; Prahlow et al. 2001).

Stab Wounds

A stab wound results when a sharp-edged instrument is forced into the skin and the underlying tissues, producing a wound that is deeper in the body than its length on the skin. Knives are the most common weapon used to inflict stab wounds. Other instruments that can cause stab wounds include scissors, forks, screwdrivers, arrows, ice picks, and any other cylindrical object that has a sharp or pointed tip (Prahlow 2010).

The size and shape of a stab wound on the skin depends on the type of the weapon, the location and orientation of the wound in the body, the movement of the victim, the movement of the weapon in the wound, and the angle of the weapon withdrawal. Stab wounds from a single-edged blade typically have a sharp end and a blunt end (Fig. 5.6) but may also have two sharp ends if the blade penetrates the skin at an oblique angle. Thus, only the sharp edge of the blade cuts through the skin and the squared-off back does not contact the skin. Stab wounds with the same knife may appear variably slit-like if the wounds are parallel to the elastic fibers (also known as Langer's lines) in the dermis of the skin, or widely gaping if the wounds are perpendicular to or oblique to the elastic fibers. Because of the effect of the elastic fibers, the edges of a gaping wound should be reapproximated when measuring the size of the wound.

Stab wounds caused by scissors or screwdrivers may have characteristic appearances. The shape of stab wounds from scissors depends on whether the scissors are open or closed. If the two blades are closed, one single stab wound will be produced. The wounds will have abraded margins and be more broad than the typical stab wound from a knife with abraded margins because the scissor blades are so much thicker. If the two blades are open, two stab wounds will be produced side-by-side. A stab wound with the appearance of four-point stars is consistent with a Phillips screwdriver due to the X-shaped point of the blade.



Fig. 5.7 Defense wound on the back of the left forearm (a) and the palm of the left hand (b)

The majority of stab wounds are homicidal (DiMiao and DiMiao 2001; Gill and Catanese 2002). Self-inflicted stab wounds are uncommon and frequently accompanied by incised wounds. The distinction between homicide and suicide in sharp force injuries requires the analysis of autopsy findings and a comparison with other results from the death scene investigation. “Defense injuries” may be of value in differentiating between homicide and suicide. “Defense injuries” are incised wounds or stab wounds sustained by victims as they are trying to protect themselves from an assailant. They are usually on the upper extremities, most commonly found on the palm of the hands due to an attempt to grasp the knife or on the back of the forearms and upper arms in an attempt to ward off the knife (Fig. 5.7a, b). It has been reported that approximately 40–50% of homicide stabbing victims had defense wounds (Gill and Catanese 2002; Katkici et al. 1994).

Chop Wounds

A chop wound is a wound caused by a heavy instrument that has at least one sharp cutting edge wielded with a tremendous amount of force. Examples are machete, ax, bush knife, boat propeller, lawn mower blade, and a multitude of industrial and farm machinery. Chop wounds may have

features of both sharp and blunt force injuries due to the combination of cutting and crushing by the heavy thick blade. If the wounds are over bone, there frequently exist underlying comminuted fractures and deep grooves or cuts in the bone.

Firearm Injury

Firearm injury in the USA caused an average of 29,986 deaths annually between 1999 and 2007 (US Centers for Disease Control and Prevention 2011). In 2007, 31,224 people died from firearm injuries in the USA with the age-adjusted death rate of 10.2/100,000 (US Centers for Disease Control and Prevention 2010). Firearms were the third leading cause of death from injury after motor vehicle crashes and poisoning (US Centers for Disease Control and Prevention 2010). In the USA, firearms are the most common method used in homicides, followed by sharp instruments, and then blunt instruments (Karch et al. 2010). Firearm injury represents a significant public health problem, accounting for 6.6% of premature deaths in the USA (US Center for Disease Control and Prevention 2007). Firearm injury disproportionately affects young people, resulting in lives cut short.

Depending on the types of weapons used, firearm injuries include two major types of wounds: gunshot wounds and shotgun wounds.

Gunshot Wounds

Today's gunshot wounds – as opposed to shotgun wounds or those from older smooth-bore firearms – are produced by rifled weapons, such as revolvers, pistols, rifles, and many types of military weapons. The rifled weapons fire one projectile at a time through a barrel that has a series of parallel spiral grooves cut into the length of the bore (the interior) of the barrel. Rifling consists of these grooves and the intervening projections between the grooves called the lands. The purpose of the rifling is to grip the bullet and impart a gyroscopic spin to the bullet along its longitudinal axis as it moves down the barrel, which assists in maintaining an accurate trajectory.

The ammunition for rifled weapons consists of a cartridge case, primer, propellant (gunpowder), and bullet. When the firing pin of the weapon strikes the primer, it detonates the primer. This in turn ignites the propellant. The propellant burns rapidly, producing huge volumes of gas. The pressure of the gas pushes the bullet down the barrel. The materials that exit from the end of the barrel are as follows: the bullet, gas produced by combustion of the gunpowder, soot produced by the burning of the gunpowder, partially burnt and unburnt gunpowder particles, and vaporized metal from primer, cartridge case, and bullet.

Gunshot wounds can be classified into three categories based on the range of fire: (a) contact wounds, (b) intermediate wounds, and (c) distant wounds.

Contact Wounds

A contact wound is produced when the muzzle of the weapon is held against the surface of the body at the time of discharge. Contact wounds from a rifled weapon usually have a circular or ovoid bullet hole with a surrounding zone of sealed, blackened skin. Soot in varying amounts is also deposited around the bullet hole, depending on how tightly the gun is held against the body. If the weapon is pressed tightly into the skin (tight contact), all the material exiting from the end of the barrel (muzzle) enters the body. The muzzle of the weapon can leave an ecchymosed or abraded muzzle imprint on the skin around the entrance wound (Fig. 5.8). If the muzzle is held loosely against the skin, gas discharged from the weapon can escape from the temporary gap between the end of the muzzle and the skin with the deposition of soot around the entrance wound.

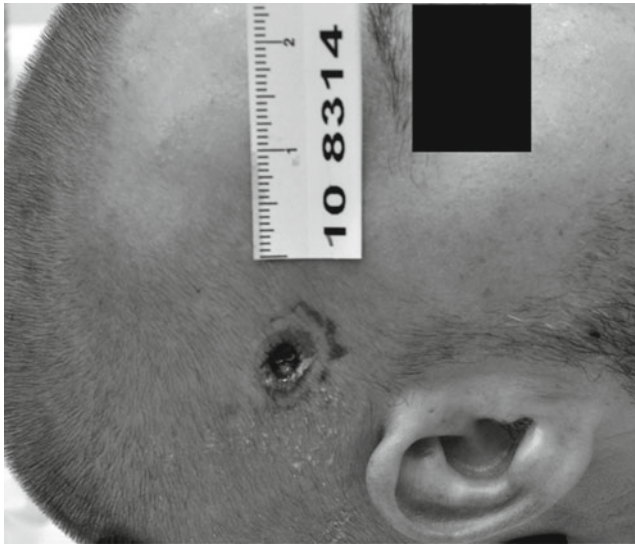


Fig. 5.8 Contact gunshot wound of right temple shows muzzle imprint around the bullet hole and the absence of soot or powder stippling around the entrance wound

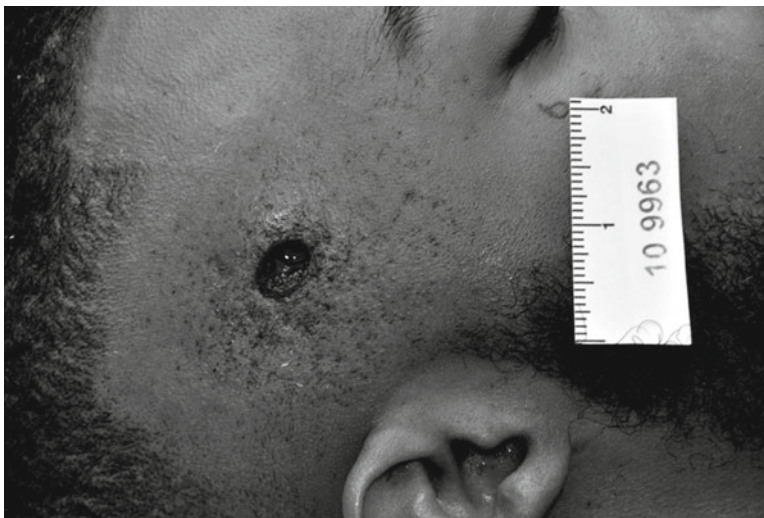


Fig. 5.9 Intermediate shot of the right temple. Note gunpowder stippling scattered over a diameter of 2 in. on the skin around the entrance site, indicating a shot fired from a distance of several inches

Intermediate-range Wounds

An intermediate-range (close-range) gunshot wound is characterized by a central circular/ovoid bullet hole with a margin of abraded skin and presence of gunpowder stippling (tattooing) on the skin around the entrance site. Gunpowder stippling is characterized by reddish-brown punctate abrasions caused by the impact of partially burnt and unburnt gunpowder particles (Fig. 5.9). Stippling is important in that the diameter of the stippling distribution can help determine the range of fire.

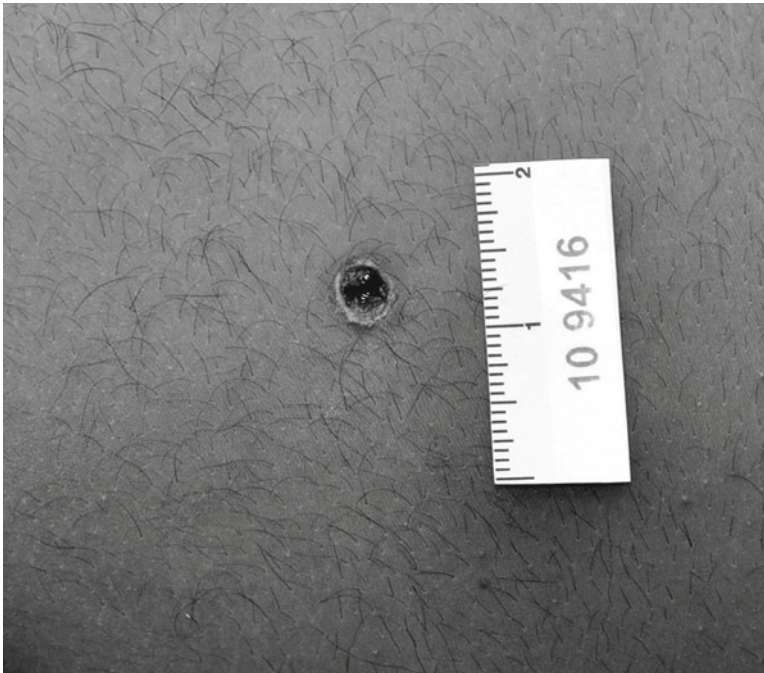


Fig. 5.10 Classic distant entrance wound. Note a round central defect surrounded by a thin margin of abrasion

An intermediate-range gunshot wound is one in which the muzzle of the weapon is away from the body at the time of firing yet is sufficiently close so that powder grains emerging from the muzzle along with the bullet powder result in stippling on the skin. For most handguns, intermediate range is a distance from the muzzle to the body surface at 24 in. (60 cm) to 42 in. (105 cm), depending on the type of weapon and the type of ammunition used (DiMiao 1985). For rifles, this may reach several feet. Soot can also be deposited on the skin around the intermediate-range gunshot wound. In handguns, soot can be identified in shots fired from a distance within 6 in. (Spitz 1993). Increasing the range of fire will increase the distribution area of stippling but decrease the density of the stippling. The patterns of stippling, however, vary widely with different weapons and type of ammunitions. Test firing with the particular weapon, using the same type of ammunition as that used in the shooting case under consideration, is recommended to estimate the range of an intermediate shot.

Distant Wounds

A distant wound is produced when the weapon is fired from a distance at which gun smoke will not reach the target. In other words, soot and gunpowder stippling disappear at the distant entrance wound. Generally, in the case of most handguns, gunpowder stippling disappears at a distant shot beyond 2–3 ft depending again on the weapon and the type of ammunition. A distant entrance wound has a round to ovoid bullet hole with an abraded margin without soot deposition or gunpowder stippling (Fig. 5.10). If a gunshot entrance wound lacks features that define an intermediate-range or contact wound, no distinction with respect of distance can be made between one distant shot and another, e.g., the appearance of a gunshot wound produced from a distance of 5 ft will be the same as one from 10 or 20 ft. In evaluating a gunshot wound, consideration must be given to any intermediary target that may filter out gunpowder or soot. For example, a shot fired through a door or a



Fig. 5.11 Two exit wounds of the back of the head. Note the irregularly shaped wounds with ragged edges and no abraded ring

window from a close range (a few inches distance), wounding a person on the other side of the door or window, will produce a wound that lacks all the features of close-range firing. Clothes can shield the skin and filter out the gunpowder and soot, too. Therefore, examination of the victim's clothes is imperative to ascertain the presence of soot or gunpowder around the entrance wound.

Gunshot wounds can be either penetrating or perforating. If the bullet enters the body and remains inside, it is a penetrating wound. If the bullet passes through the body and exits, it is a perforating wound. An exit wound is typically larger and more irregular than an entrance wound, with no abraded margin (abrasion ring around the bullet hole). The edges of exit wounds are usually torn, creating a stellate configuration or ragged appearance (Fig. 5.11). Exit wounds may be slit-like, resembling a stab wound. The edges of the skin of an exit wound usually can be reapproximated. Occasionally, there may be an abraded margin around the exit wound. This occurs when the exit site is in contact with or shored by a firm surface of another object as the bullet is attempting to exit the body, thereby slapping the skin against a hard surface that produces abrasions around the exit wound. These "shored" exit wounds usually show irregular configurations with much wider and more irregular abraded margins.

Shotgun Wounds

Shotgun wounds are produced by smooth-bore weapons that are designed principally to fire a shell containing multiple pellets down the barrel rather than a single projectile. There are two common



Fig. 5.12 A 34-year-old man was found in the front seat of his vehicle with a shotgun in his hand and pointed toward his head. Note the blowout and extensive destruction of the top of the head

types of shots loaded in shotgun shells: birdshot (tiny lead or steel pellets) and buckshot (larger lead or steel pellets). In addition to birdshot and buckshot, a shotgun can be loaded with a single large projectile called a slug. The shotgun is used mainly for hunting game. Wounds produced on the human body by a shotgun are usually devastating, especially if the shotgun is fired at a contact or close range (Fig. 5.12).

The Role of Forensic Pathology in Public Health and Safety

Traditionally, the emphasis of work done by medical examiners, coroners, and the death investigation community has been viewed as serving the criminal justice system. During the last several decades, however, the role of medical examiners and coroners has evolved from criminal justice service to a broader involvement that now significantly benefits public health and safety (Hanzlick 2006). The public service goal of forensic pathology is to investigate death for the benefit of the living by the development of strategies to prevent injury, disease, and death. Specific involvement of forensic pathology in public health and safety are as follows:

1. Death certification is a public health surveillance tool and a valuable source of information at the national and local levels. Among activities that benefit from the availability of cause of death and manner of death statistics obtained from death certificates are the monitoring of the health of populations, the setting of priorities, and the targeting of intervention. Such statistics are also the keystone of much epidemiological study. Medical examiners and coroners certify approximately 20% of the deaths in the USA and therefore are a major contributor to national mortality, especially in regard to nonnatural deaths and sudden, unexpected natural deaths (Hanzlick and Parrish 1996).
2. Death investigation can be the early warning system for dangerous hazards in the community. Patterns of preventable death may be identified in the workplace and on the road and associated with recreation, disease, or injury. Identifying these patterns, because they may be occurring over large geographic areas and over time, requires that death investigation be handled systemically

and detailed information be collected. Data collected during forensic death investigation has a proven ability to detect clusters and unusual deaths. In addition, death investigation data can yield timely and specific information about an unfolding epidemic and can also be used to discern risk factors that are the key to developing preventive interventions. The detailed investigation of deaths caused by injuries constitutes a substantial forensic contribution to injury prevention and improvement in public health and public safety.

3. Knowledge gained from forensic autopsy can contribute to the evaluation of poorly understood diseases and new medical therapies and surgical techniques and procedures. It can also assist families by providing a factual basis for genetic counseling of relatives if diseases with genetic components are identified. Subject to observance of relevant law and in accordance with local customs (which may include ensuring that the consent of the next of kin is obtained), tissue available as a consequence of the autopsy may be retained for medical research and be used for therapeutic purposes (corneas, aortic valves, bones, and skin).
4. Medical examiners and coroners form an important part of the complex response to a known bioterrorist event and emerging infectious diseases. Bioterrorism is the use or threatened use of biological agents or toxins against civilians with the objective of causing fear, illness, or death. Deaths as a consequence of a known bioterrorist or terrorist attack are homicides, so they fall under the jurisdiction of medical examiners and coroners. All five fatalities due to anthrax inhalation in 2001 were referred to medical examiners, and all five victims were autopsied (Borio et al. 2001; Centers for Disease Control and Prevention 2001). Medical examiners might see fatalities that have not been seen by other health providers. For example, in 1993, medical examiners were the first to recognize an outbreak of a fatal respiratory disease, which led to a rapid multiagency investigation and the identification by the CDC of an emerging infectious disease, Hantavirus pulmonary syndrome (Nolte et al. 1996). Medical examiners and coroners also have played an important role in recognizing outbreaks and cases of fatal plague (Jones et al. 1979; Kellogg 1920) and malaria (Helpern 1934).
5. Medical examiners and coroners also play a pivotal role in a number of ongoing surveillance programs (Hanzlick 2006) that have a public health and safety focus: (a) Drug Abuse Warning Network. The Substance Abuse and Mental Health Services Administration administers this surveillance system for collecting information on emergency room visits and deaths related to nonmedical use of drugs and substances that may have resulted from homicide, suicide, or accident, and in cases in which the circumstances could not be determined. Data are collected periodically from medical examiners and coroners; (b) Medical Examiners and Coroners Alert Project (MECAP). The Consumer Product Safety Commission administers this program to collect timely information on deaths involving consumer products. After follow-up on reports, unsafe consumer products can be recalled or standards may be developed to improve the safety of products; (c) National Violent Death Reporting System (NVDRS). The goals of this state-based program are to inform decision makers about characteristics of violent deaths and to evaluate and improve state-based violence prevention. Medical examiner and coroner records are crucial to the NVDRS project because much of the NVDRS data are derived from such records in conjunction with police and crime laboratory records; (d) Child Death Review Teams (also known as Child Fatality Review Teams). The teams are also state-based. Core membership generally includes representatives from the medical examiner/coroner's office, law enforcement, prosecutorial agencies, child protective services, and public health agencies. The teams examine all child fatalities, especially those deaths in which medical examiner/coroner's services are involved. Systematic multiagency reviews consist of agencies sharing information to improve case management, promote child health and safety, increase criminal convictions of perpetrators, and protect surviving siblings.

In summary, forensic pathology, as a branch of medicine, applies principles and knowledge of medical sciences and technologies to problems in the court of law. Medicolegal death investigation serves the criminal justice system by detecting criminal activity or collecting evidence and

developing opinions for use in criminal or civil law proceedings. During the last several decades, however, the role of medical examiners and coroners has evolved from criminal justice service to a broader involvement that now significantly benefits public health and safety. Medicolegal death investigation has played an important role in satisfying the needs and protection of public health, public safety, education in medicine, research, and the development of strategies to prevent injury, disease, and death.

References

- Batalis, N. I. (2010). Blunt force trauma. eMedicine. <http://emedicine.medscape.com/article/1068732-overview> Accessed 27 Dec 2010.
- Borio, L., Frank, D., Mani, V., et al. (2001). Death due to bioterrorism-related inhalational anthrax: report of 2 patients. *Journal of the American Medical Association*, 286, 2554–2559.
- Brunel, C., Fermanian, C., Durigon, M., & de la Grandmaison, G. L. (2010). Homicidal and suicidal sharp force fatalities: autopsy parameters in relation to the manner of death. *Forensic Science International*, 198(1–3), 150–4.
- Centers for Disease Control and Prevention. (2001). Update: investigation of anthrax associated with intentional exposure and interim public health guidelines. *Morbidity and Mortality Weekly Report*, 50, 889–893.
- Collins, K. A., & Nichols, C. A. (1999). A decade of pediatric homicide: a retrospective study at the Medical University of South Carolina. *American Journal of Forensic Medicine and Pathology*, 20(2), 169–72.
- Demirci, S., Dogan, K. H., & Gunaydin, G. (2008). Throat-cutting of accidental origin. *Journal of Forensic Sciences*, 53(4), 965–7.
- DiMiao, V. J. (1985). Gunshot wound: practical aspects of firearms, ballistics, and forensic techniques (pp. 111–120). New York: Elsevier.
- DiMiao, V. J., & DiMiao, D. (2001). *Forensic pathology*. New York: CRC Press LLC.
- Fukube, S., Hayashi, T., Ishida, Y., Kamon, H., Kawaguchi, M., Kimura, A., & Kondo, T. (2008). Retrospective study on suicidal cases by sharp force injuries. *Journal of Forensic and Legal Medicine*, 15(3), 163–7.
- Gill, R., & Catanese, C. (2002). Sharp injury fatalities in New York City. *Journal of Forensic Sciences*, 47, 554–557.
- Godwin, T. A. (2005). End of life: natural or unnatural death investigation and certification. *Disease-a-Month*, 51(4), 218–277.
- Hanzlick, R. (2003). Overview of the medicolegal death investigation system in the United States: workshop summary. <http://www.nap.edu/openbook.php> Accessed 24 Oct 2010.
- Hanzlick, R. (2006). Medical examiners, coroners, and public health: a review and update. *Archives of Pathology and Laboratory Medicine*, 130(9), 1274–1282.
- Hanzlick, R., & Parrish, R. G. (1996). The role of medical examiners and coroners in public health surveillance and epidemiologic research. *Annual Review of Public Health*, 17, 383–409.
- Helpern, M. (1934). Malaria among drug addicts in New York City. *Public Health Reports*, 49, 421–423.
- Hunsaker, D. M., & Thorne, L. B. (2002). Suicide by blunt force trauma. *American Journal of Forensic Medicine and Pathology*, 23(4), 355–9.
- Jones, A. M., Mann, J., & Braziel, R. (1979). Human plague in New Mexico: report of three autopsied cases. *Journal of Forensic Sciences*, 24(1), 26–38.
- Karch, D. L., Dahlberg, L. L., & Patel, N. (2010). Surveillance for violent deaths—National Violent Death Reporting System, 16 States, 2007. *Morbidity and Mortality Weekly Report Surveillance Summaries*, 59(4), 1–50.
- Karger, B., Niemeyer, J., & Brinkmann, B. (2000). Suicides by sharp force: typical and atypical features. *International Journal of Legal Medicine*, 113(5), 259–62.
- Karger, B., Rothschild, M., & Pfeiffer, H. (2001). Accidental sharp force fatalities – beware of architectural glass. *Forensic Science International*, 123, 135–9.
- Katkici, U., Ozkök, M. S., & Orsal, M. (1994). An autopsy evaluation of defence wounds in 195 homicidal deaths due to stabbing. *Journal of the Forensic Science Society*, 34(4), 237–40.
- Kellogg, W. H. (1920). An epidemic of pneumonic plague. *American Journal of Public Health*, 10, 599–605.
- Lee, C. K., & Lathrop, S. L. (2010). Child abuse-related homicides in New Mexico: a 6-year retrospective review. *Journal of Forensic Sciences*, 55(1), 100–3.
- Mason, J. K., & Purdue, B. N. (2000). *The pathology of trauma* (3rd ed.). London: Arnold.
- National Association of Medical Examiners. (2002). A guide for manner of death classification. First Edition. <http://thename.org/index2.php> Accessed 24 Oct 2010.

- Nolte, K. B., Simpson, G. L., & Parrish, R. G. (1996). Emerging infectious agents and the forensic pathologist: the New Mexico model. *Archives of Pathology and Laboratory Medicine*, 120(2), 125–128.
- Platt, M. S. (1993). History of forensic pathology and related laboratory sciences. In W. U. Spitz (Ed.), *Medicolegal investigation of death* (pp. 3–13). Springfield: Charles C. Thomas.
- Prahlw, J. A. (2010). Sharp force injuries. *eMedicine* (2010). <http://emedicine.medscape.com/article/1680082-overview> Accessed 28 Dec 2010.
- Prahlw, J. A., Ross, K. F., Lene, W. J., & Kirby, D. B. (2001). Accidental sharp force injury fatalities. *American Journal of Forensic Medicine and Pathology*, 22(4), 358–66.
- Spitz, W. U. (1993). Gunshot wound. In W. U. Spitz (Ed.), *Medicolegal investigation of death* (pp. 311–381). Springfield: Charles C. Thomas.
- The Office of the Chief Medical Examiner of the City of New York. (1967). Report by the committee on public health. New York Academy of Medicine. *Bulletin New York Academy of Medicine*, 43, 241–249.
- US Center for Disease Control and Prevention. (2007). Years of potential life lost (YPLL) before age 65. <http://webapp.cdc.gov/cgi-bin/broker.exe>. Accessed 6 Jan 2011.
- US Centers for Disease Control and Prevention (2010). Deaths: Final Data for 2007. *National Vital Statistics Reports*, 58(19), 11 (available at http://www.cdc.gov/nchs/data/nvsr/nvsr58/nvsr58_19.pdf) Accessed 5 Jan 2011.
- US Centers for Disease Control and Prevention. (2010). QuickStats: death rates for the three leading causes of injury death† – – United States, 1979–2007. *Morbidity and Mortality Weekly Report*, 59(30), 957.
- US Centers for Disease Control and Prevention (2011) WISQARS Leading Causes of Death Reports, 1999 – 2007. Available at: <http://www.cdc.gov/ncipc/wisqars/> Accessed 4 Jan 2011.

Chapter 6

Determination of Injury Mechanisms

Dennis F. Shanahan

Introduction

Investigations of aircraft and automobile crashes are generally conducted by government entities for the express purpose of determining the cause of the crash. Determination of the cause of injuries incurred in the crash is frequently not considered or is given only minimal emphasis. Traditionally, this emphasis on crash cause determination was to identify and fix systemic problems that led to the crash and that might contribute to future crashes if not corrected or, less commonly, to affix blame. In theory, focus on correction of systemic causes of crashes could ultimately lead to elimination of crashes. While a laudable and necessary goal, total reliance on this concept ignores the fact that transportation is a human endeavor and, as such, is inherently fallible – a zero crash rate will never be achieved in spite of all efforts to the contrary. Consequently, it is equally important to investigate injury mechanisms in crashes to understand how injuries occur and, from this understanding, develop improved means of mitigating crash injury. Working toward both goals simultaneously is the best way to minimize casualties in any transportation system. This chapter discusses a methodology of determining injury mechanisms in vehicular crashes.

Injury Mechanisms

An injury mechanism is a precise mechanistic description of the cause of a specific injury sustained in a particular crash. As an example, a restrained, adult passenger of an automobile who was involved in a 48-kph (30-mph) crash into a tree sustains a rupture of the large bowel with associated mesenteric tears and a large, horizontal linear contusion at the level of the umbilicus. This situation is frequently described in the medical and engineering literature and is part of what is often referred to as the “seat belt syndrome” (Garrett and Braunstein 1962; Williams et al. 1966; Smith and Kaufer 1967; Williams 1970; Anderson et al. 1991). The abdominal contusion is often referred to as a “seat belt sign,” and its location as well as the underlying large bowel injuries is consistent with the lap belt riding above the level of the iliac crests and impinging on the soft abdominal wall instead of remaining on the pelvis as it was intended to do in a frontal crash (Thompson et al. 2001). This is a

D.F. Shanahan, MD, MPH (✉)
Injury Analysis, LLC, 2839 Via Conquistador, Carlsbad, CA 92009-3020, USA
e-mail: dfshanahan@aol.com

situation known as “submarining” the lap belt (Department of the Army 1989). The foregoing summary constitutes a description of the mechanism of injury and involves analysis of data obtained from the person, the crash, and the vehicle. Mechanistic descriptions not only provide the information necessary to understand how this serious abdominal injury was caused but also provide a basis upon which to develop mitigation strategies for this imminently preventable injury.

Determining injury mechanisms in a series of crashes allows epidemiological researchers, vehicle manufacturers, and government agencies to quantify the prevalence of injuries and associated injury mechanisms for various types of crashes as well as provide objective data upon which to base mitigation priorities and strategies. The term often applied to the ability of a vehicle and its protective systems to prevent injury in a crash is “crashworthiness.” The absence of epidemiologic data on injury mechanisms in crashes leads either to a stagnation of improvements in crashworthiness design for a particular vehicle or class of vehicles or it leaves decision makers with no option but to establish priorities based on anecdotal impressions rather than objective data. The first scenario allows unnecessary and potentially preventable injuries to continue, and the latter leads to inefficiencies of cost and manpower.

Unfortunately, injury mechanism data are not collected for all forms of transportation. Currently, this type of data is only consistently collected and analyzed for motor vehicle crashes in the USA by the National Highway Traffic Safety Administration (NHTSA) through the National Accident Sampling System Crashworthiness Data System (NASS–CDS) and the Crash Injury Research and Engineering Network (CIREN). Other Department of Transportation agencies as well as the National Transportation Safety Board (NTSB) do not routinely collect or analyze injury data or determine injury mechanisms. Lack of injury data has been a major impediment to developing effective safety regulations as well as improved crashworthiness designs in general aviation aircraft and helicopters (Baker et al. 2009; Hayden et al. 2005). This problem was identified in a study commissioned by the DOT and conducted by Johns Hopkins University Bloomberg School of Public Health with the participation of diverse injury experts from around the country. The authors of the study recommended in 2006 that all transportation modes institute programs similar to the NASS–CDS to systematically determine how injuries are occurring, to provide an objective basis for more effective safety guidelines and regulations, and to provide a basis for the initiation of programs to mitigate those injuries. Unfortunately, this recommendation has yet to be implemented by the DOT or any of its agencies.

Reasons why crash investigation agencies do not place a greater emphasis on determining injury causation are numerous. First, there is a general lack of understanding of the importance of crash injury analysis and its various applications. A second issue is that funding is always a problem in government, and increasing the scope of investigations and data collection admittedly leads to greater costs and increased time and manpower. Thirdly, there are very few investigators trained in the field of biomechanics and injury cause determination. Finally, there is a general lack of coordination between investigative agencies and medical caregivers and medical examiners, which inhibits the free flow of vital information related to injuries sustained in crashes.

As one can infer from the preceding discussion, determination of injury mechanism is a complex process that requires the synthesis of analyses of all aspects of a crash related to the person, the vehicle, and the crash.

The Person

A full analysis of the occupants of a vehicle involved in a crash forms the basis for the determination of crash injury mechanisms. Additional information obtained in the analysis of the occupants should include seating location, physical position at the time of the crash (i.e., sitting, kneeling on the floor, etc.), restraint use, physical condition, age, sex, and clothing worn.

Table 6.1 Classification of crash injury mechanisms

(A) Mechanical injury
1. Acceleration
2. Contact
3. Mixed
(B) Environmental injury
1. Fire
2. Drowning
3. Heat/dehydration
4. Cold
5. Chemical exposure (fuel, cargo)

Classification of Traumatic Injuries

At the risk of oversimplifying the issue, it is useful from a mechanistic standpoint to divide injury suffered in vehicular crashes into mechanical injury and environmental injury. Mechanical injury may be further subdivided into contact injury and acceleration injury (Shanahan and Shanahan 1989). Environmental injury includes burns, both chemical and thermal, cold and heat exposure injuries, and other events related to the environment such as drowning or inhalational injuries (Table 6.1).

In a strict sense, both acceleration and contact injuries arise from application of force to the body through an area of contact with a potentially injurious surface. In the case of acceleration injury, the application of force is more distributed so that the site of force application usually does not receive a significant injury and the site of injury is distant from the area of force application (Shanahan and Shanahan 1989). In this case, injury is due to the body's inertial response to the acceleration, which is simply a manifestation of Newton's Third Law of Motion – for every action there is an opposite and equal reaction (King and Yang 1995). An example of an acceleration injury is laceration of the aorta in a high-sink-rate crash of an aircraft. Here the application of force is through the individual's thighs, buttocks, and back where his body is in contact with the seat. The injury itself is due to shearing forces at the aorta imparted by the downward inertial response of the heart and major vessels to the abrupt upward acceleration of the body.

A contact injury, on the other hand, occurs when a portion of the body comes into contact with a surface with sufficient force that injury occurs at the site of contact (“secondary collision”) (King and Yang 1995; Shanahan and Shanahan 1989). This contact may result in superficial blunt force injuries including abrasions, contusions, and lacerations or incised wounds, depending on the physical nature of the contacted object as well as deeper injuries to organs or the skeletal system. Relative motion between the contacting surface and the body is required for blunt force injuries and may be due to motion of the body toward the object, motion of the impacted object toward the occupant, or a combination of both. An example of this type of injury is a depressed skull fracture due to impact of the head into an unyielding object within the vehicle. Here the contact is of sufficient force that the object penetrates, at a minimum, the outer table of the skull. A mixed form of injury may also occur wherein there are both contact injury and acceleration (inertial) injury resulting from a single impact. An example of a mixed form of injury would be a primary depressed skull fracture with an associated contracoup injury to the brain resulting from the inertial motion of the brain within the skull secondary to the initial contact.

Distinction is made between the major forms of traumatic injury since, mechanistically, they are quite different, and, as a result, mitigation of these injuries involves distinctly different intervention strategies. The basic method of preventing acceleration injuries is to provide means within the vehicle structure or seating system to absorb a portion of the energy of a crash so that that energy is not

transmitted to occupants. Structural crush zones within a vehicle, energy-attenuating seats, and energy-absorbing landing gear or wheels are designed to provide this function. The primary strategy employed to prevent contact injury, on the other hand, is to prevent occupant contact with internal or intruding structures. This can be accomplished through a variety of methods including improved occupant restraint or relocation of the potentially injurious object. If a potentially injurious object cannot be practically moved such as the case of vehicle controls such as steering wheels or aircraft controls, injury can be mitigated by reducing the force of body contact through such strategies as padding of the object, making the object frangible (breakaway) such that the object yields before injury can occur, or providing the occupants with impact-mitigating protective equipment such as crash helmets (Department of Defense 2000).

For crashes where there is generally good preservation of the occupant compartment, sometimes referred to as survivable or potentially survivable crashes, acceleration injuries are relatively rare. This is because crash accelerations that are in excess of human tolerance most often result in significant collapse of occupied spaces. In these cases, the occupants receive very significant contact injuries that mask potential acceleration injuries. In a study of US Army helicopter crashes, it was determined that in survivable crashes, contact injuries exceeded acceleration injuries by a ratio of over seven to one (Shanahan and Shanahan 1989). Most of the identified acceleration injuries were vertebral fractures related to high-sink-rate crashes.

Injury Identification

Injury identification most often relies on review of existing records. Sources of information include police crash reports and photographs, first responder records (fire department and ambulance), hospital records including photographs and imaging studies, and, in the case of fatally injured occupants, autopsy reports, photographs, and imaging studies obtained by the medical examiner or coroner. The value of photographs of the occupants taken at the scene before their removal from the vehicle cannot be overstressed in the process of injury mechanism determination.

It should be remembered that a single injury mechanism may result in a group of injuries occurring at various anatomic levels and locations. These concurrent injuries should be grouped together since they were caused by the same general mechanism. For instance, a single, distributed blunt impact to the anterior chest may result in superficial skin injuries, anterolateral or posterior rib fractures with associated bleeding, and cardiac contusion or laceration with associated bleeding. These are markedly different injuries from a care standpoint, but they all result from a single blunt force impact, although the injury may be distant from the site of impact. The anterolateral or posterior rib fractures noted above are a good example of this phenomenon. The thoracic cage, from an engineering perspective, forms a hoop or ring. Compression of a hoop generates stresses distant from the area of compression, and failure will occur at the area of greatest force concentration or at the weakest part of the hoop (Crandall et al. 2000; Yoganandan et al. 1993; Love and Symes 2004; Kallieris et al. 1998). This is why a distributed anterior chest compression frequently results in anterolateral or posterior rib fractures. A similar situation exists for compression injuries of the pelvis where an impact to one side may result in fractures on the contralateral side (Tile 1996). For some spinal injuries, an analogous situation exists. For instance, a distributed impact to the top of the head may result in cervical spine fractures, while the skull at the area of impact does not fracture. This is because the skull has considerable tolerance to well-distributed impact forces, while the cervical spine under certain orientations of the head-neck complex has considerably lower tolerance to transmitted force (Alem et al. 1984).

An interesting aspect of injury mechanism determination is that minor, superficial injuries are very frequently more helpful in identifying the mechanism of injury than the more serious internal injuries. Superficial injuries often provide detailed information about the nature of an object contacted

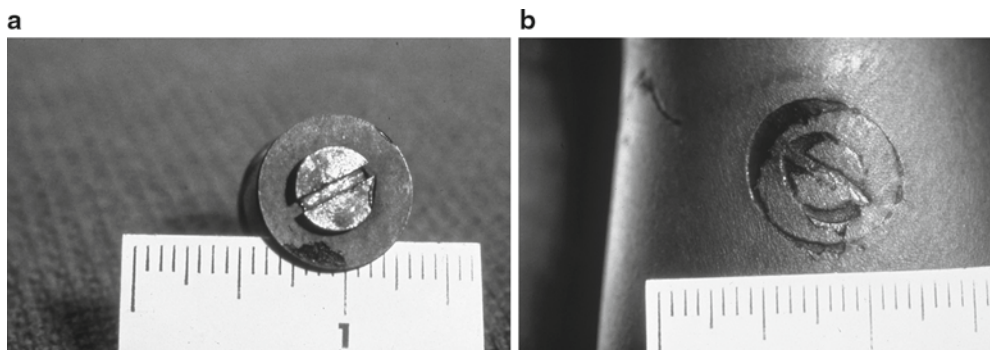


Fig. 6.1 Patterned contusion on the calf of a pedestrian from impact by a license plate screw

as well a relative estimate of the forces involved. In this respect, so-called patterned abrasions and contusions are probably the most useful types of wound. These occur when a distinctly shaped object leaves its imprint in the flesh of the impacted occupant. Figure 6.1 shows how a screw securing a front license plate on an automobile left an imprint on the leg of a pedestrian after striking the pedestrian. Although patterned contusions are gratifying to the analyst, such determinative evidence is uncommon in crash investigations. More often, contact surfaces leave considerably less distinct evidence on the body. More commonly seen are less definitive abrasions and contusions such as those often seen on occupants from loading into their belt restraint systems.

The nature of superficial injuries can also provide a great deal of information about the impact surface. Abrasions to the outboard side of the face of an automobile crash victim that are linear and contain embedded grit that is often black in color suggest a rollover collision where the face impacted the road surface creating a so-called “road rash” injury. When the scratch pattern is more random and finer with imbedded dirt or plant materials, it suggests the contact was to an unpaved surface over which the vehicle traveled during the crash sequence. Therefore, careful analysis of external injuries can provide very significant information in decoding injury mechanisms.

As critical as superficial injuries are to the determination of injury mechanisms, they are probably the most poorly documented types of injury in crashes. This is because first responders are trained to stabilize and transport the patient, and they have little time or resources to identify injuries that are essentially inconsequential to the care of the patient. Also, many superficial injuries including abrasions and contusions take time to develop and may not be visible immediately after the crash, particularly under poor lighting conditions. The emphasis on potentially life-threatening injuries carries over to the emergency or trauma department staff. As a result, many superficial injuries will not be identified in emergency department (ED) or trauma team records. Usually, the most reliable and consistent source for identification of superficial injuries is the nurses’ notes because they have the opportunity to observe the wounds at 24 h when most contusions are fully developed and local inflammation surrounding abrasions and lacerations is maximal.

Since caregivers rarely provide an identification and detailed description of all superficial injuries and because pictures are far more descriptive than words, investigators or other parties interested in obtaining a reconstruction of injury causation are urged to obtain photographs of the entire body of crash victims within a few days of injury and, preferably, before surgery.

Once all injuries are identified, it is often useful to create an “injury map” for each occupant. This is simply a diagrammatic representation of an individual with, as a minimum, front and back views sometimes supplemented with side views or regional views. All injuries are then recorded on the depiction. When a large number of injuries occur, it is helpful to create separate injury maps for superficial injuries and internal injuries. This allows an analyst to visually assess the distribution and

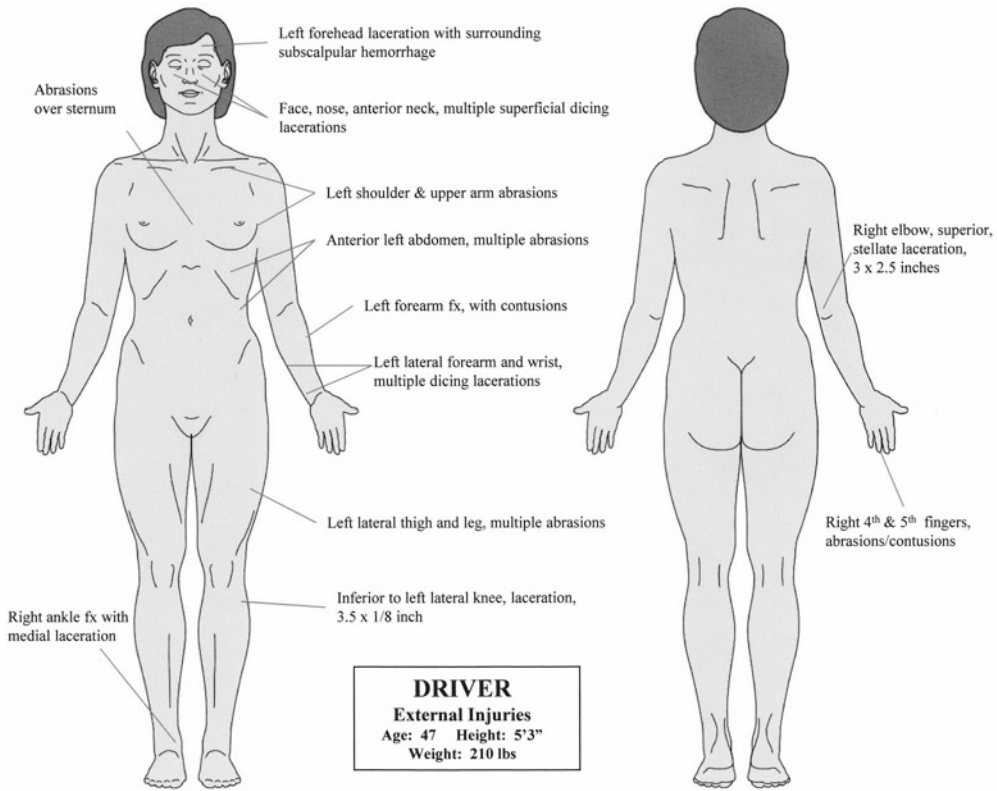


Fig. 6.2 Injury map showing predominance of left-sided injuries in a left, frontal collision

types of injuries sustained by the victim, which can be highly useful in the analytical process. As an example, when the most significant injuries occur on one side of the body, a side impact to the most injured side of the body is suggested (Fig. 6.2).

Injury Severity

The abbreviated injury score (AIS), which is an injury scale predictive of threat to life, is almost universally used by epidemiological and biomechanical researchers to classify the severity of injuries in crashes (AAAM 1990). It is an ordinal scale ranging from one to six where AIS 1 is a minor injury and AIS 6 is a uniformly fatal injury. AIS 9 or NFS (not further specified) would refer to an injury of undetermined severity. The AIS-coding manual defines AIS level for various types of injuries broken down by body region using rigorously defined criteria. Since the AIS has undergone a number of revisions in the past few decades, it is important to identify the version used to code injuries in a particular study or database. The injury severity score (ISS), which mathematically combines the greatest AIS injury in each of three body regions to yield a single score, may be used to predict overall outcome for a multiply injured patient (Baker et al. 1974; Baker and O'Neill 1976). The ISS is less often seen in the biomechanical and epidemiological literature to describe overall injury severity than the maximum AIS (MAIS) either applied to the whole person or to a particular body region. The "HARM" scale applies a weighting factor to injury severity to assess cost of injuries (Digges et al. 1994). These methods are more completely described in Chap. 14.

Police accident reports frequently use the observational KABCO scale to classify injury severity. This scale was proposed by the National Safety Council in 1966 as a means of classifying injury severity in police crash reports. This is a subjective, ordinal scale where K=killed, A=incapacitating injury, B=nonincapacitating injury, C=possible injury, and O=no injury (Compton 2005). Some states have added an additional code, “U,” which indicates injury, severity unknown (KABCOU) (Kindelberger and Eigen 2003). NASS–GES reports injury severity using the KABCO scale.

Because of the subjectivity of the KABCO scale and because the injuries are assessed by nonmedical personnel, usually at the crash scene, this method is very imprecise at all levels of the scale. Several authors have studied the correlation of police-reported KABCO scores with the AIS classification of injuries assigned by NHTSA investigators (Compton 2005; Farmer 2003). These studies found significant misclassification of injury severity in police reports, prompting Farmer to recommend caution in utilizing unverified KABCO scores in analytical studies. However, Compton concluded that the KABCO scale “appears to be an appropriate tool for planners to use to discriminate the more serious crashes from the multitude of minor crashes.” Regardless of these arguments, KABCO scores have significant error, and they probably should not be relied upon except to determine gross differences in injury severity.

Human Tolerance to Acceleration and Blunt Force Impact

To determine injury mechanisms that occur in a crash, it is important for investigators and researchers to have a general concept of how much acceleration the body can withstand as well as how much force various parts of the body can bear without serious injury (Snyder 1970a, b; Department of the Army 1989). This allows investigators to compare the injury with the forces calculated in the crash by reconstructionists to ensure a proposed injury mechanism appropriately correlates with the dynamics of the crash.

Also a detailed knowledge of human tolerance is vital in developing vehicular crashworthiness designs. This is because the design of protective equipment invariably involves a compromise between designing for the greatest amount of protection for the greatest number of crashes and the practical and economic realities of having limited space, weight, technology, and money to implement a particular improvement. A good illustration of this relates to the development of ejection seats in tactical military aircraft (Latham 1957; Clarke 1963; Levy 1964). Ideally, an ejection seat would be able to safely eject a pilot at all potential speeds, altitudes, and orientations of the aircraft. To accomplish this, among other things, the seat would have to accelerate extremely rapidly once initiated to get the occupant clear of the aircraft with sufficient altitude to deploy a parachute in the shortest time possible. Unfortunately, humans are limited by the amount of acceleration they can tolerate, a reality which forces designers to compromise the rate of acceleration with known human tolerance. To maximize the capability of the seat for all potential occupants, designers have to accept a certain percentage of minor or moderate injury to some occupants due to the variability of tolerance among the potentially exposed pilot population. Similar concerns and limitations apply to the development of any crashworthiness concept or item of protective equipment.

Whole-Body Acceleration Tolerance

Human tolerance may be conceptualized in a number of ways. In this section, we will consider tolerance to whole-body abrupt acceleration. In the field of biomechanics, distinction is made between exposures to abrupt (transient) or impact acceleration and sustained acceleration since tolerances for these exposures are considerably different (Brinkley and Raddin 1996; Cugley and Glaister 1999).

Table 6.2 Factors determining tolerance to abrupt acceleration

-
1. Magnitude
 2. Duration
 3. Rate of onset
 4. Direction
 5. Position/restraint/support
-

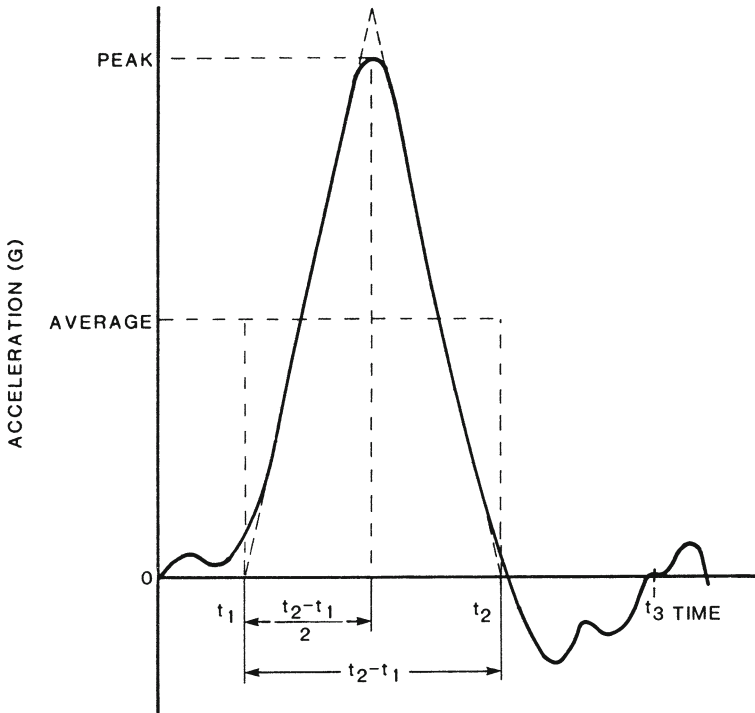


Fig. 6.3 Crash pulse showing magnitude and duration of acceleration (Department of the Army 1989)

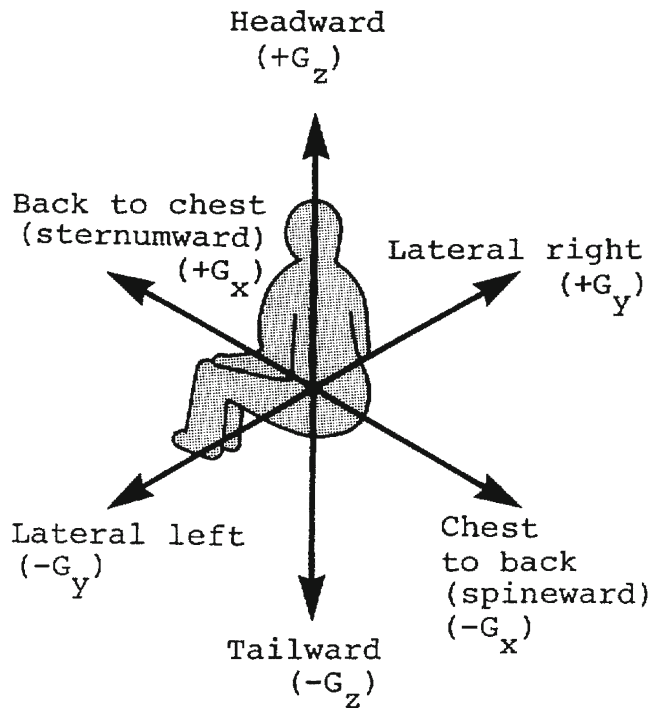
Generally, abrupt acceleration refers to accelerations of short duration with high rates of onset as occurs in a vehicular crash. Long-duration exposures (sustained) are those typically associated with maneuvering of tactical aircraft or those encountered in space flight. Most crash impacts result in pulse durations of less than one-quarter of a second (250 ms). As an example, it is rare for car-to-car impacts, which are considered long-duration impacts, to exceed a duration of 180 ms, and most barrier impacts are over in as little as 90–100 ms (Agaram et al. 2000). Most aircraft impacts experience similar pulse durations. Consequently, for purposes of considering human tolerance to impact, approximately 250 ms may be considered the upper limit duration of impact.

Tolerance to acceleration is dependent on a number of distinct factors, some related to the nature of the acceleration and others related to the exposed individual.¹ Table 6.2 is a summary of these factors.

An applied abrupt acceleration or crash pulse has magnitude, duration, and slope (rate of onset) and is generally depicted graphically as a plot of acceleration versus time (Fig. 6.3). For most

¹ Note that a deceleration is a negative acceleration, and many texts will refrain from the practice of using the term deceleration and instead refer to negative acceleration.

Fig. 6.4 Axes of seated human (Department of the Army 1989)



impacts, the shape of the pulse is essentially triangular or haversine. Although not completely true for all impacts, this assumption results in simpler calculations than attempting to apply more complex waveforms that may more precisely describe the actual crash waveform. More precise waveform descriptions rarely add significant benefit in predicting survival or confirming proposed injury mechanisms since the variation in tolerance from person to person is so large.

Magnitude of acceleration is probably the most critical factor in determining tolerance. For a given magnitude of acceleration, the longer the duration, the more likely injury will ensue since a longer-duration pulse involves greater energy than a shorter-duration pulse of the same magnitude. However, for a given impact energy, tolerance can be increased by increasing the duration of the impact which, in turn, lowers the magnitude of the acceleration. This occurs when crush zones are added between the impact point on the vehicle and the occupant compartment. Regarding rate of onset of acceleration, it has been shown experimentally in humans and animal surrogates that the more rapidly the acceleration is applied (higher jolt), the less tolerable that impact will be, all other parameters being equal (Department of the Army 1989).

The orientation of the body with respect to the applied acceleration vector is generally considered to affect one's tolerance to the acceleration. For purposes of description, both vehicles and humans are arbitrarily assigned coordinate axes. The coordinate system applied to the seated human is illustrated in Fig. 6.4 wherein the x-axis applies to fore-aft accelerations, the y-axis applies to transverse accelerations, and the z-axis applies to accelerations directed parallel to the spine (vertical). Each axis is assigned a positive and negative direction, which varies among different commonly accepted coordinate systems. The system illustrated here is the system developed by the Society of Automotive Engineers, which is the most commonly used system today (SAE 1995). Any force or acceleration may be described according to its components directed along each of the orthogonal axes, or the components may be mathematically combined to determine a resultant vector. In accordance with Newton's Third Law of Motion, an accelerated mass has an inertial response that is opposite and equal to the applied acceleration. It is the body's inertial response to an acceleration that results in

injury if that acceleration exceeds the tolerance of the exposed occupant, and the body's response to the acceleration is always opposite the direction of the applied acceleration.

The final factor (Table 6.2) related to human tolerance to whole-body abrupt acceleration encompasses a number of elements that are primarily related to the occupant and how he is packaged in the vehicle as opposed to the initial four factors which are related to the acceleration pulse determined by characteristics of the vehicle and the impacted object. This final factor is critical since it accounts for most of the variability in tolerance seen in crashes and, therefore, outcome for a given crash. It relates to how well the occupant is restrained and supported by his seat and restraint system and the degree to which crash loads are distributed over his body surface. It also encompasses the various occupant intrinsic factors, or factors directly related to the individual subjected to the impact, that in large part determine his tolerance to an impact. These factors explain the observed biological variability between different humans subjected to similar crash impacts and include:

1. Age

Testing as well as real-world crash investigations have repeatedly demonstrated that younger adults are less likely to be injured in a given impact than their older counterparts. This principle is reflected in military crashworthiness and protective equipment design criteria, which typically permit more severe accelerations than similar equipment designed for the general population (Department of the Army 1989). Additionally, children and infants demonstrate marked differences in impact response compared to adults of either sex (Burdi et al. 1969; Tarriere 1995).

2. General health

Chronic medical conditions such as heart disease and osteoporosis clearly degrade one's ability to withstand impact accelerations. History of previous injuries may also adversely affect one's tolerance.

3. Sex

There are clearly sex differences in tolerance to acceleration. Women have a different mass distribution and anthropometry than men as well as generally lower muscle mass and strength. This has been of particular concern for neck tolerance since women have approximately one-third less muscle mass than men of comparable age and stature.

4. Anthropometry

Anthropometric considerations involve differences in mass, mass distribution, and size related to sex, age, and individual variation. From a protective design standpoint, equipment design must account for the range of anthropometries of people expected to utilize the vehicle or equipment. A commonly accepted design range includes the so-called 5th percentile female to the 95th percentile male.

5. Physical conditioning

Physical conditioning appears to have a modest effect on tolerance to abrupt acceleration apparently related to muscle mass and strength. Conditioning is also thought to be a factor in recovering from injuries.

6. Other factors

Other intrinsic factors considered to have a possible effect include obesity, the phase of the cardiac cycle when impact occurs, and other unidentified factors (Viano and Parenteau 2008).

Restraint, Seating, and Support

The primary extrinsic factors determining one's tolerance to a particular acceleration relate to restraint, seating, and support. In many crashes, injuries are attributed to lack of restraint or to failures or inadequacies of existing restraint systems. For this reason, it is imperative that investigators of injury mechanisms have a thorough understanding of restraint theory and application. Belt

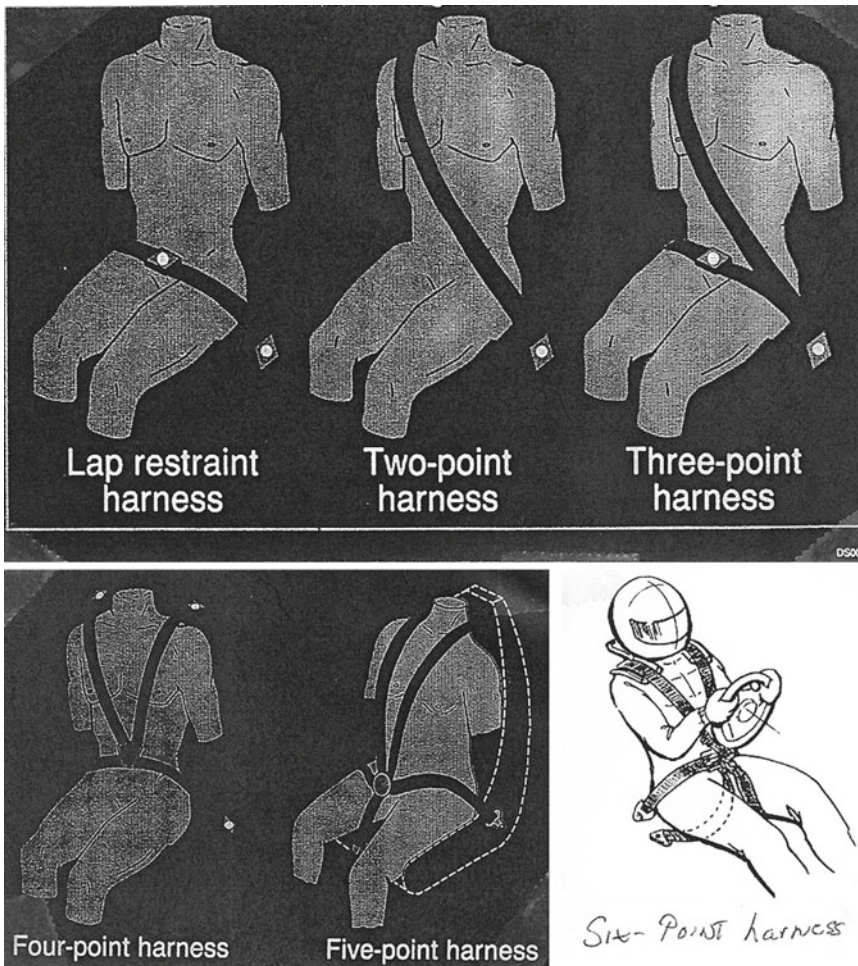


Fig. 6.5 Six belt restraint systems

restraint systems come in many forms and are well described in numerous textbooks and journal articles (Department of the Army 1989; Chandler 1985; Chandler 1990). Typically, belt restraints are described by their “points” of attachment. A lap belt system is referred to as a two-point system referring to the number of anchor attachments, one on each side of the pelvis. Another two-point system, often referred to as a “sash,” involves a single belt that is applied diagonally across the torso and anchored above one shoulder and at the contralateral hip. A three-point system adds to the lap belt a single, diagonal shoulder belt and its attachment point above the shoulder of an occupant. A four-point system adds a second shoulder belt with its separate anchor, and five- and six-point systems add a single or dual tie-down strap (“crotch strap”), respectively (Fig. 6.5). Each additional belt adds a degree of safety to the system, but frequently results in a decrease in convenience as well. It is for this reason that most automobiles are equipped with three-point lap/shoulder belt systems, although at least one automobile manufacturer is considering offering a four-point system to its customers, probably as an option (Rouhana et al. 2006).

A restraint system, as the name implies, is composed of a system of components designed to restrain the occupant in a crash or other sudden event. As such, it includes the entire tie-down chain for the occupant including the belt restraint system, the seat, and the anchoring mechanism securing the seat and, sometimes the belt restraint, to the vehicle. Air bag systems have been added to motor

vehicles and some aircraft to add supplemental restraint to the belt restraint system. Restraint systems serve multiple functions:

1. Prevent ejection

The original function of restraints when first developed in the early days of aviation was to prevent ejection from the aircraft during aerobatic maneuvering. Particularly in open cockpit aircraft, the consequences of not being restrained while performing aerobatic maneuvers were rather severe! Subsequently, prevention of ejection was also shown to be highly beneficial to survival in crashes.

2. Minimize the “second impact”

Restraint systems are designed to prevent the occupant from striking interior objects such as the steering wheel, dash, windshield, or other interior structures. Prior to the introduction of upper torso restraints and air bags, these contacts were frequent and often deadly.

3. Couple the occupant to the vehicle

A belt restraint system serves to couple the occupant to the vehicle during a crash allowing the occupant to benefit from the energy management provided by the crush of vehicle structure, thereby allowing the occupant to “ride down” the forces of the crash in unison with the vehicle.

4. Distribute crash loads across the body

Not only does a restraint system restrain the body, but it also serves to distribute the loads of a crash over the portions of the body that are most capable of sustaining high loads such as the pelvis, shoulder, and thoracic cage.

To better understand the principles of restraint, it is important to recognize that a crash is a dynamic event that is essentially governed by Newton’s Laws of Motion (King and Yang 1995). For simplicity, the following illustration will discuss a frontal crash of an automobile although the same principles apply to any crash in any direction. Newton’s First Law states that an object in motion will remain in motion in the same direction and velocity until acted on by external forces or objects. In a crash of a vehicle, an occupant will be moving at a given velocity with respect to the ground prior to an impact. At impact, the vehicle decelerates rapidly while an unrestrained occupant will continue moving forward in a sitting position until his chest impacts the steering wheel, his knees impact the lower portion of the dashboard, and his head strikes the windshield or windshield header. The force of these second collisions is determined by Newton’s Second Law, which states that force is equal to the product of the effective mass of the impacted body segment and the acceleration of that segment. The acceleration of each impacting segment is determined by the compliance or stiffness of the body segment and of the vehicle structure impacted. The more compliant the structures are, the lower the acceleration and, therefore, the more tolerable the impact. The addition of padding to interior structures serves to increase the duration of the impact by resisting deformation with a tolerable force thus increasing the stopping distance, which, in turn, decreases the acceleration. Of course, this example also illustrates the importance of designing restraint systems to prevent the second collision in the first place.

Restraints are designed to couple the occupant to the vehicle to prevent the development of a relative velocity between the occupant and the vehicle, thus allowing the occupant to “ride down” the crash with the vehicle. This is a rather complex concept related to a situation known as “dynamic amplification” or “dynamic overshoot.” Vehicular crashes involve considerable kinetic energy transfer based on their mass and velocity at impact ($E = 1/2 mv^2$). Much of the energy of the crash may be dissipated through deformation and crushing of structures outside of the occupant compartment. It is beneficial to tightly couple occupants to the vehicle so that they may profit from the energy attenuation afforded by crushing vehicular structures. This requires that the occupants be effectively coupled to the vehicle through their restraints, their seat, and the seat’s attachment to the vehicle, the so-called tie-down chain.²

²The seat is part of the tie-down chain, and dislodgement of the seat from the vehicle can be just as serious as a failure of the belt restraint system. In either case, the occupant becomes a projectile as he flies into bulkheads or other interior structures.

To the extent that an occupant is not effectively coupled to the vehicle through his restraint chain, he will have slack in the system that delays his deceleration with respect to the deceleration of the vehicle. For instance, in a frontal crash with a loose restraint system, the vehicle immediately decelerates as the occupant continues forward at his initial velocity. By the time he begins to be restrained by the loose restraint system, there may be a significant velocity difference between the occupant and his restraint that is tightly coupled to the vehicle. Since the mass of the vehicle is so much greater than the mass of the occupant, when he suddenly becomes restrained by his belt restraint, he must immediately assume the lower velocity of the vehicle. The impact of the occupant with his restraint system results in an acceleration spike that may be multiples of what the center of mass of the vehicle experienced. This is referred to as dynamic amplification or dynamic overshoot and is reflective of poor coupling of the occupant to the vehicle and often leads to serious, preventable injuries. This is why passengers are always urged to secure their belts tightly in any vehicle. To assist in reducing slack, many belt restraint systems now employ pretensioners, which are usually pyrotechnic devices that reel in slack from the webbing when activated by a crash. In practice, all seats and restraint systems are subject to some degree of dynamic amplification due to the inherent compliance of the body, cushions, webbing, and other elements of the restraint chain. The objective of restraint designers is to minimize dynamic amplification in most crash scenarios. However, they accomplish this to a variable degree, and it is imperative that the injury investigator be able to determine injuries resulting from poor restraint design.

Another important concept in restraint system design is distribution of force across the body. One of the problems of lap belt-only restraint systems, aside from lack of control of the upper torso and a tendency to facilitate submarining, is that in a frontal crash, the force of the crash is concentrated in a 48-mm (1.9-in.) band across the occupant's pelvis. Not only does the addition of one or more upper torso harness decrease the likelihood of a secondary impact to the upper torso, but it also allows the loads of the crash to be distributed over a greater area of the body than a lap belt alone. This reduces the loads over any particular area of the body and, thus, decreases the probability of injury from seat belt loading. Using rear-facing seats will maximize load distribution in a frontal crash. In this configuration, the forces of the crash are distributed across the entire posterior surface of the head, torso, and thighs, eliminating force concentration and decreasing the probability of injury. Rear-facing seats are particularly important for infants in automobiles (Car-Safety.org 2009)

Air bags are designed to distribute loads, absorb energy, and provide additional ride down for occupants by controlling occupant deceleration through collapse of the bag (Crandall et al. 2003). Air bags come in numerous varieties and serve multiple functions in preventing injury in automobile crashes. Frontal air bags serve to reduce the loads borne by the belt restraint by further distributing loads across the anterior upper torso. Knee air bags provide the same function for the knees and help prevent knee, thigh, and hip injuries when lap belts fail to prevent contact of the knees with underdash structures (Estrada et al. 2004; Yoganandan et al. 2001; Sochor et al. 2003; Rupp et al. 2008, 2010). Side protection air bags include torso bags, head bags, combination head/torso bags, and side curtain air bags.

Early air bag systems inflated at very high rates, sometimes resulting in serious injuries particularly to short-statured individuals who sat closer to the steering wheel than the NHTSA-recommended 10 in. or who were otherwise out of position, were unrestrained, or to children strapped into child restraint systems in the passenger seat (NHTSA 1997, 1999). Recent changes to safety regulations (FMVSS 208) have allowed manufacturers to decrease inflation rates, which should decrease the number of occupants injured in this manner. Regardless of these changes, if an occupant has any part of his body in the zone of inflation of a frontal air bag, these systems still have the capability of causing serious harm. In most cases, air bag impacts cause telltale abrasions to the face, neck, upper torso, or upper extremities in the area of contact as well as deeper, more serious injuries in extreme impacts.

Side curtain air bags are very effective in reducing injury in rollover collisions of motor vehicles. Roll-activated side curtain air bag systems were not introduced into automobiles until approximately 2002 due to developmental issues surrounding roll sensing and inflation of the curtain. Most air bags are only required for a small amount of time after inflation, usually less than 100 ms since frontal and side impacts are generally over in about 100 ms. Consequently, they have large vents to vent the gas when an occupant loads into the bag. Air bags intended to protect in a rollover must remain inflated on the order of 3–5 s since rollover crashes often involve numerous rolls that occur over a period of several seconds before the vehicle comes to rest. Therefore, these bags are not vented, many are coated to reduce their porosity, or they use cold inflators to avoid cooling of the gas which helps maintain bag pressure.

Sensing an impending roll is more complex than sensing an impact and requires significant development and testing. Nevertheless, most automobile manufacturers have overcome these issues and offer roll-activated side air bag systems particularly in SUVs which are much more prone to rollover than passenger cars (NHTSA 2003). These systems have been shown to significantly reduce partial ejections of the head and upper torso and to reduce injuries in rollover collisions. Such technology is of vital importance since NHTSA has shown that although only about 3% of tow-away crashes involve a rollover, approximately one-third of fatalities occur in these rollover crashes (NHTSA 2005).

Tolerance

The above discussion illustrates that there are numerous variables, both intrinsic and extrinsic, that influence one's tolerance to abrupt acceleration. This leads to a wide variation in tolerance among individuals exposed to similar crashes. Nevertheless, testing combined with crash investigations has provided the basis for establishing general estimates of human tolerance. The determination of human tolerance to impact has been impeded by the obvious limitations in testing live subjects at potentially injurious levels of acceleration. This led to the use of various human surrogates including cadavers, primates, and other animal surrogates, all of which have their own limitations in biofidelity or how well they mimic a live human.

The earliest systematic testing of live volunteer subjects was performed by the US Air Force under the direction of John P. Stapp, M.D., beginning in the late 1940s (Stapp 1961a, b).³ Dr. Stapp and his team used a number of devices to expose volunteer subjects, usually themselves, to various acceleration pulses. They also performed a number of tests using animal surrogates. In 1959, Eiband compiled what was then known about tolerance including the work of Dr. Stapp as well as other data from a variety of studies performed on various animal models (Eiband 1959). Based on these data, he generated curves of acceleration versus time showing different levels of tolerance (voluntary, minor injury, and severe injury) for any combination of average acceleration and duration. He generated separate plots for each of the three orthogonal axes (Fig. 6.6). These plots provide the basis for current estimates for human acceleration tolerance used by the designers of aircraft and aviation protection systems. The US Army updated the Eiband data in the Aircraft Crash Survival Design Guide (Department of the Army 1989). Table 6.3 provides tolerance without serious injury estimates in terms of average acceleration along each axis for pulse durations of 100 ms (0.1 s) for fully restrained (lap belt plus upper torso harness) subjects.

³ Earlier testing was performed by the German military during World War II on prisoner subjects. What little reliable data these tests generated is not generally available and, for ethical reasons, has not been utilized by researchers in the field.

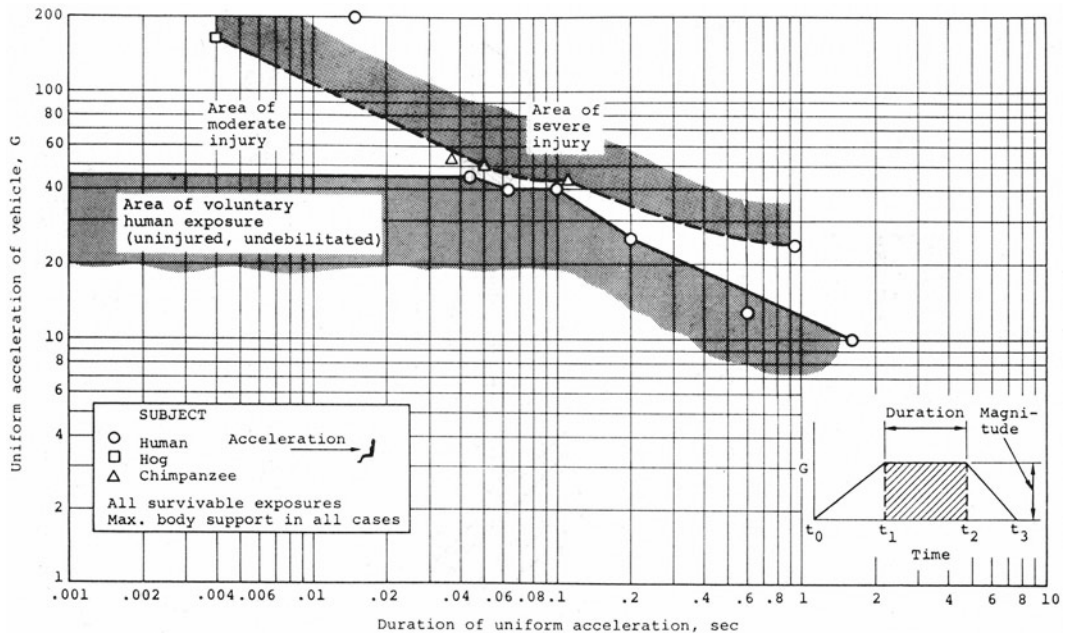


Fig. 6.6 Eiband curve showing tolerance for $-x$ -axis accelerations (Department of the Army 1989)

Table 6.3 Human tolerance limits

Direction of accelerative force	Occupant's inertial response	Tolerance limit
Headward (+Gz)	Eyeballs down	20–25 G
Tailward (-Gz)	Eyeballs up	15 G
Lateral right (+Gy)	Eyeballs left	20 G
Lateral left (-Gy)	Eyeballs right	20 G
Back to chest (+Gx)	Eyeballs in	45 G
Chest to back (-Gx)	Eyeballs out	45 G

Reference: Crash Survival Design Guide, TR 89-22; 100 ms crash pulse; full restraint

It should be noted that the x -axis has the greatest tolerance for accelerations under typical impact durations. The limit of 45 G provided the basis for cockpit strength design often referred to as the “45-G cockpit.” The intention was to preserve occupied space for all impacts below this limit to provide the best chance for crash survival without providing costly excessive protection.

The limits shown in Table 6.3 have provided the basis for vehicle crashworthiness design for many decades. Recent studies of Indianapolis-type racecar (Indy car) crashes demonstrate that these limits may be quite conservative. In a cooperative effort between General Motors and the Indianapolis Racing League (IRL), Indy cars have been equipped with onboard impact recorders to record impact accelerations in the cockpit (Melvin et al. 1998). Also, most Indy car crashes are videotaped by sports media, which provides additional data on the crash. This surveillance program provides laboratory quality data on the impact tolerance of humans for crashes that could not be performed for research purposes due to the risk of injury for the drivers. More than 260 crashes have been recorded and analyzed (Melvin et al. 1998). Peak accelerations as high as 60–127 G have been recorded in frontal, side, and rear crashes with durations similar to those experienced in highway crashes.

Average accelerations in excess of 100 G have been recorded in side impacts, and average accelerations have exceeded 60 G in front and rear impacts without any serious torso injuries to the involved drivers. These results indicate that young, well-conditioned subjects under the idealized conditions of an Indy car cockpit can survive much more serious accelerations than previously thought possible. Although similar protective systems are not practical in most other vehicles, the Indy car results show that a higher level of protection is than currently available potentially achievable in other types of vehicles.

Regional Tolerance to Impact

Vehicular crashes rarely result in inertial injuries primarily because structural collapse into occupied spaces usually occurs significantly before whole-body acceleration limits are exceeded. Consequently, contact injury to one or more locations on the body is a far more common occurrence in vehicular crashes. By definition, these injuries occur due to contact of a part of the body with interior structures due to inadequate restraint (flailing), due to intrusion of structures into the occupant compartment, or due to a combination of both mechanisms. Different regions of the body demonstrate different sensitivities to blunt impact as well as different injury mechanisms and rates of injury. Epidemiological studies of frontal automobile crashes demonstrate that the most frequent seriously injured ($\text{AIS} \geq 3$) body regions are the extremities, thorax, and head for restrained drivers (Stucki et al. 1998). In fatal crashes, head injuries predominate (Alsop and Kennett 2000; Huelke and Melvin 1980). The distribution and severity of blunt impact injuries is related to the type of crash as well as to use of restraint and air bag systems.⁴

In summary, injury mechanism determination in a crash must begin with a detailed record of occupant injuries and injury severity gleaned from various sources. Additionally, the injury investigator must have knowledge of human tolerance to acceleration and impact and the various factors that affect an occupant's tolerance including intrinsic and extrinsic factors. As will be seen, this information is combined with an analysis of various crash factors as well as evidence determined from a vehicle inspection to finally reconstruct mechanisms of injury.

The Crash

To determine injury mechanisms in crashes, it is essential for investigators to have detailed knowledge of the crash circumstances since injury and occupant kinematics are influenced by the type of crash (frontal, side, rear, rollover, multiple events) and various characteristics of the crash including velocity change, principal direction of force (PDOF), and the crash pulse. It is also important to establish the final resting point of the vehicle and the location of ejected occupants and vehicle parts, bloodstains, or other biological materials on the ground and their relationship to the vehicle path. In this regard, it is instructive to inspect the scene in person, particularly if this can be done shortly after the crash. If this is not possible, one must rely on the examination of photographs taken shortly after the crash if such photographs are available. When photographs are not available, to acquire the required information, one must rely on official investigation reports as well as witness interviews, transcribed witness statements, and depositions of witnesses when available.

⁴The reader is referred to textbooks of biomechanics as well as to the Society of Automotive Engineers (SAE 2003) for general information on regional tolerances. More specific information may be found by searching the medical and engineering literature (Pub Med, ASME, and SAE).

Since most injury investigators do not perform crash reconstructions, detailed data relating to crash parameters generally must be obtained from a crash reconstruction expert, many of whom can also download stored crash data from the air bag sensors in some automobiles. Crash data recorders are generally not available in general aviation aircraft.

Delta v and Principal Direction of Force

The change in velocity (delta v) of the predominant crash impact is a key indicator of the severity of a particular crash. In automobile crashes, which are usually planar, a resultant delta v in the horizontal plane is usually determined. Since aircraft crashes are usually three-dimensional, delta v's for the three orthogonal axes are usually separately determined. When a delta v is determined for automobile crashes, the direction of the crash vector within the horizontal plane is also determined. This is referred to as the PDOF and is often based on a clock direction with 12 o'clock being straight ahead. A frontal crash is generally defined as occurring from 10 to 2 o'clock and a rear crash from 4 to 8 o'clock. A right side impact would then be considered to occur from 2 to 4 o'clock and a left side impact from 8 to 10 o'clock. PDOF may also be given as an angle with 0° being straight ahead. Each hour on the clock encompasses 30° of angle.

Delta v has often been used in automobile applications to classify crashes as to severity with respect to the potential for occupant injury. One should be very cautious in applying delta v to predict injury or to compare the severity of different crashes because delta v is related to total kinetic energy of the crash and human impact tolerance is not dependent on the kinetic energy of the crash per se. Consider two automobiles of the same model year and type traveling at a speed of 72 kph (45 mph). The driver of vehicle 1 observes the traffic light ahead of him turn to yellow. He applies the brakes and comes to a stop at the stop line. The driver of vehicle 2 falls asleep and runs off the road at 72 kph crashing head on into a concrete bridge abutment. Even though both drivers experienced approximately the same delta v, their outcomes were considerably different. Driver 1 drives away when the light changes to green, while driver 2 receives fatal injuries. The primary differences between the two scenarios are stopping time and distance, which determine the acceleration experienced by the occupants. In the first example, the vehicle stops over a period of several seconds and a distance of many meters, while vehicle 2 comes to stop in approximately 100 ms and a distance on the order of 1 meter, basically the crush distance experienced by the front of his car. This results in an acceleration of less than 1 G for vehicle 1 and about 40 G for vehicle 2! Although a rather extreme example, this illustration demonstrates the inherent problem of using delta v to predict injury in a crash. Estimating the crash pulse (magnitude and duration of acceleration) is a far more reliable method of predicting injury in a crash, particularly when comparing crashes involving dissimilar vehicles or different crash conditions (Woolley and Asay 2008; Cheng et al. 2005; Berg et al. 1998).

Occupant Kinematics

Occupants within a crashing vehicle move with respect to the interior of the vehicle according to the dynamics of the vehicle and their restraint status. Currently in the USA, approximately 84% of occupants use their restraint systems (NHTSA 2009). Since a significant number of occupants do not use the available restraints and since restraint status is a major factor in the determination of injury mechanisms, it is essential for investigators to determine the restraint status of vehicle occupants. In this regard, it is important to realize that ejection from a vehicle does not necessarily indicate that the occupant was unrestrained. Although rare, there have been numerous examples of restrained

occupants being ejected from vehicles while restrained usually as a result of misuse of the restraint or due to failure of a component of the belt restraint system or the seat. It is advisable for the injury investigator to seek physical evidence on the person and in the vehicle to support any determination of restraint status regardless of witness testimony.

How an occupant moved within the vehicle during a crash is a significant factor in determining injury (Backaitis et al. 1982; Biss 1990; Bready et al. 2002; Estep and Lund 1996). Determination of occupant kinematics is essential in order to determine the possibilities for occupant contact within a vehicle and to rule out those objects that do not fall into the occupant's potential strike zone considering the dynamics of the crash. Recall that according to Newton's First Law, a vehicle occupant will continue to move in the same direction and velocity as he was before the beginning of the crash sequence. Contact of the vehicle with an external object will cause the vehicle to decelerate and, frequently, change its direction of travel while the occupant continues along his previous travel vector until acted on by an external force, usually his restraint system and/or internal objects. For instance, in a direct frontal crash, the occupants will move forward with respect to the vehicle interior as the vehicle slows. For a rear impact, the vehicle will be accelerated forward causing the seat to accelerate into the occupant, and the occupant will appear to load rearward into the seat. Similarly, a side impact will cause the impacted side to accelerate toward the occupant and load into the side of the occupant. These examples suggest the general rule that an occupant initially moves toward the area of impact with respect to the vehicle interior. This general rule is somewhat modified by the fact that many impacts result in rotation of the vehicle around its yaw axis immediately after impact (Cheng and Guenther 1989). This causes the vehicle to rotate under the occupant so that the occupant appears to move opposite the direction of rotation with respect to the vehicle interior. As an example, a left frontal impact will cause the vehicle to slow longitudinally and rotate in a clockwise direction. With respect to the vehicle interior, a driver will initially move forward toward the impact, but due to the rotation of the vehicle under him, his trajectory will be modified so that he moves in an arc farther to the left than if there were no rotation of the vehicle. This can be explained by the phase delay between the movement of the vehicle and the corresponding movement of the occupants. If an occupant is restrained, his relative movement within the vehicle will be restricted by his restraint, whereas an unrestrained occupant will move unrestricted within the vehicle and impact the interior according to the vehicle dynamics.

Rollover crashes have been increasing in frequency over the past few decades as the proportion of small trucks and SUVs proliferates due to the inherent instability and rollover propensity of these vehicles compared to automobiles (Kallan and Jermakian 2008; Robertson 1989; NHTSA 2003, 2005). Rollover collisions of motor vehicles involve some rather special considerations in regard to occupant kinematics (Adamec et al. 2005; Newberry et al. 2005; Praxl et al. 2003; Takagi et al. 2003; Howard et al. 1999). Most of the injuries in these crashes are associated with head and upper torso impacts with interior structures aggravated by deformation of structures into the occupant compartment (Digges et al. 1994; Ridella et al. 2010).

Occupant kinematics in rollovers are usually described for the prerollover phase, the trip, and the rollover phase. The motions of occupants for the prerollover phase are determined the same way as for any planar crash. Occupant motion in the trip is determined by the direction of the trip, both near side and far side, and by the magnitude of the force causing the trip. In a near-side trip, the occupant tends to move laterally toward the trip based on deceleration caused by friction of the wheels with the roadway and, as the vehicle rolls, by the increasing force of gravity tending to move him toward the low side of the vehicle. A near-side occupant's motion is restricted by his seat belt, which limits hip movement away from the seat bottom, and by the side of the vehicle, which limits motion of his upper torso. In a far-side trip, the occupant is held in close proximity to the seat bottom by the lap belt, but his upper torso will move inboard since there are no surfaces to restrict this movement. If the forces are sufficient, the occupant may slip out of his upper torso restraint, which may result in subsequent upper torso flailing during the rollover phase of the crash (Obergefell et al. 1986).

After trip, the vehicle transitions into the rollover phase. Accelerations in a rollover crash are invariably low to moderate in relation to the tolerance levels of restrained humans; consequently, occupants are not seriously injured as long as they do not forcefully contact potentially injurious objects inside or outside the vehicle. Serious injuries to properly restrained occupants occur when structures intrude into occupied areas causing severe contact or flailing injuries, when restraint systems fail to provide adequate restraint, when occupants lose their restraint through a variety of mechanisms, when roof deformations expose occupants' heads or other body parts outside the vehicle, or through a combination of these mechanisms. Occupants, who are unrestrained, inadequately restrained, or become unrestrained during the collision sequence, frequently receive serious injuries from flailing into internal structures or from being partially or completely ejected from the vehicle and striking external surfaces or structures.

Finally, it should be noted that there are several sophisticated computer simulation programs available to help the investigator determine occupant kinematics in a crash (Prasad and Chou 2002). The two most frequently used general body models are MADYMO and the articulated body model (ATB). MADYMO is a program developed by TNO in the Netherlands and provides sophisticated 2-D and 3-D visual interfaces. This program is used primarily by the automobile industry. ATB is a program developed by the US Air Force and is frequently used for aviation applications. There are also numerous body segment models available (Prasad and Chou 2002). These programs have the advantage of being highly repeatable and allow variation of numerous factors related to the occupant, the seat, the restraint system, the vehicle, and the crash. They also provide timing for various occurrences that can answer such questions as "was the side of the car still deforming when the occupant struck it?" Computer simulations can be extremely valuable in reconstruction of occupant kinematics and injury mechanisms. Unfortunately, most of these occupant simulation programs are quite complex and require the services of a highly trained and experienced operator. Also, like all simulation models, they are only as accurate as the data supplied to the system by the investigator. Consequently, simulation outputs should always be checked against physical evidence to ensure a close correlation before the output of the simulation program is accepted.

Scene Inspection

A scene inspection will give the injury investigator a general impression of the terrain at the crash site including relative elevations and the presence of features such as gullies, drainage ditches, bridges, and other surface features or obstructions that may have played a role in occupant kinematics or injury. Tire marks and impact gouges on the roadway or on the ground as well as furrowing in soil can help an investigator to visualize the position of the vehicle throughout the crash sequence. One can also get an impression of wreckage distribution and ejected occupant resting positions either by observing the scene and wreckage prior to scene clean up or by visualizing wreckage distribution and body positions using a crash scene survey that is frequently produced by police investigators and reconstructionists.

In summary, inspection of the scene combined with a review of a complete crash reconstruction can provide valuable information to assist an injury investigator in determining mechanism of injury for the injured occupants of any vehicular crash. Scene investigation and review of the crash reconstruction primarily provides the injury investigator with an understanding of the scene, the vehicle dynamics, and the forces involved in the crash, all of which provide valuable insight into the probable occupant kinematics for each occupant of the vehicle. An understanding of occupant kinematics in the crash is essential in determining impact points within and without the vehicle.

The Vehicle

When the vehicle is available, it should be inspected for evidence of exterior impacts, intrusion into the occupant compartment, deformations within the occupant compartment potentially caused by occupant contact, deposits of hair, blood, or other tissues inside or outside the vehicle, seat condition and position, and restraint status including air bag deployments and belt restraint condition. The position of the seat with respect to its adjustments and the position of controls and switches may also be useful, with the caveat that the positions may have been altered prior to your investigation.

Survival Space

Hugh De Haven was one of the first engineers to articulate the basic principles of crashworthiness in the late 1940s and early 1950s. He compared the principles of human protection to already established principles of packaging, relating human protection in automobiles to “the spoilage and damage of people in transit” (De Haven 1952). According to De Haven, the first principle of packaging states “that the package should not open up and spill its contents and should not collapse under expected conditions of force and thereby expose objects inside it to damage.” This principle is today frequently referred to as preservation of occupant “survival space.” Franchini in 1969 stated that it is “essential to ensure a minimum residual space after collision, for the vehicle occupant” in order to prevent occupant crushing (Franchini 1969). This is essential because, when the clearance between an interior surface and the occupant is significantly reduced, the occupant, regardless of restraint status, can flail into the intruding structure, be impacted by the intruding structure, or a combination of both. Contact injuries caused by these impacts can be extremely serious particularly when the contact occurs to the head or thorax. Consequently, part of an examination of a crashed vehicle should include an assessment of occupant compartment intrusions noting the location and degree of intrusion in relation to the injured occupant and presumed occupant kinematics in order to determine the likely source of contact injuries. Placing a surrogate of the same height and weight into the crashed vehicle or into a similar, intact vehicle to visualize and measure clearances from suspected injurious structures will facilitate this assessment. If a surrogate seated in the vehicle with a locked retractor can reach a suspected area of contact with the same portion of his body that was injured on the subject, it can be safely assumed that that area could be reached under the dynamic conditions of a crash as long as the forces in the crash are consistent with occupant movement in that direction. The amount of movement of a restrained occupant under dynamic conditions may be greater than can be replicated with static testing, and the greater the forces of the crash, the greater the potential excursion. In most planar impacts, dynamic conditions produce greater tissue compression from restraints, more ligamentous and other soft tissue stretching, and more payout and stretching of the belt restraint system. All these factors lead to greater occupant excursion. There is no hard rule to estimate excursion beyond that demonstrated with a static test, but additional excursions of the torso and head of 5–10 cm (2–4 in.) would not be excessive in moderate to severe planar crashes.

Crash Survivability

Crash survivability is a very useful concept that allows investigators to estimate whether a particular crash was potentially survivable for the occupants of a crashed vehicle. This concept is widely used

in aviation crash investigation, but not very often in motor vehicle crashes. Survivability of a crash is based on two subjective factors:

1. The forces in the occupant compartment were within the limits of human tolerance.
2. Survival space was maintained throughout the crash sequence.

As discussed in section “Human Tolerance to Acceleration and Blunt Force Impact,” the first criterion requires a reconstruction of the crash forces and the crash pulse and comparison of these parameters against accepted human tolerance standards. Clearly, this is a highly subjective determination that may be facilitated by applying the guidelines provided in Table 6.3. The US Army uses a limit of no more than 15% dynamic deformation into occupied spaces during the crash sequence to meet the second criterion. This determination is also somewhat subjective since one has to consider that most vehicle structures are metallic, and after metals deform, they tend to rebound back toward their original shape when the deforming force is removed. Consequently, the residual deformation (plastic deformation) seen by investigators may be as much as 20% less than actually occurred during the crash (elastic deformation). When both survivability factors are met, the crash is classified as “survivable.” When neither criterion is met, the crash is considered to be “nonsurvivable.” When both are met for some parts of the occupant compartment but not others, the crash may be classified as “partially survivable” (Department of the Army 1994).

The primary utility of the concept of survivability is that its determination is completely independent of the outcome of the occupants since it is based only on the crash and the vehicle. Consequently, there may be a survivable crash where all the occupants died, or there may be a nonsurvivable crash where all the occupants survived. In the first case, since the basic criteria for survival were present, it raises the question of why the occupants did not survive. The answer is frequently due to the failure of the occupants to use their restraint systems or due to failure of one or more components of the occupant protection system. If people are consistently dying in survivable crashes, it should alert responsible parties to the fact that there is a problem that needs identification and mitigation.

It is also very instructive to thoroughly examine nonsurvivable crashes where one or more occupants survive without serious injuries. This suggests that either serendipitous factors were at play or that there was something extraordinary in the design of the vehicle that is protecting occupants in spite of a very severe crash. Determination of these factors can lead to crashworthiness improvements.

Designating a crash as nonsurvivable does not mean that the vehicle could not have been designed to render the crash survivable.

Deformations Caused by Occupant Contact

The interior and exterior of the crashed vehicle should be carefully examined to detect evidence of contact of occupants with vehicle structures. This information is useful in establishing occupant kinematics as well as identifying structures potentially responsible for blunt force injuries. Many interior structures deform significantly when impacted by occupants. These include controls and switches, side panels, roof panels, and other trim panels and padding. Also, fabric headliners and other interior fabrics are easily scuffed by human contact during a crash. Visualizing injurious and noninjurious contact points will yield useful trajectory information, which should be correlated with proposed occupant kinematics when reconstructing injury mechanisms.

An illustration of deformation frequently seen in crashes is deformation caused by head impact. In most cases, if head contact is sufficient to cause deformation to internal vehicle structures, it is also sufficient to leave evidence on the scalp and/or result in more serious injuries to the head or cervical spine. Figure 6.7 illustrates an area of head contact with the upper portion of the B-pillar and

Fig. 6.7 Head imprint on upper B-pillar and upper window frame due to a left-sided impact to the vehicle. Also note damage to the plastic trim



the upper driver's window frame. This impact resulted in abrasive injuries to the scalp of the driver as well as a severe flexion–compression injury to the lower cervical spine.

Body Fluids and Tissues

The presence of bodily fluids and tissues within the vehicle is a powerful clue for determining occupant kinematics and occupant contacts. The vehicle interior and exterior should be carefully inspected for blood or tissue depositions. Air bags, particularly frontal air bags, should be carefully inspected for the presence of saliva, blood, and cosmetic products such as lipstick, facial powders, and eye shadow. A criminologist or forensic pathologist may be consulted for advanced detection methods if required (James et al. 2005; Wonder 2007).

Steering Wheel, Seats, and Restraints

Another item which may provide clues to injury mechanism is the steering wheel. The steering wheels of most automobiles are designed to deform when forcefully impacted (Phillips et al. 1978; Horsch et al. 1991). Forward deformation of the steering wheel rim is indicative of occupant loading.

This loading may come from the hands of the operator or from head or upper body impact. When an impact is sufficiently forceful to cause steering wheel deformation, corresponding injuries on the hands, arms, head, chest, or upper abdomen of the operator may also be found. Under severe loading, the steering wheel shaft is designed to stroke in a manner similar to a shock absorber. Additionally, steering wheels are supported by brackets on either side of the shaft that attach the shaft to the dashboard through a sliding mechanism (capsule) that releases the steering column when the wheel is forcefully impacted by the driver (Phillips et al. 1978; Horsch et al. 1991).

Seats should be inspected for position and evidence of loading. Most front seats in automobiles and pilot seats in aircraft are adjustable. The position of these adjustments should be documented so that the seat of an exemplar vehicle may be placed in the same position during a surrogate inspection.

Examination of seats for deformations can also provide clues as to occupant kinematics as well as the forces involved in a crash. Seatbacks in some automobiles are very weak so that in rear-end collisions, the seat can deform significantly rearward, sometimes allowing the seat and/or the occupant to strike a person seated behind the seat. This mechanism has led to numerous instances of child death due to an adult striking the head and/or chest of the child in a rear-end collision. Rearward deformation of a seat can also allow the occupant to slide rearward under his restraint system and impact his head in the rear of the vehicle. This mechanism has caused serious head and cervical spine injuries to front seat occupants involved in severe rear-end collisions. Inspection of the involved seat will reveal a bent seat frame and/or damaged seat recliner mechanisms.

Deformations to seats can provide useful information regarding the direction and magnitude of crash forces. Downward deformation of the center and rear portion of the seat pan should alert the investigator to a crash with a relatively high vertical velocity component, a so-called “slam-down” crash. High vertical velocity crashes may occur in automobiles when they run off the road and travel off an embankment or when a vehicle frontally impacts an upward slope. These crashes frequently lead to thoracolumbar compression or anterior wedge fractures as often observed in aircraft crashes when significant vertical forces are involved. Downward deformation of the front of the seat pan is often seen in severe frontal crashes where the occupant slides forward and downward in response to the frontal impact. This finding can be a clue to look for anterior injuries to the chest, head, and lower extremities of the occupant.

A frequent issue in vehicle crashes is whether the occupant was restrained. Inspection of a belt restraint system will usually reveal evidence of dynamic loading in a severe crash, particularly in frontal crashes where the belts tend to be most heavily loaded (Felicella 2003; Bready et al. 1999, 2000; Heydinger et al. 2008; Moffatt et al. 1984). All restraint components should be visually inspected and photographed during an inspection to help determine belt use (Figs. 6.8 and 6.9). When visual examination is inconclusive, the seat belt may be removed and inspected under magnification by an expert. Microscopic inspection frequently helps to clarify the issue. Comparison with other restraints within the vehicle is also useful to determine differences between those belts that were known to be worn or not worn during the crash and the belt in question. When restraint systems are loaded to the point, they leave the abovementioned telltale signs; the occupant will usually sustain abrasions and contusions along the belt path consistent with body loading into the restraint.

Finally, the status of pretensioners should be determined in automobiles that are equipped with them. Pretensioners may be located at the buckle or within the retractor. Buckle pretensioners work by shortening the buckle stalk which is attached to the seat frame or floor, thus pulling down on both the lap belt and shoulder belt. The stalk covering material is usually accordion-shaped, and the folds will be compressed after firing when compared to a pretensioner that did not fire. When retractor-mounted pretensioners are activated, they frequently lock the retractor in place. This provides a good indicator as to whether the belt was worn. If the belt is locked in the stowed position, it clearly was not worn during the crash. If it is locked in a partially extended position, it is clear that it was worn,



Fig. 6.8 Latch plate loading by the webbing resulting in partially melted plastic on the load bearing surface



Fig. 6.9 Plastic transfer to the webbing material due to friction at the load bearing surface of the D-ring

and the amount of extension can be used to determine whether it was worn properly, by making measurements and comparing those to a surrogate in an exemplar vehicle or by putting the surrogate into the subject restraint system.

Analysis

Once all acute injuries have been identified and the scene and the vehicle are inspected, the process of determining injury mechanisms can begin. This process involves correlating identified injuries or concurrent groups of injuries with evidence from the vehicle and the scene and with the crash

conditions determined by a reconstruction expert. Through knowledge of human tolerance and the general mechanisms required to produce certain types of injuries, the investigator can correlate the identified injuries with the crash, body location and position, the vehicle condition, restraints condition, and forensic evidence within the vehicle. Once a tentative mechanism of injury for a particular injury or group of injuries has been established, the diagnosis should be supported by published studies or, sometimes, through specialized testing. This correlation process requires intimate knowledge of human response to various loading conditions most often acquired through experience and study. Although a probable general mechanism of injury can be determined for most major injuries, it is not uncommon that mechanisms for other injuries may not be determinable from the existing evidence.

Conclusion

All injuries incurred in vehicular crashes have a cause or mechanism beyond the obvious descriptions of “traffic accident,” “airplane crash,” or even, blunt force injury. The determination of a detailed mechanism of injury is a process that requires the acquisition of considerable information about the injury itself, the circumstances of the crash, and the vehicles involved in the crash. Injury mechanism data are vital for effective surveillance of transportation systems as well as for identifying and prioritizing crashworthiness improvements and for developing appropriate government safety regulations.

References

- AAAM. (1990). *The abbreviated injury scale, 1990 revision*. Barrington, IL: Association for the Advancement of Automotive Medicine.
- Adamec, J., Praxl, N., Miehl, T., Muggenthaler, H., & Schonpflug, M. (2005, September 21). *The occupant kinematics in the first phase of a rollover accident – experiment and simulation*. 2005 International IRCOBI conference on the biomechanics of impacts (pp. 145–156), IRCOBI, Prague.
- Agaram, V., Xu, L., Wu, J., Kostyniuk, G., & Nusholtz, G. S. (2000, March 6). *Comparison of frontal crashes in terms of average acceleration*. SAE 2000 World Congress. SAE paper no. 2000-01-0880 (pp. 1–21), SAE, Warrendale, PA.
- Alem, N. M., Nusholtz, G. S., & Melvin, J. W. (1984, November 6). *Head and neck response to axial impacts*. Proceedings of the 28th Stapp Car Crash Conference. SAE paper no. 841667, SAE, Warrendale, PA.
- Alsop, D., & Kennett, K. (2000). Skull and facial bone trauma. In A. M. Nahum & J. W. Melvin (Eds.), *Accidental injury: biomechanics and prevention* (pp. 254–276). New York: Springer.
- Anderson, P. A., Rivara, F. P., Maier, R. V., & Drake, C. (1991). The epidemiology of seatbelt-associated injuries. *The Journal of Trauma*, *31*, 60–67.
- Backaitis, S. H., DeLarm, L., & Robbins, D. H. (1982, February 22). *Occupant kinematics in motor vehicle crashes*. SAE International Congress and Exposition. SAE paper no. 820247 (pp. 107–155), SAE, Warrendale, PA.
- Baker, S. P., & O’Neill, B. (1976). The injury severity score: an update. *The Journal of Trauma*, *16*, 882–885.
- Baker, S. P., O’Neill, B., Haddon, W., Jr., & Long, W. B. (1974). The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *The Journal of Trauma*, *14*, 187–196.
- Baker, S. P., Brady, J. E., Shanahan, D. F., & Li, G. (2009). Aviation-related injury morbidity and mortality: data from U.S. health information systems. *Aviation, Space, and Environmental Medicine*, *80*, 1001–1005.
- Berg, F. A., Walz, F., Muser, M., Burkle, H., & Epple, J. (1998, September 16). *Implications of velocity change delta-v and energy equivalent speed EES for injury mechanism assessment in various collision configurations*. International IRCOBI conference on the biomechanics of impacts. 1998-13-0004 (pp. 57–72), IRCOBI, Bron, France.
- Biss, D. J. (1990, February 19). *Relating three-point belted vehicle occupant kinematics and dynamics to case accident injury patterns and forensic evidence*. 42nd Annual meeting of American Academy of Forensic Sciences, American Academy of Forensic Sciences, Cincinnati, OH.

- Bready, J. E., Nordhagen, R. P., & Kent, R. W. (1999, March 1). *Seat belt survey: identification and assessment of noncollision markings*. SAE International Congress and Exposition. SAE paper no. 1999-01-0441 (pp. 1–13), SAE, Warrendale, PA.
- Bready, J. E., Nordhagen, R. P., Kent, R. W., & Jakstis, M. W. (2000). *Characteristics of seat belt restraint system markings*. SAE 2000 World Congress. SAE paper no. 2000-01-1317 (pp. 1–11), SAE, Warrendale, PA.
- Bready, J. E., Nordhagen, R. P., Perl, T. R., & James, M. B. (2002, March 4). *Methods of occupant kinematics analysis in automobile crashes*. SAE 2002 World Congress. SAE paper no. 2002-01-0536 (pp. 1–6), SAE, Warrendale, PA.
- Brinkley, J. W., & Raddin, J. H., Jr. (1996). Biodynamics: transient acceleration. In R. L. DeHart (Ed.), *Fundamentals of aerospace medicine*. Philadelphia, PA: Lippincott Williams & Wilkins.
- Burdi, A. R., Huelke, D. F., Snyder, R. G., & Lowrey, G. H. (1969). Infants and children in the adult world of automobile safety design: pediatric and anatomical considerations for the design of child restraints. *Journal of Biomechanics*, 2, 267–280.
- Car-Safety.org. (2009). Why rear-facing is safest. <http://www.car-safety.org/rearface.html>. Accessed 13 June 2009.
- Chandler, R. F. (1985). Restraint system basics. *Sport Aviation*, 1985, 35–39.
- Chandler, R. F. (1990). Occupant crash protection in military air transport. AGARD-AG-306.
- Cheng, P. H., & Guenther, D. A. (1989, February 27). *Effects of change in angular velocity of a vehicle on the change in velocity experienced by an occupant during a crash environment and the localized Delta V concept*. SAE International Congress and Exposition. SAE paper no. 890636 (pp. 39–54), SAE, Warrendale, PA.
- Cheng, P. H., Tanner, C. B., Chen, H. F., Durisek, N. J., & Guenther, D. A. (2005, April 11). *Delta-V barrier equivalent velocity and acceleration pulse of a vehicle during an impact*. SAE 2005 World Congress. SAE paper no. 2005-01-1187, SAE, Warrendale, PA.
- Clarke, N. P. (1963). Biodynamic response to supersonic ejection. *Aerospace Medicine*, 34, 1089–1094.
- Compton, C. P. (2005). Injury severity codes: a comparison of police injury codes and medical outcomes as determined by NASS CDS Investigators. *Journal of Safety Research*, 36, 483–484.
- Crandall, J., Kent, R., Patrie, J., Fertile, J., & Martin, P. (2000). Rib fracture patterns and radiologic detection – a restraint-based comparison. *Annual Proceedings of the Association for the Advancement of Automotive Medicine*, 44, 235–259.
- Crandall, J., Kent, R., Viano, D., & Bass, C. R. (2003). The biomechanics of inflatable restraints – occupant protection and induced injury. In R. Kent (Ed.), *Air bag development and performance. New perspectives from industry, government and academia* (pp. 69–110). Warrendale, PA: SAE.
- Cugley, J., & Glaister, D. H. (1999). Short duration acceleration. In J. Ernstring, A. N. Nicholson, & D. J. Rainford (Eds.), *Aviation medicine*. London: Arnold.
- De Haven, H. (1952, January 14). *Accident survival – airplane and passenger car*. SAE Annual Meeting. SAE paper no. 520016 (pp. 1–7), SAE, Warrendale, PA.
- Department of Defense. (2000, February 10). Standard practice for system safety. MIL-STD-882D, pp. ii–26.
- Department of the Army. (1989). *Aircraft crash survival design guide: volume 2 – Aircraft design crash impact conditions and human tolerance*. USAAVSCOM TR 89-D-22B.
- Department of the Army. (1994). Army accident investigation and reporting. Department of the Army.
- Digges, K. H., Malliaris, A. C., & DeBlois, H. J. (1994, May 24). *Opportunities for casualty reduction in rollover crashes*. 14th International Technical Conference on the Enhanced Safety of Vehicles. Paper no. 94-S5-O-11 (pp. 863–868), NHTSA, Washington, DC.
- Eiband, A. M. (1959, June 1). *Human tolerance to rapidly applied accelerations: a summary of the literature*. NASA Memo 5-19-59E (pp. 1–93), NASA, Washington, DC.
- Estep, C. R., & Lund, A. K. (1996, May 13). *Dummy kinematics in offset-frontal crash tests*. 15th International Technical Conference on the Enhanced Safety of Vehicles. Paper no. 96-S3-W-12 (pp. 502–510), NHTSA, Washington, DC.
- Estrada, L. S., Alonzo, J. E., McGwin, G., Jr., Metzger, J., & Rue, L. W., III. (2004). Restraint use and lower extremity fractures in frontal motor vehicle collisions. *The Journal of Trauma*, 57, 323–328.
- Farmer, C. M. (2003). Reliability of police-reported information for determining crash and injury severity. *Traffic Injury Prevention*, 4, 38–44.
- Felicella, D. J. (2003). *Forensic analysis of seat belts*. Salem, OR: Kinetic Energy.
- Franchini, E. (1969, January 13). *Crash survival space is needed in vehicle passenger compartments*. International body engineering conference and exposition. SAE paper no. 690005, SAE, Warrendale, PA.
- Garrett, J. W., & Braunstein, P. W. (1962). The seat belt syndrome. *The Journal of Trauma*, 2, 220–238.
- Hayden, M. S., Shanahan, D. F., Chen, L. H., & Baker, S. P. (2005). Crash-resistant fuel system effectiveness in civil helicopter crashes. *Aviation, Space, and Environmental Medicine*, 76, 782–785.
- Heydinger, G. J., Uhlenhake, G. D., & Guenther, D. A. (2008, April 14). *Comparison of collision and noncollision marks on vehicle restraint systems*. SAE 2008 World Congress. SAE paper no. 2008-01-0160, SAE, Warrendale, PA.

- Horsch, J. D., Viano, D. C., & DeCou, J. (1991, November 18). *History of safety research and development on the General Motors energy-absorbing steering system*. 35th Stapp Car Crash Conference. SAE paper no. 912890 (pp. 1–46), SAE, Warrendale, PA.
- Howard, R. P., Hatsell, C. P., & Raddin, J. H. (1999, September 28). *Initial occupant kinematics in the high velocity vehicle rollover*. International body engineering conference and exposition. SAE paper no. 1999-01-3231 (pp. 1–18), SAE, Warrendale, PA.
- Huelke, D. F., & Melvin, J. W. (1980, February 25). *Anatomy, injury frequency, biomechanics, and human tolerances*. Automotive engineering congress and exposition. SAE paper no. 800098, SAE, Warrendale, PA.
- James, S. H., Kish, P. E., & Sutton, T. P. (2005). *Principles of bloodstain pattern analysis: theory and practice*. Boca Raton, FL: CRC.
- Kallan, M. J., & Jermakian, J. S. (2008). SUV rollover in single vehicle crashes and the influence of ESC and SSF. *Annals of Advances in Automotive Medicine*, 52, 3–8.
- Kallieris, D., Conte-Zerial, P., Rizzetti, A., & Mattern, R. (1998, May 31). *Prediction of thoracic injuries in frontal collisions*. 16th International technical conference on the enhanced safety of vehicles. Paper no. 98-S7-O-04 (pp. 1550–1563), NHTSA, Washington, DC.
- Kindelberger, J., & Eigen, A. (2003). *Younger drivers and sport utility vehicles* (Report no. DOT HS 809 636). Washington, DC: NCSA.
- King, A. I., & Yang, K. H. (1995). Research in biomechanics of occupant protection. *The Journal of Trauma*, 38, 570–576.
- Latham, F. (1957). *A study in body ballistics: seat ejection* (pp. 121–139). Farnborough: R.A.F. Institute of Aviation Medicine.
- Levy, P. M. (1964). Ejection seat design and vertebral fractures. *Aerospace Medicine*, 35, 545–549.
- Love, J. C., & Symes, S. A. (2004). Understanding rib fracture patterns: incomplete and buckle fractures. *Journal of Forensic Sciences*, 49, 1153–1158.
- Melvin, J. W., Baron, K. J., Little, W. C., Gideon, T. W., & Pierce, J. (1998, November 2). *Biomechanical analysis of Indy race car crashes*. 42nd Stapp Car Crash Conference. SAE paper no. 983161 (pp. 1–20), SAE, Warrendale, PA.
- Moffatt, C. A., Moffatt, E. A., & Weiman, T. R. (1984, February 27). *Diagnosis of seat belt usage in accidents*. SAE International Congress and Exposition. SAE paper no. 840396, SAE, Warrendale, PA.
- Newberry, W., Carhart, M., Lai, W., Corrigan, C. F., Croteau, J., & Cooper, E. (2005, April 11). *A computational analysis of the airborne phase of vehicle rollover: occupant head excursion and head-neck posture*. SAE 2005 World Congress. SAE paper no. 2005-01-0943, SAE, Warrendale, PA.
- NHTSA. (1997). NHTSA announces new policy for air bags. *NHTSA Now*, 3, 1–3.
- NHTSA. (1999). *Fourth report to congress: effectiveness of occupant protection systems and their use*. Washington, DC: NHTSA.
- NHTSA. (2003). *Initiatives to address the mitigation of vehicle rollover*. Washington, DC: NHTSA.
- NHTSA. (2005). NPRM roof crush resistance. Docket No. NHTSA-2005-22143 (pp. 1–93). Washington, DC: NHTSA.
- NHTSA, NCSA. (2009). Seat belt use in 2009 – overall results. Traffic Safety Facts DOT HS 811 100.
- Obergefell, L. A., Kaleps, L., & Johnson, A. K. (1986, October 27). *Prediction of an occupant's motion during rollover crashes*. 30th Stapp Car Crash Conference. SAE paper no. 861876 (pp. 13–26), SAE, Warrendale, PA.
- Phillips, L., Khadilkar, A., Egbert, T. P., Cohen, S. H., & Morgan, R. M. (1978, January 24). *Subcompact vehicle energy-absorbing steering assembly evaluation*. 22nd Stapp Car Crash Conference. SAE paper no. 780899 (pp. 483–535), SAE, Warrendale, PA.
- Prasad, P., & Chou, C. C. (2002). A review of mathematical occupant simulation models. In A. M. Nahum & J. W. Melvin (Eds.), *Accidental injury: biomechanics and prevention*. New York: Springer.
- Praxl, N., Schonpflug, M., & Adamec, J. (2003). *Simulation of occupant kinematics in vehicle rollover dummy model versus human model*. 18th International technical conference on the enhanced safety of vehicles, NHTSA, Washington, DC.
- Ridella, S. A., Eigen, A. M., Kerrigan, J., & Crandall, J. (2010). *An analysis of injury type and distribution of belted, non-ejected occupants involved in rollover crashes*. SAE Government/Industry Meeting and Exposition.
- Robertson, L. S. (1989). Risk of fatal rollover in utility vehicles relative to static stability. *American Journal of Public Health*, 79, 300–303.
- Rouhana, S. W., Kankanala, S. V., Prasad, P., Rupp, J. D., Jeffreys, T. A., & Schneider, L. W. (2006). Biomechanics of 4-point seat belt systems in farside impacts. *Stapp Car Crash Journal*, 50, 267–298.
- Rupp, J. D., Miller, C. S., Reed, M. P., Madura, N. H., Klinich, K. D., & Schneider, L. W. (2008). Characterization of knee-thigh-hip response in frontal impacts using biomechanical testing and computational simulations. *Stapp Car Crash Journal*, 52, 421–474.
- Rupp, J. D., Flannagan, C. A., & Kuppa, S. M. (2010). Injury risk curves for the skeletal knee-thigh-hip complex for knee-impact loading. *Accident Analysis & Prevention*, 42, 153–158.

- SAE. (1995). Surface Vehicle Recommended Practice. Instrumentation for impact test-Part 1-Electronic Instrumentation. SAE J211, SAE, Warrendale, PA.
- SAE. (2003). *Human tolerance to impact conditions as related to motor vehicle design*. SAE J885 REV 2003_12. SAE, Warrendale, PA.
- Shanahan, D. F., & Shanahan, M. O. (1989). Injury in U.S. Army helicopter crashes October 1979-September 1985. *The Journal of Trauma*, 29, 415-422.
- Smith, W. S., & Kaufner, H. (1967). A new pattern of spine injury associated with lap-type seat belts: a preliminary report. *University of Michigan Medical Center Journal*, 33, 99-104.
- Snyder, R. G. (1970a). The seat belt as a cause of injury. *Marquette Law Review*, 53, 211-225.
- Snyder, R. G. (1970, May 13). *Human impact tolerance*. International automobile safety conference. SAE paper no. 700398 (pp. 712-782), SAE, Warrendale, PA.
- Sochor, M. C., Faust, D. P., Wang, S. C., & Schneider, L. W. (2003, March 3). *Knee, thigh and hip injury patterns for drivers and right front passengers in frontal impacts*. SAE 2003 World Congress. SAE paper no. 2003-01-0164, SAE, Warrendale, PA.
- Stapp, J. P. (1961a). Biodynamics of deceleration, impact, and blast. In H. G. Armstrong (Ed.), *Aerospace medicine*. (pp. 118-165). Baltimore, MD: Williams & Wilkins Co.
- Stapp, J. P. (1961b). Human tolerance to severe, abrupt acceleration. In O. H. Gauer & G. D. Zuidema (Eds.), *Gravitational stress in aerospace medicine* (pp. 165-188). Boston, MA: Little Brown.
- Stucki, S. L., Hollowell, W. T., & Fessahaie, O. (1998, May 31). *Determination of frontal offset conditions based on crash data*. 16th International technical conference on the enhanced safety of vehicles. Paper no. 98-S1-O-02 (pp. 164-184), NHTSA, Washington, DC.
- Takagi, H., Maruyama, A., Dix, J., & Kawaguchi, K. (2003). *Madymo modeling method of rollover event and occupant behavior in each rollover initiation type*. 18th international technical conference on the enhanced safety of vehicles, NHTSA, Washington, DC.
- Terriere, C. (1995). *Children are not miniature adults*. International Research Council on the Biomechanics of Impacts. Paper no. 1995-13-0001 (pp. 15-29), Automobile Biomedical Department, Renault Research and Development Division.
- Thompson, N. S., Date, R., Charlwood, A. P., Adair, I. V., & Clements, W. D. (2001). Seat-belt syndrome revisited. *International Journal of Clinical Practice*, 55, 573-575.
- Tile, M. (1996). Acute pelvic fractures: I. Causation and classification. *The Journal of the American Academy of Orthopaedic Surgeons*, 4, 143-151.
- Viano, D. C., & Parenteau, C. S. (2008, April 14). *Crash injury risks for obese occupants*. SAE 2008 World Congress. SAE paper no. 2008-01-0528, SAE, Warrendale, PA.
- Williams, J. S. (1970, November 17). *The nature of seat belt injuries*. 14th Stapp Car Crash Conference. SAE paper no. 700896 (pp. 44-65), SAE, Warrendale, PA.
- Williams, J. S., Lies, B. A., Jr., & Hale, H. W., Jr. (1966). The automotive safety belt: in saving a life may produce intra-abdominal injuries. *The Journal of Trauma*, 6, 303-315.
- Wonder, A. Y. (2007). *Bloodstain pattern evidence: objective approaches and case applications*. Amsterdam: Elsevier Academic.
- Woolley, R. L., & Asay, A. F. (2008, April 14). *Crash Pulse and DeltaV comparisons in a series of crash tests with similar damage (BEV, EES)*. SAE 2008 World Congress. SAE paper no. 2008-01-0168, SAE, Warrendale, PA.
- Yoganandan, N., Pintar, F. A., Skrade, D., Chmiel, W., Reinartz, J. M., & Sances, A. (1993, November 8). *Thoracic biomechanics with air bag restraint*. 37th Stapp Car Crash Conference. SAE paper no. 933121 (pp. 133-144), SAE, Warrendale, PA.
- Yoganandan, N., Pintar, F. A., Gennarelli, T. A., Maltese, M. R., & Eppinger, R. H. (2001). Mechanisms and factors involved in hip injuries during frontal crashes. *Stapp Car Crash Journal*, 45, 1-12.

Chapter 7

Ergonomics

Steven Wiker

Introduction

Ergonomics is an interdisciplinary field of engineering and natural, physical, and social sciences that seeks to understand human performance capabilities and limitations and to apply such knowledge to the design of environments, machines, equipment, tools, and tasks to enhance human performance, safety, and health. While human productivity, work quality and job satisfaction are cardinal tenants of ergonomics, etiological analysis for understanding and prevention of accidents, injuries, and deaths has been a cardinal driver for the field's development and application.

The central thesis of this chapter is that poor ergonomic design creates excessive structural or energy demands upon the body, or through degradation of perception, information processing, motor control, psychosocial, and other aspects, produces unsafe behaviors or strategies that result in accidents and injuries. The interplay among machines, environments, task designs, and human capacities is often complex, interactive, and nonlinear; making epidemiological analysis of injury response to poor designs a challenging endeavor. Remediation efforts using administrative or engineering countermeasures for injury risk require careful ergonomic analysis to determine which options are most effective and provide the greatest rate of return for the countermeasure investment.

This chapter focuses upon overexertion injuries and their relationships with human-machine interface design, tasking, working environments, human physiological, psychological, and biomechanical tolerances. The general injury epidemiological investigation process advocated here is, however, applicable with other forms of injuries and accidents that can be mediated by ergonomic design quality.

Understanding Ergonomic Design Impact

Accidents and injuries, initially attributed to human error or willful unsafe behavior, are later linked to poor or improper Task-Human-Environment-Machine (THEM) system design far more often than not. Understanding the interplay of the interfaces, within the context of overexertion injury risk, helps to point to exposure metrics that should be initially considered in the injury analysis effort; leaving data collection practicality, cost, intrusiveness, and other factors to drive or shape the final scope and nature of the model. The Venn diagram below graphically characterizes the THEM

S. Wiker, PhD, CPE (✉)
Ergonomic Design Institute, Seattle, WA, USA
e-mail: wiker@ergo.org

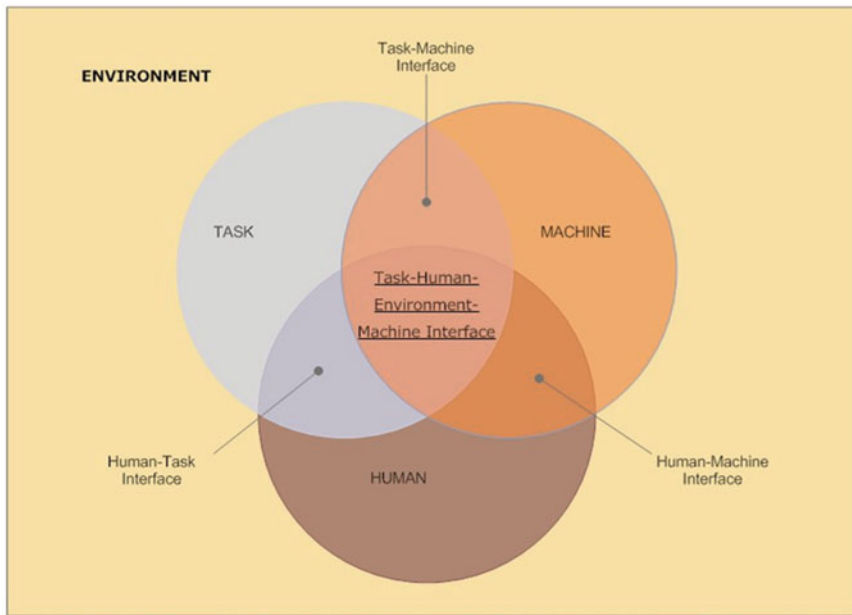


Fig. 7.1 The task–human–environment–machine interface model that should be considered when formulating injury models addressing ergonomic design questions

interfaces that should be considered when designing equipment, tasks, and work environments for human performance, safety, and health (Fig. 7.1).

To understand the THEM system, one typically follows a general process:

- (1) Understand the formal and informal objectives and goals of the system, and their impact upon allocation of activities to humans and machines. Often informal goals, which are not documented, are important drivers of hazard exposures.
- (2) Understand the range of environments that the human–machine system will be expected to operate in. Performance and operating environment requirements typically drive allocation of performance responsibilities and, thereby, determine the human’s perceptual, cognitive, motor control, biomechanical, and physiological burdens.
- (3) Perform activity, functional, decision, and action flow analyses, simulation and mockup-analysis, study of comparable systems, or focus group analysis to determine human performance demands and physical stressor exposures within the array of working environments exposures (Chapanis 1965; Niebel and Freivalds 2002).
- (4) Examine each task to determine: (a) specific perceptual, cognitive, and motor performance demands, (b) physiological workloads, biomechanical stressors, and other information needed to support other system development activities, and (c) personnel selection and training requirements are evaluated and specified. A detailed description of performance task durations, frequency, allocation, complexity, clothing and equipment used, and environmental conditions are recorded. The degree of molecularization of the analysis depends upon the nature of the question you wish to answer.
- (5) Perform failure modes and effects analyses (FMEA) or comparable studies to determine how expected and unanticipated failures impact human performance and physical exertion exposures.
- (6) Where work is more unstructured, link analysis is often used to understand the pattern of interface between humans and their work environments (Chapanis 1965; Niebel and Freivalds 2002).

A link is any useful connection between a person and a machine or part, two persons, or two parts of a machine. Evaluation of workplace layouts and tool use can be made to assess distances traversed during typical or unusual operations, crowding or disbursement of activities, lifting, carrying, pushing, pulling, and other physical interactions with humans, equipment, loads, and so forth. Often excessive or unnecessary distances for walking, carrying, pushing, or pulling are determined.

- (7) If there are voids in needed information, then design, forensic, and other forms of testing are needed to gather unknown information.
- (8) Consider study of comparable system designs and implementations to ensure that your understanding of the system is not myopic.

Standards and Design Guidelines

Design metrics in standards, testing methods, data, analyses, findings, interpretations, and rationale are often useful in creating ratio scalar, nominal, or other exposure metrics for injury modeling. If injury outcomes are associated with such metrics, then industry will be well prepared to evaluate risk based upon their degree of compliance with the standards.

One should understand that consensus guidelines or standards often provide compromised scope of application, specifications, or guidelines. They may be completely consensual, but they are more likely to be a “middle-ground” outcome that groups of members have agreed to accept because their recommendations represent an improvement over the status quo. This may leave you with a metric that allows one to gage risk, but not at the quality that one might hope.

Standards can also be conflictive in nature. Conflicts often develop when amendments to related standards are not considered at the same time due to panel schedule constraints or normal society standards committee schedules. Talking to members of committees may be helpful to determine the span of the metrics that were considered, or the bases for conflicts, before using such information to shape the span and nature of predictor variables to be used in injury modeling.

Focus Groups

Often focus groups shed light upon differences within the workforce’s and management’s perceived risk of injury, bases for injury incidence or severity, and types of countermeasures that they believe could or could not be of value in reducing risk of injury. Opinions may differ within and among focus groups. Careful and active listening will help with understanding the bases for the opinions and differences in perspectives – information that will be helpful in shaping the etiological model or future countermeasures. Focus group membership should be representative of the group at risk of injury, and small enough to encourage equal expression of opinions and insights (e.g., 5–10 members) about design issues and needs (Greenwood and Parsons 2000).

Personnel Selection and Training Factors

Previous efficacy of personnel selection and training factors should be considered when evaluating injury response. If extant personnel selection or training have no impact upon injury incidence or severity, such information is important to know when evaluating candidate etiological models and considering administrative control design options.

Design Features

Once you have a functional task analysis that bounds the interfaces among the human, machines, tasks, and environment, one should search for perceptual, information processing, motor demands, physiological demands (e.g., aerobic power demands), mechanical stresses acting upon the body, and other stressors that are mechanistically relevant to the injury model of interest. Each of these features have features that are used by ergonomists to evaluate or drive the design of machine–environment–task–human interface.

Hazard Recognition

Humans are often initially blamed for causing their injury by failing to pay heed to the “obvious” hazard or a hazard that they were trained to recognize. Subsequent analysis frequently reveals that recognition error was designed into the system and the human was unable to reliably perceive the hazard and, thereby, behave prophylactically.

Recognition of hazardous situations requires adequate sensory stimulus intensity for critical features of hazards, effective recipient sensory sensitivity and decision criteria for acknowledging the presence of stimuli and, finally, the capacity to accurately and reliably interpret and classify a stimulus ensemble as intended or expected by the designers. If the hazard cannot be reliably recognized, then hazard exposure becomes insidious. Injury assessment models should assess those perceptual or recognition issues to rule them in or out of the injury epidemiological model.

Adequate Stimulus Intensity

The psychometric relationship between the physical intensity of a stimulus and its perceived intensity follow a power function (Stevens 1957):

$$\psi = bX^P, \quad (7.1)$$

where:

- ψ = perceived magnitude of sensory perception
- b = empirically derived coefficient
- X = physical intensity of the stimulus
- P = exponent power

The coefficient b and exponent p are determined experimentally and for a number of stimuli, or features, have been cataloged for use by designers (Mowbray and Gebhard 1958; Van Cott and Warrick 1972). When the stimulus exponent is much less than unity, the human must experience large increases in stimulus intensity before the stimulus is detected, or just noticeable differences (JNDs) can be detected. If the exponent is much greater than 1, then small changes in the physical stimulus produce large changes in perceived intensity. Exponents for palm force, perceived biceps force, and heaviness are 1.1, 1.7, and 1.45, respectively.

Sensory thresholds are used by ergonomists when considering the type and magnitude of stimulus cues that are required for a given THEM system (Davis 2003; Fechner et al. 1966;

Gescheider 1976; Gescheider 1984; 1966; Stevens 1951; Stevens 1975; Yost et al. 1993).¹ The type of stimulus threshold that must be considered depends upon the nature and type of hazard(s):

- (1) Absolute threshold (AL): the smallest amount of energy that can be sensed by a normal young adult who is focused upon the source of the stimulus in the absence of other stimuli.
- (2) Recognition threshold (RL): the level at which a stimulus can be both detected and recognized.
- (3) Differential threshold (DL): the Just Noticeable Difference (JND) or Difference Limens (DLs): the amount of increase that is required in a sensed stimulus before one can detect a JND. This magnitude depends upon the existing stimulus intensity.
- (4) Terminal threshold (TL): the upper end of the detectable stimulus range results from an inability of sense organs to respond to further increases stimulus intensity.

Inspection of the power function shows that equal increments of physical energy do not produce equal increments of sensation. A stimulus DL complies approximately with its particular fraction referred to Weber's law or fraction:

$$K = \frac{\Delta I}{I}, \quad (7.2)$$

where:

- K = ratio or Weber fraction or constant
- I = current intensity of stimulation
- ΔI = change in stimulus intensity from the reference level I

As the stimulus intensity increases, greater amounts of physical stimulus intensity are required for detection. Some warning systems are designed to increase in intensity as hazard risk increases. However, if the perceived changes in stimulus intensity do not match the actual risk, then the "warning" miscues the recipient with regard to the true risk.

The Weber fraction for weight is approximately 10%. Thus, if a 100 pound load is lifted, one can expect that adding 10 pounds will result in a JND in half of the population, or in 50% of trials for an individual. An increase of load of approximately 30 pounds would be required if one wanted to 99% of the population, or 99 of 100 exertions, would result in the detection of the additional weight. Thus, when loads are heavy, a worker may not be able to reliably detect a substantial increase in the load (e.g., feed bags that are not evenly filled) and can lift excessive loads accordingly.

Not all tissues provide reliable Weber Fractions. For example, lumbar spinal disks do not provide direct sensory feedback regarding the magnitude of disk compression stress. Ancillary cues are provided such as abdominal or thoracic pressure, muscle tension, and other factors that contribute to or are correlated with disk compression. The ensemble of ancillary cues varies with postures selected or compelled, external force production, and other factors.

If the human is not intensely focused on the stimulus, or the stimulus is dynamically changing, then threshold multiples exceeding 20 may be required before hazard feature detection can be reliable. Many investigators have contributed to our understanding of the specific types and ranges of physical energy, typically referred to as stimuli, which fall within human perceptual capabilities. Mowbray, Gebhard, Van Cott, and Warrick have cataloged much of the early work (Mowbray and Gebhard 1958; Van Cott and Warrick 1972).

¹Unless otherwise specified, thresholds, limens, or JNDs are 50% detection rate thresholds and are determined based upon focused attention of young adults.

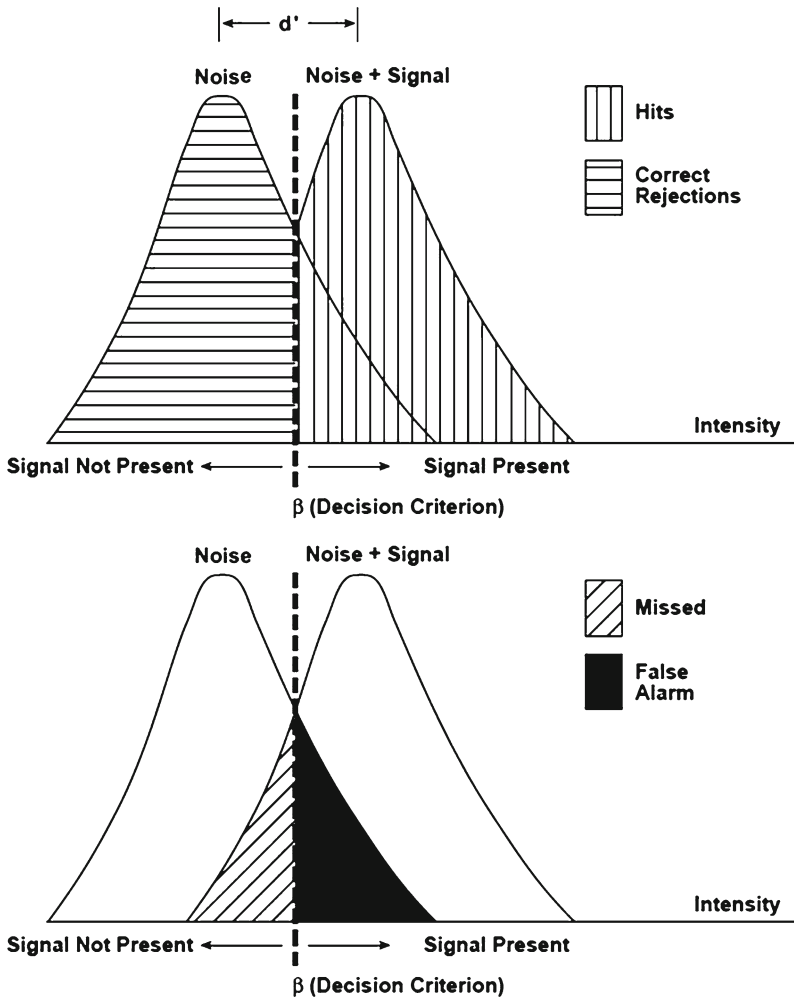


Fig. 7.2 Signal detection theory paradigm demonstrating the impact of d' and β upon hazard detection performance

Hazard Feature Detection

Signal detection theory (SDT) was developed to help designers understand why humans fail to detect suprathreshold features of stimuli (e.g., hazard characteristics) in a reliable manner (Green and Swets 1966; Green and Swets 1974; Hancock and Wintz 1966; Helstrom 1960; McNicol 1972; Poor 1994; Swets 1996; Wickens 2002).

Workers may detect the presence of hazards (Hits), their absence (Correct Rejections), fail to detect hazards (Misses) or report the presence of hazards when they are absent (False Alarms). The frequency of hits, misses, correct rejections and false alarms depends collectively upon the magnitude of a worker's perceptual signal:noise ratio, or perceptual sensitivity (d'), and at what level of stimulus physical intensity the worker requires to decide that the stimulus is present (i.e., beta (β)). The worker's sensitivity, or inherent ability to detect the hazard, can change with age, with fatigue, and other factors that degrade perception. Their betas are influenced by the frequency or probability of encountering a hazard, and the consequences of hits, misses, false alarms, and correct rejections.

If workers are not adequately trained to recognize and deal with the hazard, then their effective d' will approach zero and detection of the hazard becomes ineffective. If the effective payoff matrix manipulates the beta in a direction that is hazardous, then effective training and high d' s will be overwhelmed. Efforts to warn humans of hazards will be ineffective if their observed or predicted d' and $1/\beta_{\text{optimal}}$ are small (Fig. 7.2).

If β changes, in the face of a constant d' , material differences will result in the observer's detection of a hazard. Humans adjust their β based upon their expectation for existence of the hazard or hazard's cues (e.g., beliefs or recent phenomena that challenge their beliefs), and the values of making Hit, Correct Rejection and absolute values of costs associated with False Alarms and Misses. The product of the optimal β is, thus, adjusted by observer's payoff matrix:

$$\text{BetaOptimal} = \frac{\text{Pr(N)}}{\text{Pr(S)}} \left(\frac{\text{Value(CR)} + \text{Cost(FA)}}{\text{Value(Hit)} + \text{Cost(Miss)}} \right), \quad (7.3)$$

where:

Pr(N)	=	probability of encountering no hazard or noise
Pr(S)	=	probability of encountering a hazard or signal
Value(CR)	=	value of making a correct rejecting the presence of a hazard
Value(Hit)	=	value of detecting the hazard or signal
Cost(FA)	=	cost of responding to a hazard when in fact it is not present
Cost(Miss)	=	cost of failing to detect the presence of a hazard

For example, if a worker is assigned a manual materials handling job and they believe that the company would assign a hazardous lift, then their optimal beta would be gaged as very low (i.e., they would place their beta at very low levels and believe that most loads were hazardous):

$$\text{Beta}_{\text{Optimal}} = \frac{0.01}{0.99} = 0.01. \quad (7.4)$$

However, if the workers are rebuked in front of others or future employment loss threats are made if they do not perform the manual materials handling tasks as assigned, the value of a hit (\$1) can be far less than that of a correct rejection (\$10). The cost of a FA may be viewed as potential loss of job (salary loss of \$40,000), and the cost of a miss or injury while maintaining a job while injured could be much less (medical costs are covered by employer, \$0). Thus, the comparatively large cost of a false alarm and minimal values and cost of a miss results in a large shift in the beta to:

$$\text{Beta}_{\text{Optimal}} = 0.01 \times \left[\frac{10 + 40,000}{1 + 0} \right] = 400.1. \quad (7.5)$$

The payoff matrix shifts the beta from 0.1 (very low) to 400.1 (very high); producing a material risk of rejecting even strong hazard cues.

The observed operator's sensitivity or d' by estimating the distance between signal plus noise distribution (SN) and the lesser intense noise distribution (N) using z-scores or Receiver Operating Characteristics (ROCs). Tail probabilities for miss and false alarm rates are used to compute z-scores for distances from distribution means to the observer's beta (Fig. 7.3).

A hazard's critical feature ROC is produced by plotting the worker's hit rate against their false alarm rate for trials when the worker employs different β s. Various signal detection trials in which different expected frequencies of signals, or variations in payoff matrices, are individually or collectively used to manipulate the observer's β s. Each different combination of expected frequencies

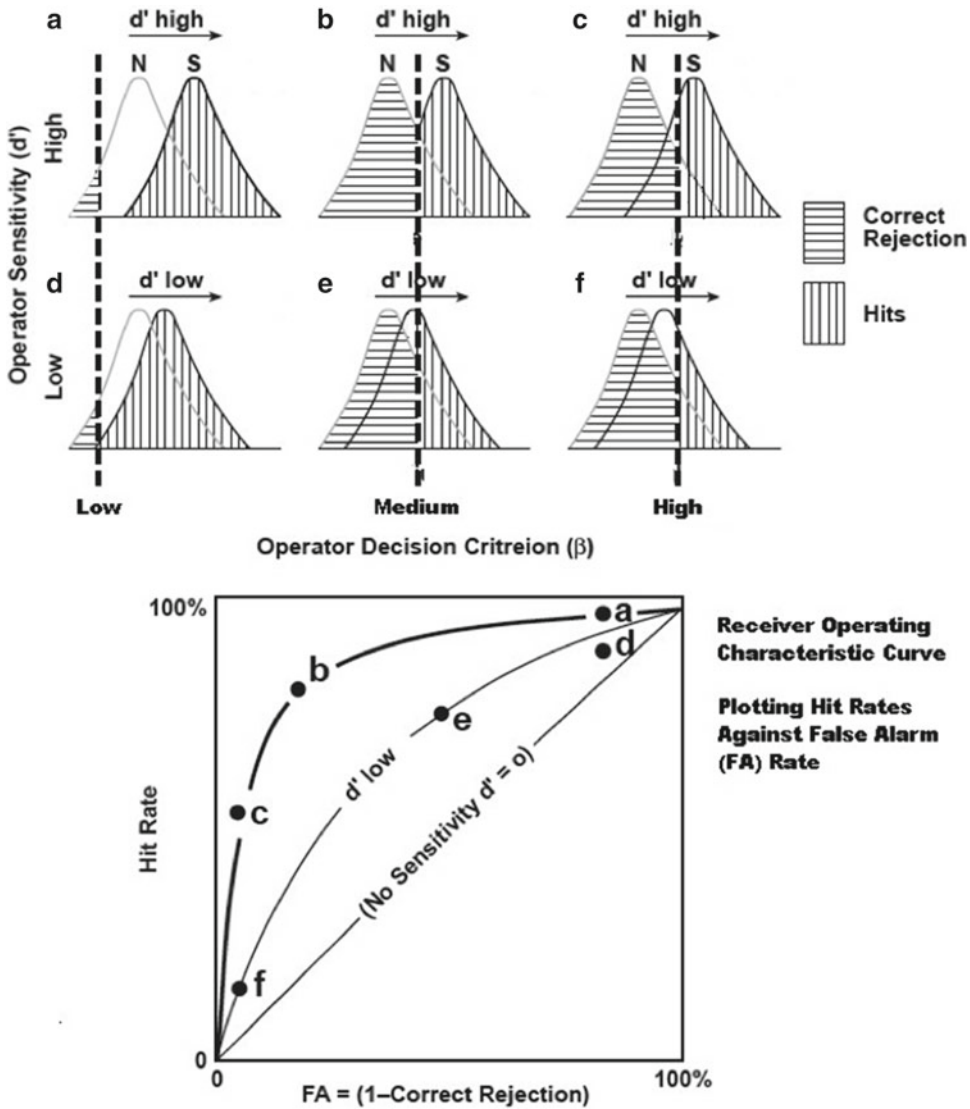


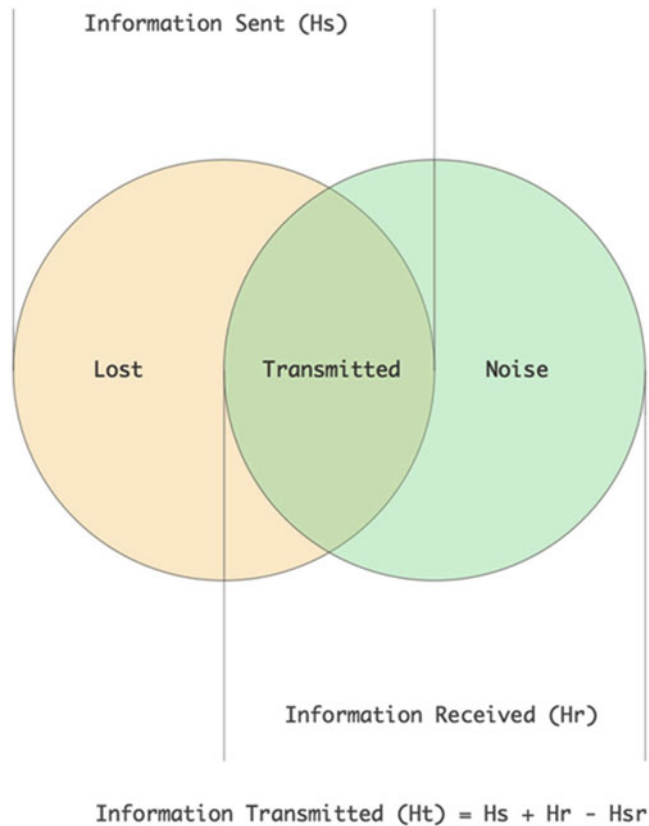
Fig. 7.3 Receiver operating characteristic (ROC) curves and their relationships to SDT operator sensitivity and response criteria

and decision payoffs produces an individual point on the ROC. The greater the area bounded between the ROC and the diagonal line (i.e., zero sensitivity or d'), the greater the signal:noise ratio and the observer's capacity to detect the signal or hazard for any given β .

The tangent to the fitted ROC line, the hit versus false alarm rate, provides an approximation of observed β s. As the tangent along the ROC increases the worker's β increases. As the area between the ROC and diagonal decreases, the observer's capacity to detect the worker's d' increases.

For example, as shown in the ROC figure, observers *a*, *b*, and *c* have greater d' or sensitivities, than *f*, *e*, and *d*; all of which do not share the same β or decision criterion. Some observers are very liberal in that they seek to increase hit rates at the expense of increased false alarms (FA) (e.g., *a* and *d*), while others are very conservative in their decision criterion (e.g., *f* and *c*) where greater intensities of signals are needed before they are willing to claim the stimulus is a signal. The conservative β s reduce hits and false alarms.

Fig. 7.4 An information theoretic model for evaluating hazard information transmittal and recognition



The injury investigator should understand that the nature and characteristics of hazards influence the effective d' and β of the person at risk. Small d' and high β s increase opportunities and risks for accidents and injuries. Some overexertion cues produce very limited d' s, corporate-workforce norms and morays concerning workload expectations and risks of overexertion injuries that materially alter individual β s.

Hazard Equivocation or Masking

If workers reliably detect a hazard's features or cues (i.e., high d' and low β s), they may still fail to recognize the presence of hazard because the cues are confusing or promote inappropriate interpretations. The information intended to be transmitted by the hazard's perceptual feature set may be equivocated or confused with other hazards or nonhazards, and extraneous perceptual input can convey additional information that produces confusion (e.g., noise).

Information theory provides a conceptual and computational framework for evaluating the risk of confusion, or noise, in creating misperceptions, misclassifications, or failures to appropriately recognize hazards. Ideally, the human receives (H_{received}) all critical information that is sent (H_{sent}) by the hazard or warning. If that occurs, then the two circles in the Venn diagram above superimpose (i.e., all information is that is sent is received); producing perfect information transmission (H_t) with no loss (H_{loss}) and no noise (H_{noise}). Hazards that present significant equivocation or noise will demonstrate very poor recognition rates. Others have provided an in-depth discussion of information theory (De Greene and Alluisi 1970) (Fig. 7.4).

To determine the amount of information that was transmitted by the intended hazard cues, one can calculate the information that was sent, received, and that represents the union of sent and received information. Information is represented in terms of bits or binary states. If we have a single event i , we can determine the information sent by that stimulus using the formula provided below with $N=1$. If we have a large number of events with each possessing a different event probability (P_i) of occurrence, we can compute the average amount of information presented to the observer by the following formula for the average sent information:

$$H_{\text{average}} = \sum_{i=1}^N P_i \log_2 \left[\frac{1}{P_i} \right]. \quad (7.6)$$

For example, a group of 1,000 supervisors, who are tasked with enforcing safe lifting practices, are presented with four images of lifting postures and asked to rate each image in terms of risk of injury. Their ratings require them to rank order the risk associated with the four postures. From the distribution of responses, we compute the information content of the hazard cues (i.e., marginal probabilities of posture cues), information content of the information received (i.e., marginal probabilities of hazard intensity judgments), and the information content within the following “confusion matrix” (Fig. 7.5).

The average information sent ($H_s=2.0$ bits) is computed from the column marginal probabilities. The information received is computed from the marginal probabilities of the rows ($H_r=1.81$ bits). Information content inside the matrix ($H_{s,r}=2.65$ bits) is subtracted from the sum of H_s and H_r to determine the information that was transmitted by the observed postures by subtraction ($H_t=H_s+H_r-H_{s,r}=1.16$ bits). The lost information ($H_s-H_t=2.0-1.16=0.84$ bits) due to confusion and errors attributed to noise ($H_r-H_t=1.81-1.16=0.65$ bits) were determined. An ideal situation is where transmission is perfect with zero equivocation and zero noise.

Information theory provides a tool to allow us to evaluate the capacity of humans to recognize the presence of hazards. Understanding the quality or capacity of hazard recognition is useful in modeling injury incidence. Moreover, studying the confusion matrix helps one to understand what phenomena are confused with the hazard. Adding or eliminating certain hazard features can reduce confusion and improve hazard recognition.

Administrative controls aimed at injury prevention, such as employee or supervisor hazard recognition and severity classification may not be as effective as anticipated. Improving hazard recognition may require changes in worker safety training content, tagging hazards with stronger and more discriminating perceptual cues, or recognition that administrative controls cannot be effective under certain conditions; requiring engineering out the hazard.

Affordances

An affordance is a form of communication that conveys purpose and operation of a perceptual cue. It can also cue behaviors that are to be avoided. Gibson described an affordance as an objective property of an object, or a feature of the immediate environment, that indicates how to interface with that object or feature (Gibson 1966). Norman refined Gibson’s definition to refer to a perceived affordance; one in which both objective characteristics of the object are combined with physical capabilities of the actor, their goals, plans, values, beliefs, and interests (Norman 1988).

Affordances are powerful design elements that can be useful if used wisely, and punishing if it motivates inappropriate or injurious behaviors. Cognitive dissonance may also develop and lead one to use of the object in an unexpected, hazardous, or injurious manner.



Postures Observed by Supervisors

Rated Risk of LBI	Posture 1	Posture 2	Posture 3	Posture 4
Nominal	1000	0	0	500
Low	0	0	0	500
Medium	0	700	800	0
High	0	300	200	0

Sums

Rated Risk of LBI	Posture 1	Posture 2	Posture 3	Posture 4	Sums
Nominal	1000	0	0	500	1500
Low	0	0	0	500	500
Medium	0	700	800	0	1500
High	0	300	200	0	500
Sums	1000	1000	1000	1000	4000

Determine Probabilities

Rated Risk of LBI	Posture 1	Posture 2	Posture 3	Posture 4	Marginals
Nominal	0.25	0	0	0.125	0.375
Low	0	0	0	0.125	0.125
Medium	0	0.175	0.2	0	0.375
High	0	0.075	0.05	0	0.125
Marginals	0.25	0.25	0.25	0.25	

Compute Information (bits)

Rated Risk of LBI	Posture 1	Posture 2	Posture 3	Posture 4	Sums
Nominal	0.50	-	-	0.38	0.53
Low	-	-	-	0.38	0.38
Medium	-	0.44	0.46	-	0.53
High	-	0.28	0.22	-	0.38
Sums	0.50	0.50	0.50	0.50	

$$H_t \text{ (Transmitted)} = H_s + H_r - H_{sr} = H_t = 1.16 \text{ bits}$$

Fig. 7.5 A confusion matrix showing the level of confusion among perceived and actual risk of spinal overexertion injury for different postural and load cue presentations to supervisors

For example, handles placed on boxes or loads can convey affordances indicating where to grasp or handle the object. The handle placements imply that the center of mass is located between the handles and that the load will be stable when lifted. However, if that is not the case, the worker is miscued and can encounter unexpected exertions, loss of balance, and impulsive forces acting upon the body.

Affordances created by designers may differ from those of the workforce due to differences in knowledgebase or experiences. Injurious behaviors may be promoted by strong affordances that the investigator must ferret out or reject through careful testing and evaluation. With slip and fall accidents, victim perceptions and selection of gait patterns are often heavily mediated by design-induced affordances (e.g., expectations of coefficients of friction, step rise:run ratios, and lack of surface discontinuities). When percepts are incorrect, trips, slips, and falls often occur (Tisserand 1985).

Cognition Errors

Even if hazards are recognized without error, poor ergonomic design can produce excessive mental workloads that challenge the worker's capacity to integrate hazard information and make preferred decisions regarding avoidance of injury. Under such conditions, workers feel time-stressed and pursue strategies that reduce mental effort. Overexertion injury risk may increase if workers feel that they have inadequate time to perform tasks using prescribed postures or methods (i.e., performing squat lifts in a slow and controlled manner). Playing catch-up with manual materials handling tasks promotes the use of higher velocities of load handling, greater use of momentum, greater metabolic burden, thermal strain, and tradeoff paradigms that trade speed against risk.

Humans process information in a "rated-limited" manner. Individually, or collectively, excessive cognitive processing span and velocity provokes filtering of critical information, slows response, and promotes flawed decision making. Information processing rates typically increase when the number of concurrent tasks performed are increased, or when information flow to a human increases, memory burdens are elevated, or when motor performance speed-accuracy trade-off requirements become excessive (Gaillard 1993; Hancock and Caird 1993; Hancock et al. 1990; Hockey and Sauer 1996; Hockey et al. 1989; Horrey and Simons 2007; Loft et al. 2007; Wierwille 1979).

Absolute workload assessments can be made when directly measuring primary and secondary tasks that burden the same resource pool. Relative resource demand assessment can be made when using comparative evaluation of indirect measures such as physiological strain. Which type of workload measurement should be used depends upon a number of factors (Wickens and Hollands 2000; Wickens et al. 2004).

In preliminary analyses, it may be useful to use timeline analysis to estimate the mental workload, or time-sharing demands, that an operator may experience. Here you simply count the number of tasks that are being performed or monitored concurrently. The sum becomes the mental workload metric. This analysis is particularly useful when evaluating changes in workloads when failures occur or when operating under unusual or stressful conditions.

Mental workload measurement has been classed into three categories: subjective (i.e., self-report) measures, performance measures, and physiological measures (O'Donnell and Eggemeier 1986). Performance metrics can be made on the primary or actual work task, secondary operational tasks, or nonoperationally relevant secondary tasks that tap the same resources that primary tasks do.

In the face of fast-paced or mentally taxing work, humans often seek to reduce mental workloads or challenges of decision making by either ignoring information, shortening their attentive period, making poor decisions too quickly, and short-cutting activities that slow performance (Craik and Salthouse 2007; Durso and Nickerson 2007; Hancock 1999; Lamberts and Goldstone 2005). Accidents that lead to injuries and deaths are often associated with high-mental workloads and excessive decision making demand.

Memory is an important tool for the detection of overexertion injury risk factors, risk assessment, recalling risk handling protocols, learning new material, and for learning from mistakes and near misses (Bower 1977; Cermak and Craik 1979; Estes 1975; Gardiner 1976; Hockey and Sauer 1996; Manzey et al. 1998; Shanks 1997; Veltman and Gaillard 1998). Memory failures are often direct or indirect causes for human performance failures and subsequent injuries. Memory may be phase classified as: (1) sensory, (2) short term, and (3) long term.

Sensory memory acts like limited buffers for sensory input. Visual iconic sensory memory is briefly present for visual stimuli (e.g., a visual “snapshot” that fades very quickly). Aural stimuli produce *echoic sensory memory* that requires silent rehearsal. Other sensory modes have rapid decay of sensory information unless the stimulus is reinforced by continuous visual, aural, haptic, olfactory, or gustatory stimulation.

Stimuli captured by sensory memory must move rapidly into short-term memory through attention. If the stimuli are not attended, the sensory information is effectively filtered. Sensory memory is very susceptible to masking disturbance (i.e., extraneous stimuli that compete more effectively for attention than the stimuli of interest).

Short-term memory also decays rapidly if not sustained by continuous stimulus or rehearsal (e.g., continuously looking at a visual image, or rehearsing the phone number while waiting to dial) and is the locus for coupling incoming information from long-term store. The short-term store process is often referred to as the “work-bench” where low- and high-level associations are developed and sensory patterns are imbued with characteristics that were never sensed. This process is also a component required for the development of new associations and creation of augmented long-term store.

Long-term storage has been classified as episodic memory (i.e., storage of events and experiences in a serial form) or as semantic memory (i.e., organization record of associations, declarative information, mental models or concepts, and acquired motor skills). Information from short-term memory is stored in long-term memory if rehearsal is adequate, and if associative structures or hooks are available (i.e., some prerequisite information is available in long-term store). For example, rehearsal of an equation is of little value if one does not have any knowledge of the underlying phenomena linked to the equation.

Poorly designed injury prevention training programs fail to produce adequate long-term store of essential information related to the recognition of overexertion risk and selection of appropriate response behaviors. Poor training programs are characterized by excessive information flow, failure to allow adequate and distributed rehearsal of information, or are designed to capitalize upon prerequisite long-term store or associative structures that are absent. Administrative controls associated with training are ineffective if the training program is not designed properly for the intended population. Learning is most effective if elaborative rehearsal is distributed across time. Frequent and distributed training is more effective than providing training only at the hiring stage.

Companies often rely too heavily upon learning and recall on the part of the worker to prevent errors in sequences of operations, to support choice, diagnostic or predictive decisions associated with exertion work behaviors. Injury or accident investigators may also expect too much from injured workers when asking them to recall events leading to, or occurring during or after an accident. Marked differences in investigator and injured worker semantics during discourse or questioning, or the brevity of the accident or injury event and injury process, can produce material differences in the capacity to accurately recall and record events for subsequent injury analysis.

Poor design is typically characterized by substantial recall demands without memory aids such as checklists, increased display times, electronic to do lists, attention cues, and other tools that promote accurate recall and sequencing of information (Hancock 1987; Manzey et al. 1998; Wise et al. 2010). Injury investigators should seek objective corroborators of human recall wherever possible. Black box recorders are used in nearly all vehicles where accidents can be either frequent or public disasters. Those instruments present control, display, and operator behavior information prior to and during accidents that is more reliable and objective than human memory.

The injured may attempt to fill the recall voids with “puzzle pieces” until an accident or injury scenario develops which they have shaped based upon the capabilities or limitations of their associative memory. This outcome leads to reporting of “facts” that fit their theory, and rejection of facts that do not. This behavior is not malevolent; it is simply the result of an honest attempt to try to understand what happened and acceptance of “facts” that may be provided by coworkers or others who have expressed theories about the injury etiology. The sooner the investigator queries the

injured, and stresses the benefit of reporting only immediately available facts, the less bias one will encounter in the injury investigation process. In any event, objective corroboration is important when dealing with human recall of an injury or illness.

Information processing and memory demands always influence decision-making quality. A decision occurs when one must choose from among options, predict or forecast outcomes, or when one must perform diagnostics. Injuries are often associated with inappropriate decisions. Humans are not purely objective and rational decision makers. Past experience and previously successful heuristics, or rules of thumb, supplant computer-like evaluation, and selection processes (Booher and Knovel 2003; Stokes et al. 1990; Wickens and Hollands 2000; Wickens et al. 2004).

Cognitive burdens imposed by decision making are created with excessive recall and maintenance of a set of attributes and their values in working memory. A complex decision is similar to attempting to mentally solve an algebraic equation that possesses many terms and coefficients. The greater the number of terms and coefficients, the more data have to be recalled, inserted into the terms, multiplied by their coefficients, and serially aggregated to arrive at a solution. The greater the burden, the more likely errors will be made. Choice decisions typically are easier to make than prediction decisions because predictions often require additional mental algebra.

Diagnostic decisions typically produce the greatest burden because the individual has to start with a large number of potential choices. Information gathered is then used to back-chain from a current state to an array of possible etiologies. In the early stages of diagnosis, there may be hundreds of potential etiologies to contend with. Further data gathering is required until the solution space can be adequately narrowed. Even when adequately narrowed, the potential solution space may be very large and can exceed human capacity to handle without high risk of error.

Often decision makers are not given adequate time to obtain all facts. Facts do not sequence in to the decision make in an ideal rate, order, or manner. Decision makers may have to arrive at conclusions or make decisions without all of the facts. An incomplete set of facts can produce inappropriate hypotheses or perceptions. Sequencing effects can also produce inappropriate differentials in the values or weights applied to such information.

Decision makers use experience to select as few hypotheses as possible to evaluate the problem at hand. Initial hypotheses, once selected, serve as filters for subsequent information that is not germane to those hypotheses (i.e., if the information is not relevant to the hypothesis entertained, it is rejected). This stratagem reduces mental workload but may do so at the expense of decision error.

Decision errors associated with human accidents and injuries result when one or more of the following behaviors occur (Wickens and Hollands 2000):

- (1) When in doubt, correlate. Causality by correlation is a common but fallacious approach used to understand unusual phenomena.
- (2) We tend to develop “cognitive tunnel vision” and resist attending information that contradicts our beliefs.
- (3) Rules of thumb or heuristics are used to avoid mental effort and expedite decision making.
- (4) Mental statistical assessment of data is intuitive rather than objective; leading to errors in assigning weights to attributes. Examples are:
 - (a) We linearize curvilinear relationships and, thus, over or underestimate future behaviors of systems.
 - (b) We overestimate the range or magnitude of variability for larger means of sampled data.
 - (c) Modes indicate data means (e.g., higher counts of a number suggest the average).
 - (d) We do not condition probabilities based upon accepting new and relevant information; producing errors in expectation.
- (5) We bias our decisions to choose conservative decision outcomes. We increase our SDT β s, regress toward the “mean,” and avoid thinking “out of the box.”

- (6) First impressions (primacy bias) can take hold and bias all subsequent information gathering and weighting. Or, a recent material negative experience can promote negation of prior data or experiences (recency bias).
- (7) Divide and conquer in the face of overwhelming data and choices. Throwing too much information at a human often leads to filtering on their part. They seek a small set of hypotheses (<3 or 4), and then attend information that principally supports one or more of their initial guesses. Ease of recall of an initially feasible hypothesis may be used to filter additional information, or the worker can rely upon heuristics, primacy bias, and other behaviors to control their mental workload in decision making.
- (8) Overreliance upon topic experts, computers, and other sources of apparently reliable information when they are given inaccurate information or they use an inappropriate heuristic to quickly address a question.
- (9) Over confidence in one's ability to make decisions. Past success can breed unjustified feelings of consistent success.
- (10) Negative consequences outweigh positive consequences. If all benefits and costs result in a zero-sum gain, the decision maker is likely to select outcomes that are risk or cost averse.

One should also consider whether unrelated decision-making errors have inadvertently resulted in shorter time periods to accomplish tasks, poor spatial placement of equipment, material, or other types of loads to be moved or stored. Just as decision stress shapes mental workloads and drives "cost-cutting" measures to reduce mental workloads at the expense of decision quality, it imposes tradeoff decisions regarding manual materials handling and work scheduling. If the worker inadvertently paints themselves into a corner with their previous decisions and plans, they may be forced to make a decision to undo the mistake and muscle through the error "just this time."

Motor Performance Demands

Human motor performance is governed by rate-limited information processing which induces a speed-accuracy trade-off model that was developed by Fitts and colleagues from the 1950s (Fitts 1954; Fitts and Deininger 1954; Fitts and Peterson 1964; Fitts and Radford 1966; Fitts and Seeger 1953; Welford 1968). Fitts and colleagues demonstrated a reliable and powerful relationship between movement difficulty and movement indexes of difficulty (Fitts et al. 1956; Fitts 1958; Fitts and Peterson 1964):

$$MT = a + b \log_2 \left(\frac{2A}{W} \right), \quad (7.7)$$

where:

- MT = movement time
- a, b = regression model coefficients
- A = amplitude or move distance
- W = move endpoint accuracy requirement or target width
- ID = index of Difficulty = $\log_2 (2A/W)$ units are in bits

There are different degrees of molecularization of this model (Welford 1968; Wiker et al. 1989a, b). However, load-handling speed-accuracy tradeoffs, and mental and physical workloads increase, as the index of difficulty for manual transfer of loads increases. As the index of difficulty increases, precision load handling increases – leading to extended placements, less reliance upon load momentum, and increased duration of mechanical strain.

Overexertion Injuries

Overexertion injuries result from exposures to excessive whole-body force production, excessive physical work, and failure to adequately shed metabolic heat in challenging thermal environs. In the following sections, a brief discussion of the prevalence and import of each form of overexertion risk, and methods that have historically been used to characterize both stress and strain.

Force, Work, and Power

The impact of force depends collectively upon the characteristics of the force exposure, the groups of muscles and joints affected, the orientation of the body's segments and articulations as it generates force, the frequency and duration of exertions and, finally, what type of thermal environment has enveloped the worker producing such forces. Force is the produce to acceleration and mass. Work is the product of force and distance. Power is the timed aggregation of work.

Some studies have relied upon weight of objects manually handled as an indicator variable for force and biomechanical stresses. Under highly constrained circumstances, such a metric is rational. However, in most cases, one needs to fully understand the nature of forces acting upon the body. Characteristics that merit attention from an injury study team include:

- The *magnitude of the force* impacts focal pressure and rotational forces about a joint.
- The *line of action* of the force. If the line of action of the force is a projection of the force vector and it determines the magnitude of the “lever” or moment arm that is created when the force can create rotation about a point of rotation (e.g., joint). The perpendicular distance of the line of action from a joint determines the rotational forces about that joint for any given magnitude of force.
- The *sense or direction* of the force determines the direction of its effect (e.g., direction of joint rotation).
- The *point of application* also influences the magnitude of load moments by moving the line of action of the force.
- The distribution of force over an area of the body determines surface or contact pressure.
- Pressure can reduce or halt perfusion, oxygenation, and removal of metabolic waste products from tissues (Fig. 7.6).

As shown in Fig. 7.6, the moments or rotational forces acting upon the elbow are vastly different even though the forces (1, 2, and 3) are of comparable magnitudes. The effective lever arms L1, L2, and L3 are determined by the perpendicular projection to the line of action of each of the forces. L2 is of zero length because the force (2) has a line of action that projects through the point of rotation. The greater lever arm magnitude, L1, produces a greater rotational force about the elbow when compared with the shorter lever arm L3. The cross product of the vectors $L1 \times 1$ produces a clockwise rotational force of greater magnitude than the counterclockwise moment $L3 \times 3$. Thus, it is critical to know the direction and line of action of forces in manual exertion tasks, not just the magnitude of the force or load, if one wishes to adequately characterize the biomechanical rotational stress at a particular articulation.

Nonstatic Forces

The greater the acceleration acting upon a mass, the greater the force. Any rapid change in the velocity of execution will result in much higher force levels during the exertion (e.g., jerks, collisions, etc.). Hence, the tenet of exert in a slow and controlled manner without jerking.

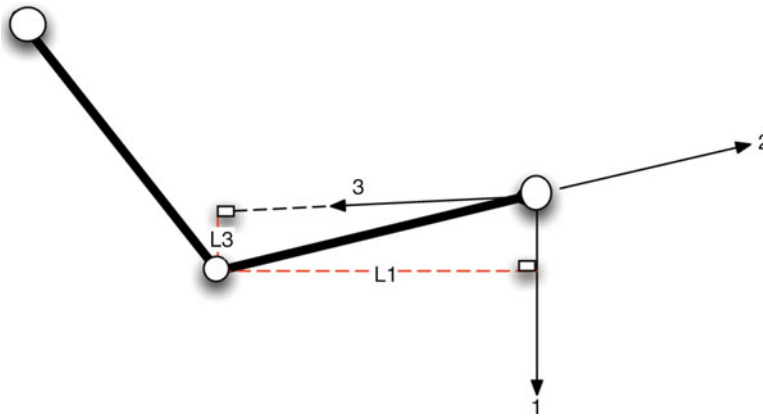


Fig. 7.6 Diagram of upper and lower arm with three forces acting upon hand in three different directions. Lever arms L1 and L3 are shown, L2 has no magnitude because the line of action passes through the elbow joint or point of rotation

Impulsive forces are created by a change in momentum and the time duration in which the change occurs as shown in the equation below:

$$\text{Force} = \frac{m_f v_f - m_i v_i}{t_f - t_i}, \quad (7.8)$$

where:

- Force = average magnitude of impulse force over time interval Δt
- $m_{i,f}$ = initial or final mass of load
- $v_{i,f}$ = initial or final velocity of load
- $t_{i,f}$ = initial or final times, or Δt

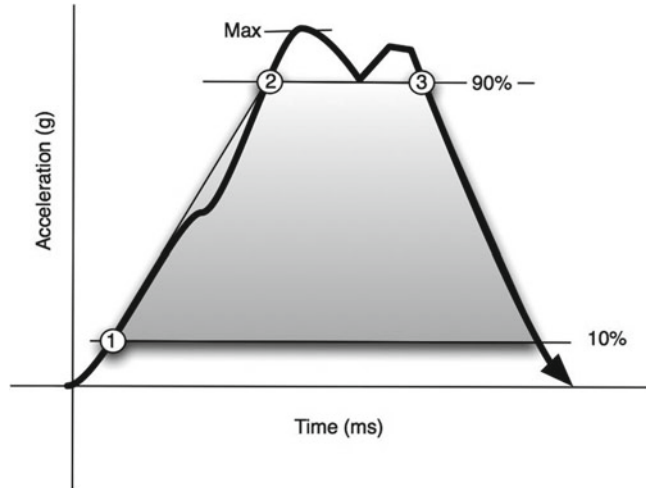
The more abrupt the change in momentum (mass \times velocity), the greater the impulsive force exposure. An impulse trapezoid is typically used to describe a measured acceleration impulse. Given that mass is constant, the magnitude of force experience Q is determined by the acceleration S . The acceleration is recorded across time. The maximum value is determined, and 90% of the maximum serves as the peak acceleration metric. Ten percent of the maximum serves as the base of the trapezoid. The slope from the 10 to 90% intersections determines the onset rate and the intersection points (Fig. 7.7).

For example, a seated human impulse received at the base of the spine with a line of action through the axis of the human spine would be expected to produce hazardous static load exposures for a 100 kg human at about 13 g. Standards for pilot ejection seat impulse exposure, based upon human subject testing during WW II, are set at 18 g peak acceleration, limited to 350 g/s onset rate, and peak duration of not more than 100 ms. If peak durations are much less than 100 ms, then impulse peaks can rise toward 27 g without injury. A historical bibliography of in vivo human testing and analytical development of the standards is available elsewhere (Wiker and Miller 1983).

Segment Rotations Create Additional Forces

If the body is moving at levels that exceed near static (e.g., slow and controlled movements), then dynamic biomechanical computations are required. The impact of material velocity and accelerations upon the load moment of a single body segment experiencing rotation (e.g., the forearm and hand link) is demonstrated below:

Fig. 7.7 Trapezoid overlay that is used to characterize a peak acceleration (point 2), onset rate (slope of 1–2), and duration (time from 2 to 3) of an acceleration duration at peak



$$M = [\text{mass} \times g \times \cos(\theta)] + [\text{mass} \times r^2 \times \ddot{\theta}] + [I_{\text{cm}} \times \ddot{\theta}], \quad (7.9)$$

$$F = [\text{mass} \times g] + [\text{mass} \times r \times \dot{\theta}^2] + [\text{mass} \times r \times \ddot{\theta}], \quad (7.10)$$

where:

M	=	dynamic moment
F	=	dynamic resultant force
mass	=	mass of segment
r	=	radial distance from joint to center of mass
g	=	gravitational constant
Velocity ($\dot{\theta}$)	=	angular velocity (rad/s)
Acceleration ($\ddot{\theta}$)	=	angular acceleration (rad/s/s)
Tangential Force	=	mass r acceleration
Centripetal Force	=	mass r velocity ²

The following plot shows that if accelerations and velocities of a single body segment (e.g., the forearm and hand segment) are low, then the difference between the dynamic and static moment (i.e., moment ratio is low). However, if the limb is rotating at high velocity and jerks, then dynamic moments can exceed static moments by a factor of three or more (Fig. 7.8).

Biomechanical models sum moments and resultant forces from the hands to the lumbar spine and then to the feet. The errors estimating joint moments and resultant forces are exacerbated if static moments are used in lieu of dynamic moments. If all body segments are in motion, then additional forces are encountered due to the differences between rotational accelerations of the two links. Under such circumstances, one must account for additional Coriolis forces when computing a joint's dynamic load moment and resultant force (Plagenhoef et al. 1971).

Exertion Stress: Strain Models

Injuries associated with excessive mechanical strain to musculoskeletal tissues are generally focal to the torso and extremities with the manual materials handling sector of industry. More than one million workers suffer back injuries each year; back injuries account for one of every five

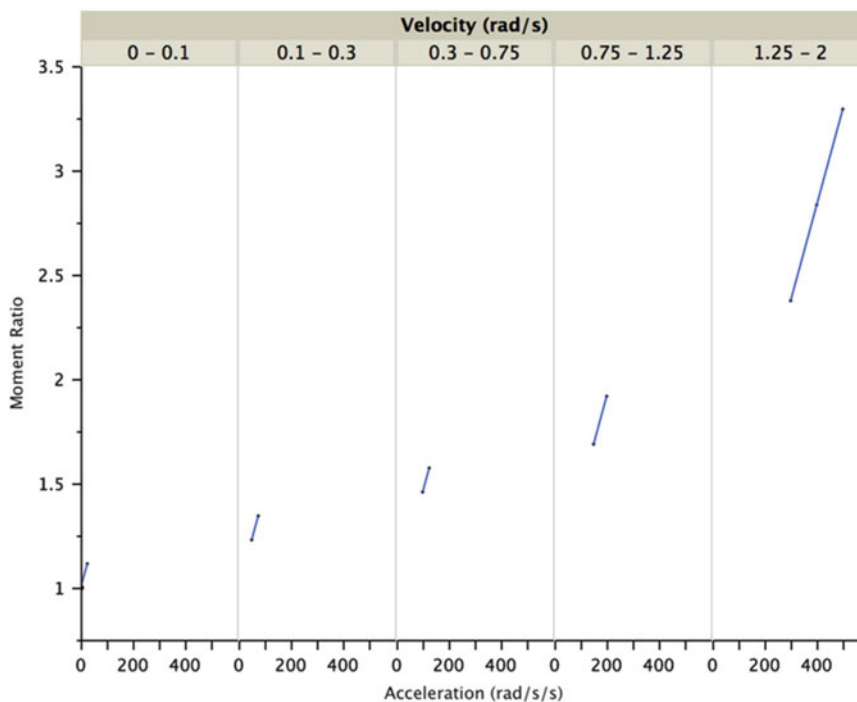


Fig. 7.8 Ratio of dynamic to static moment for a single segment (e.g., forearm) at various angular velocities and accelerations

workplace injuries or illnesses; and they account for more than 25% of all indemnity costs. Musculoskeletal overexertion injuries are broadly experienced across industrial sectors (Fig. 7.9).

Infrequent excessive exertions can strain the musculoskeletal system without taxing the cardiopulmonary system. Biomechanically safe exertions, if sufficient in frequency and duration, can prove excessive from the perspective of metabolic energy demand. Excessive metabolic demand produces excessive internal heat production with exergonic catabolism. Training effects reduce some impact of the stress. However, without proper rest, the chance of circulation problems and stroke increase. Cardiac arrhythmias can also develop with episodes of extreme physical workloads (Morris et al. 1953; Paffenbarger et al. 1970).

Unaccustomed or overexertion has demonstrated serious performance loss and muscle damage in young male populations, and materially increases the need for careful planning of recovery bouts (Bahr et al. 2003; Baumert et al. 2006; Budgett 1990; Kibler et al. 1992; Kuipers 1998; Purvis et al. 2010; Simpson and Howard 2009; Stone 1990; Teeple et al. 2006; Vetter and Symonds 2010; Wilber et al. 1995). While much of the research has been conducted upon young athletes and military personnel, there are industrial jobs that require sufficient effort to be considered comparable to athletic effort – particularly in older and less fit workforces.

If metabolic heat gain cannot be dissipated sufficiently to maintain thermal homeostasis, heat-related injuries and deaths can occur. During a 15-year period ending in 2006, a total of 423 worker deaths from exposure to environmental heat were reported in the USA (0.02 deaths per 100,000 workers); 102 (24%) occurred in workers employed in the agriculture, forestry, fishing, and hunting industries (rate: 0.16 per 100,000 workers). Of the 102 cases, 68 (67%) occurred in workers employed in the crop production or support activities for crop production sectors; resulting in an average annual fatality rate of 0.39 deaths per 100,000. Nearly all deceased crop workers were males ranging

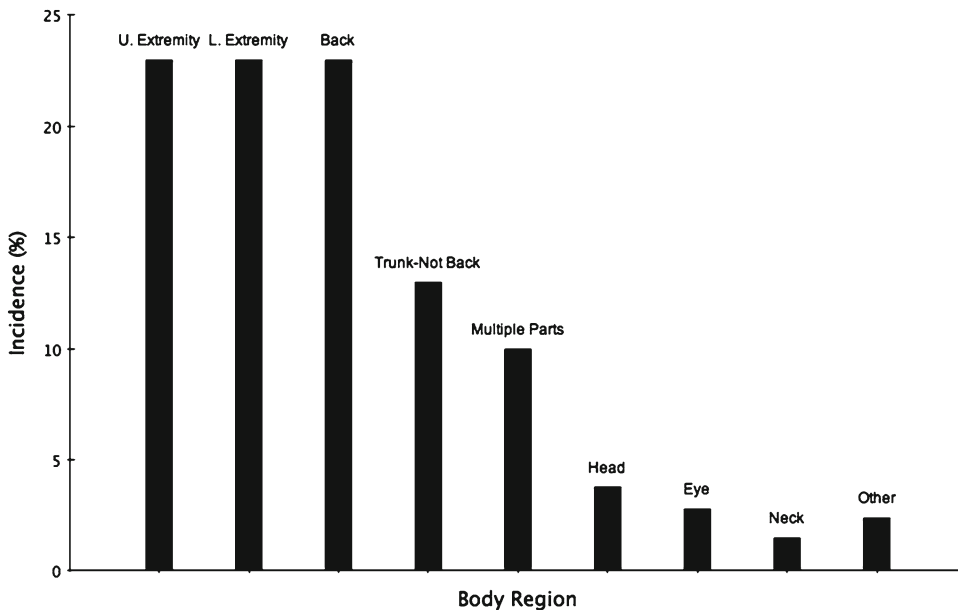


Fig. 7.9 Distribution of injuries across body parts (BLS 2005)

in age from 20 to 54 years, with 60% of the deaths occurring in the afternoons during July (Luginbuhl et al. 2008).

Biomechanical Models

A variety of biomechanical modeling tools have since developed for evaluation of various types of manual exertion tasks, dynamic activities, and composite lifting hazard assessment tools that combine biomechanical, work physiology, psychophysical strain, and injury epidemiology criteria (Chaffin 1969; Freivalds et al. 1984; Gallagher et al. 1994; Granata et al. 1997; Lavender et al. 1999; Lee and Granata 2006; Lee and Chen 2000; Marras et al. 1999; Marras et al. 1995; Mirka and Marras 1993). See Chaffin et al. (2006) for a historical review.

All biomechanical models use a set of linked rods or links to represent the bony architecture. Links are typically sized using ratios of stature (Contini et al. 1963). Each segment's mass, and its center of mass, are scaled and located, respectively, ratios of the total body mass and link lengths, respectively, following the work of Dempster and colleagues (Dempster and Gaughran 1967).

For dynamic models, population segment moments of inertia, or resistance to rotation created by the distribution of mass from the point of rotation, are added based upon in vivo and cadaver studies of body segments. Rotation of body segments requires overcoming the moments of inertia created by the distribution of segment mass distal to the point of rotation, as well as Coriolis forces associated with multi-articulated links moving at different angular velocities and accelerations.

Origins and insertions are established for muscle groups and, thereby, determining lines of action for internal forces for given joint postures. Models typically lump muscles together into groups, and other operational constraints are imposed, to constrain the number of variables to avoid indeterminate solutions.

Once the architecture and postures are defined, hand force vectors are determined and applied. Models successively compute joint load moments and resultant forces, pressures, and other phenomena at points of interest (e.g., L5/S1 disk compression, shoulder flexion strength demands, etc.).

Joint moments and resultant forces aggregate as one successively computes joint moments and forces from the wrist to locations of interest (e.g., lumbar spine). For static biomechanical models, load moments are equal to the moment at the distal end of the link plus the moment created by the weight of the link and moment created by the aggregated resultant forces acting at the distal end of the link. Body postures can change the direction of moments and, thereby, reduce or increase the magnitudes of moments as one moves through the body.

Resultant forces are successively aggregated from the hands and aggregated with body link weights without regard to link orientation. Resultant forces are passed through the body linkage system to the feet, or ground, to determine the ratio of horizontal:vertical foot forces, or required coefficients for friction.

Some biomechanical models compare articulation load moments imposed by external forces and postures against strength moments produced by muscular forces. Isometric strength moments generating functions are often used to evaluate whether humans have sufficient strength to resist or tolerate load moments. As load moments approach strength limits, risk of joint injury increases unless postures are modified in an effort to modulate strength demands (NIOSH 1981).

If activities produce body segment accelerations that exceed quasi-static behaviors, then dynamic biomechanical models should be considered. Dynamic models require one to record link orientations, as well as their translational and rotational velocities and accelerations, to produce estimates of dynamic joint moments and resultant forces. Translational and rotational velocities and accelerations are usually assessed by comparing “snapshot” samples or recordings of link spatial orientations and positions and using differencing algorithms to determine the magnitude of change per unit time sample.

Recording changes in the Cartesian loci (e.g., x , y , and z coordinates) for all computational points of interest (e.g., joint centers, centers of mass, etc.) determines the position and velocity of translational and angular movements of the body. Positional information is usually filtered to smooth out marker or sensor “jitter” or other measurement error in positional time histories. Biomechanical models typically smooth input data rather than intermediate or output data; however, some models smooth all. One should take care to determine that filtering is adequate but not excessive for the task(s) to be studied (Winter 2009).

Velocity computations are averages over the sampling duration. Shorter sampling epochs increase accuracy of postural time histories and subsequent difference equation estimates of velocity. Linear and angular accelerations are obtained by using the same differencing process for velocities; however, successive velocity estimates are used. Jerks are determined using equivalent differencing computations using successive acceleration estimates.

Hand force measurements, or accurate predictions of such, are requisite for biomechanical computations of load moments and resultant forces acting throughout the body. For static biomechanical models, force gages can be used to measure loads or hand forces along proper lines of action. Making such measurements is logistically acceptable on a small scale. However, large-scale studies should consider mechanisms that are used for dynamic models.

For dynamic biomechanical models, hand force measurements are obtained by instrumenting handles or objects, use of force plates, and other approaches to determine the lines of action and magnitudes of forces in three dimensions. While these approaches are acceptable on a small scale, they become more feasible when large numbers of different postures must be addressed and instrumentation costs shrink per measurement. For such jobs, real-time synchronized measurement of body kinematics, kinetics and hand forces, and moments becomes increasingly necessary for field research.

For static or quasi-static biomechanical models, one can record linear and angular position of the body using photographs, frame-by-frame analysis of video recordings, marker, or gyroscopic and accelerometer sensor-based kinematic recording systems. For quasi-static exertions, experienced ergonomists will typically use the most biomechanically stressful combinations of hand forces and body postures to perform biomechanical assessments (i.e., “worst-case scenarios”).

This strategy is often used with quasi-static biomechanical analyses because: (a) it materially limits the recording and computational efforts, (b) workers are trained to use slow and controlled exertions to perform manual materials handling tasks in a quasi-static manner, and (c) human strength is maximized when exertions are approximate isometric behaviors (e.g., slow, controlled, and near static exertions allow greatest force production for heavy exertions).

Static whole-body biomechanical models have been used predominantly in the past because of ease of data collection, reasonable *in vitro* intradiscal pressure correlations with predicted pressures, availability of static strength models, and cadaver spinal segment tolerance limits [see Chaffin et al. (2006) for discussions and tolerance data]. Historically, quasi-static biomechanical analyses have been applied to the “worst case scenarios” to reduce data gathering and computational demands. Experienced biomechanists can determine which epochs of hand force and body posture produce the most stressful exposures and they analyze those exertions to find maximum stress.

Dynamic whole-body biomechanical models have historically been relegated to laboratory environments due to the magnitude and nature of instrumentation required. However, goniometric recording systems, instrumented hand couples, and other field-tolerant instrumentation have made use of dynamic biomechanical models increasingly feasible. There are “camps” regarding whether one should use static or dynamic whole-body biomechanical models. In truth, both have their place. The investigators have to choose or develop the best model for their particular study.

Quasi-static exertion is a term of art that refers to physical exertions that are performed in sufficiently slow and controlled manners that accelerations are close to unity; permitting one to approximate biomechanical stresses using static equilibrium methods with weights or forces that are measured statically using a force gage or like-instruments.

Static models have been advocated because: (a) tissue tolerance to static compressive or tensile forces are easier to measure using standard materials testing methods, (b) human muscular strength is greatest when the muscle’s velocity of shortening approximates zero, (c) *in vitro* static exertion tests have produced reasonable degrees of validation of static biomechanical model predictions, (d) some studies have demonstrated reasonable relationships between static force predictions and injury incidence, and (e) measurement of postures and measurement of hand forces were comparatively easily.

Dynamic biomechanical models are most appropriate when exertions are clearly dynamic, acceleration magnitudes increase well above unity and the rates of change of acceleration are material. Under such conditions, dynamic models provide better internal force exposure assessment. There are costs associated with the greater biomechanical stress assessment fidelity: (a) tissue responses to highly dynamic forces becomes increasingly nonlinear and predicting tissue response becomes more complicated, (b) detailed postural kinematics are required to get linear and angular acceleration inputs needed by the model, and (c) time and costs associated with data collection and reduction can climb exponentially.

If sufficient, tissue compression occludes perfusion, oxygenation, and physiological homeostasis of the underlying tissue, then tissue anoxia and necrosis can develop (e.g., pressure wounds or sores), collapse of the tissue’s cellular architecture (e.g., lumbar disk herniation, endplate fractures, etc.) and/or pain. Excessive forces can cause tissues to yield or part, vascular damage and hemorrhage, a loss of architecture (e.g., muscle, tendon, or ligament tears), and pain. Under such circumstances, we are concerned with tissue “strength” or resistance to plastic behaviors (e.g., yields and fractures). See the following Fig. 7.10.

Body tissues display linear elastic behavior, defined by a linear stress-strain relationship at the lower levels of deformation. Within that range, tissues will return to their resting length following a deformation. However, beyond the linear stress-strain region, deformations become plastic in nature and will remain deformed to some extent.

For lumbar disk compression, NIOSH has recommended Action limits at 3,400 N at which place Administrative Controls must be in place. A Maximum Permissible Limit of 6,400 N has been set

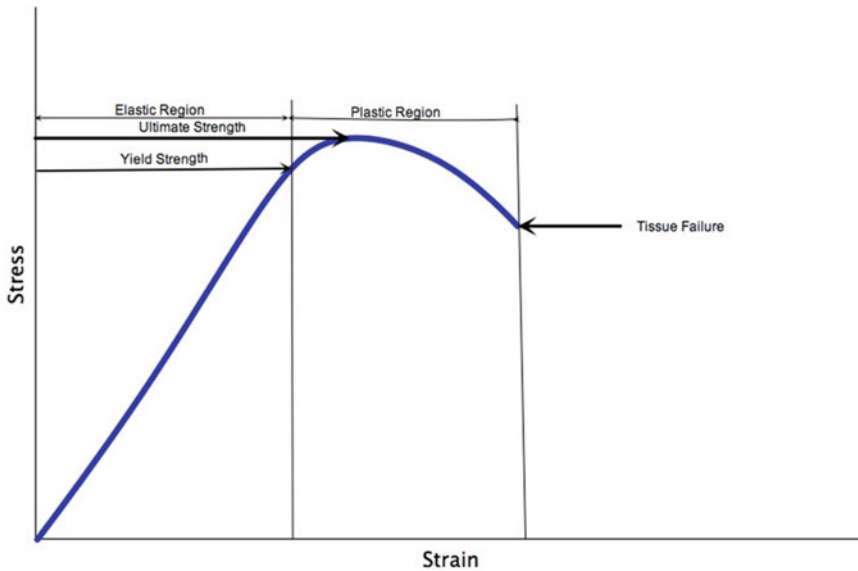


Fig. 7.10 Plot of tissue strain response to increasing stress

where engineering controls are the only acceptable mode of prophylaxis from plastic deformations of paravertebral tissues.

Hernias

Hernias in the abdominal wall, umbilical and inguinal areas are associated with heavy lifting. High Intra-Abdominal Pressures (IAPs) have been measured when performing heavy lifting have been associated with abdominal hernias. Torso posture and hip moment magnitudes have demonstrated material association with IAP (See Chaffin et al. 2006, for historical development of IAP metrics):

$$\text{IAP} = 10^{-4} \times (44 - 0.36A) \times \text{HipMoment}^{1.8}, \quad (7.11)$$

where:

- IAP = Intra-abdominal pressure (mm Hg)
- A = Relative angle (deg) between Torso and Thigh
- Hip moment = Combined load moment at the hips (Nm)

When performing lifts or exertions, abdominal wall musculature and intrathoracic pressures increase making the diaphragm stiffer. Collectively these actions serve to increase IAP. Fisher (1967) found IAP increased as the moment at the hips increased and that relative posture between the torso and thigh influenced IAP as well. As one flexes the torso, the abdominal wall musculature stretch is reduced making the wall more flaccid. As one stands increasingly erect, the abdominal wall musculature becomes more taut and, thereby, more efficient at producing tension. The magnitude of lifting stress imparted to the torso is highly correlated with the moment at the hip. As lifting stresses increase, lifters promote greater intra-thoracic and intra-abdominal pressure to stiffen the torso. Increased stiffness may act to reduce spinal compression stress and only to a comparatively small degree.

A relatively low limit of 90 mm of mercury for the value of abdominal pressure has been suggested to control risk of lifting-induced injuries to the torso (Davis et al. 1977). Values of 150 mm of mercury and higher have been reported in those who regularly lift weights (Kingma et al. 2006).

Risk of herniation of the abdominal wall is dependent upon an array of risk factors – including heavy lifting (Cobb et al. 2005; Davis et al. 1977; Dennis et al. 1975; Hemborg et al. 1983; Madden 1989; Mairiaux et al. 1984; Seidler et al. 2003; Smith et al. 1999; Stubbs 1981; Stubbs 1985; Veres et al. 2010; Wagner 2011). Peak IAPs are approximately 20% greater when compared against the sustained values (Marras and Mirka 1990). Pressures rise further if Valsalva maneuvers (breath holding during lifts) are performed (Goldish et al. 1994). Impulse loads, or unexpected changes in hand forces that can occur with team lifting, can also materially increase peak and sustained IAPs.

The relationship between IAP and disk compression has been found to range between $r=0.73$ and $r>0.92$ (Chaffin 1969). If one controls lumbar disk compression risk, then one is effectively controlling risk of herniation due to very excessive IAPs. Thus, NIOSH has handled the IAP risk for the population by setting mechanical exposure limits for the lumbar spine's disks.

Confounders

Some musculoskeletal injuries that result from excessive physical exertions result from exposure to external forces that result from falls or near falls, whole-body impacts, vibrations, or overuse syndrome. Depending upon the goal and scope of the injury model of interest, one must consider the following hazards as contributors or, when not included in the scope of the study, as confounders.

Falls

Falls are one of the top three causes of accidental deaths in the USA (Englander et al. 1996). See the following table for classifications of falls reported in 2007 (Table 7.1).

Falls produce sprains, strains, and connective tissue tears. Reflexive muscular contractions during fall recovery efforts produce sufficiently violent muscular contractions that musculoskeletal injuries may result in the torso or spine. Thus, one should address whether slips, trips, or falls have contributed to the population of overexertion injuries under analysis.

Slips occur when available frictional force is insufficient to resist the foot's shear force when walking or when pushing or pulling objects – resulting in a slip. Trips result when gait surfaces unexpectedly disrupt the gait cycle, disrupt the base of support, or present abrupt and unexpected increases in the available frictional forces (e.g., walking from hard smooth surfaces onto carpet or much greater coefficient of friction flooring material). Stumbles are typically provoked by unexpected changes in the level, slope or other geometric properties of the walking surface (e.g., uneven or inappropriate rise:run ratios of stairs).

Regardless of the type of precursor, once the body's center of mass ventures outside of the standing or gait's effective base of support, and the base of support cannot be reestablished under the center of mass in a timely manner, the individual will fall (see Fig. 7.11).

Slip resistance is gaged by the available static Coefficient of Friction (COF) as determined below:

$$\text{COF} = \frac{\text{Force}_{\text{Horizontal}}}{\text{Force}_{\text{Normal}}}, \quad (7.12)$$

Table 7.1 Bases for fall deaths in US during 2007^a

Fall deaths in 2007 by type	Deaths	Percent of total	
Same level unknown	6,076	27%	30%
Slip and trip	691	3%	
Same level ice and snow	114	1%	
Stairs and steps	1,917	8%	10%
Ladder falls	366	2%	
Fall from building	587	3%	5%
Between levels unknown cause	507	2%	
Scaffolding	68	<1%	
Furniture	984	4%	54%
Wheelchair falls	392	2%	
Trees and cliffs	155	1%	
Diving into water	32	<1%	
Carrying person	31	<1%	
Unknown	10,636	47%	
Other	75	<1%	

^aTaken from National Center for Health Statistics (NCHS), National Vital Statistics System (2007)

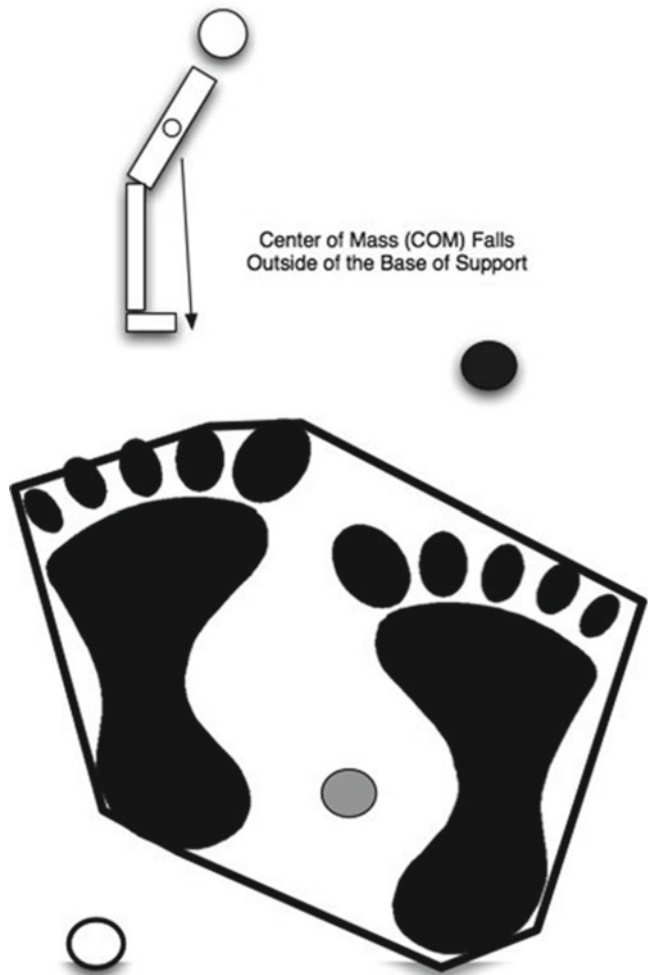
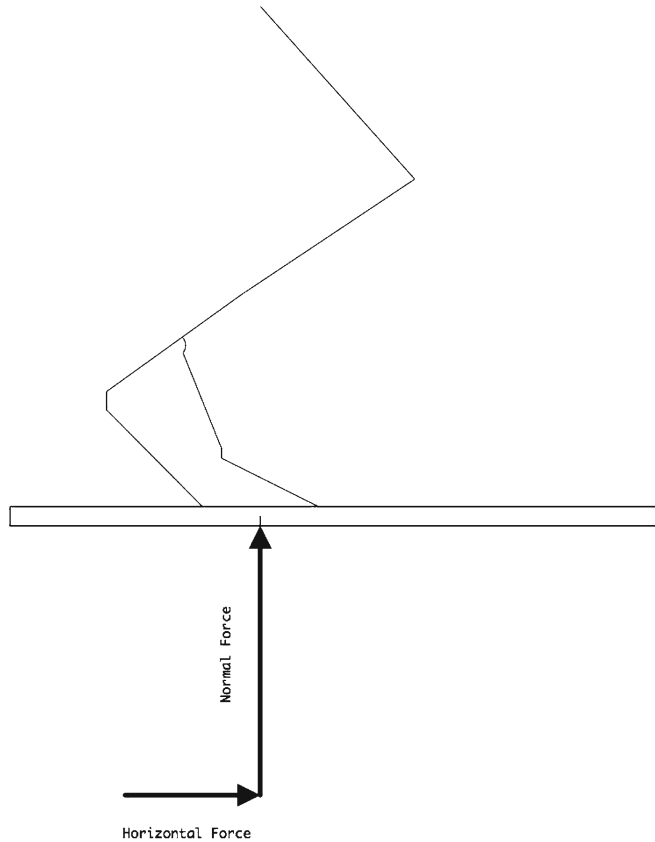


Fig. 7.11 Balance is lost if the center of mass (COM) falls out of the base of support (depicted by the polyhedron enclosing the feet in contact with the supporting surface).
Note: The circles represent potential centers of mass. The black and white COMs will result in a fall. The area COM will not produce a fall



$$COF = \frac{Force_{Horizontal}}{Force_{Normal}}$$

EQ. 12

Where:

$Force_{Horizontal}$ = Horizontal force required to initiate shoe slip (static) or to continue slip movement of the shoe (dynamic)

$Force_{Normal}$ = Force acting perpendicular to walking surface associated with shoe and loads placed atop of the shoe

Fig. 7.12 Schema for characterization of the static coefficient of friction (COF)

where $Force_{Horizontal}$ is the horizontal force required to initiate shoe slip (static) or to continue slip movement of the shoe (dynamic), $Force_{Normal}$ is the force acting perpendicular to walking surface associated with shoe and loads placed atop of the shoe (Fig. 7.12).

The horizontal force is that needed to just initiate a slip referred to as the static COF. The vertical force is the total weight of the shoe that is pulled across the walking surface. Once the slip is initiated, then horizontal forces recorded during the continuous slipping motion are used to determine the dynamic COF. Dynamic coefficients of friction can be greater or lesser than the static COF, and they typically vary in complex and nonlinear manners. Given that static COF behavior is reasonably

linear, simple to measure, provides good agreement within and between investigators, it has been adopted by the engineering community as the metric for slip resistance.

The Occupational Safety and Health Administration recommends that walking surfaces have a static coefficient of friction of not less than 0.5. Some slippage (i.e., microslips) is normal and should be maintained in walkways. Excessive COFs promote disruption of gait, trips and falls, and transition issues when moving from one COF surface to another. When surfaces are slanted such as wheelchair access or pedestrian ramps, small increases in COF are required to reduce the risk of slips (e.g., 0.6–0.8).

Typically, required coefficients of friction are determined by measurement or prediction of ground reaction forces of the feet when performing assigned work or activities (Berg and Norman 1996; Chambers and Sutherland 2002; Lord et al. 1986). If available static COFs are less than those required by gait or work activities, then slippage can be anticipated (Cham and Redfern 2001; Chambers and Cham 2007; Gao and Abeysekera 2004; Gronqvist et al. 2001; Menant et al. 2008; Redfern et al. 2001).

Mismatches between the one's mental model of a walking surface's properties or frictional resistance, rise:run ratios of steps or ladder rungs, presence of irregular step geometries or surface discontinuities, or expected COFs are probably more hazardous than the level of available COF (Tisserand 1985). If the COF is above 0.5 but one transitions from a 0.8 surface with a gait pattern that is unacceptable for the 0.5 surfaces, then a slip and fall can occur even when the surface friction is considered adequate from a frictional force perspective.

Vehicles and Impact Injuries

Aerospace, marine, and terrestrial vehicles [vehicles, trucks, snowmobiles, all-terrain vehicles (ATVs)] expose riders to musculoskeletal sprains, strains, fractures, and pain that can be confused with manual materials handling activities. Such expenses should be screened or filtered out of data sets where appropriate (Inamasu and Guiot 2007; Inamasu and Guiot 2009; Smith et al. 2005; Stemper and Storvik 2010). Excellent reviews of biomechanical criteria that are used to characterize risk of injury in vehicle collisions exist (Nordhoff 2005).

Whole-Body Vibration

As with slips and falls, musculoskeletal pain and injury studies should consider whole-body vibration exposures as possible contributors or cofounders to the overexertion injury model under consideration. An excellent and detailed review of the research performed regarding human vibration is provided by Griffin (1990).

Spinal injury and pain are associated with acute or chronic exposure to excessive whole-body vibration of seated occupants in some vehicles and machinery. The most widely used document on this topic is the Guide for the Evaluation of Human Exposure to Whole Body Vibration (ISO 2631) (Birlik 2009; Blood et al. 2010; Bovenzi and Zadini 1992; Futatsuka et al. 1998; Goglia and Grbac 2005; Griffin 1978; Ozkaya et al. 1994; Pope et al. 2002; Pope et al. 1999; Schust et al. 2010; Seidel 2005; Shoenberger 1979; Smets et al. 2010; Smith 2006; Troup 1978).

The ISO 2631 standard advocates measuring vibration exposures in the frequency range of 0.5–160 Hz using triaxial accelerometers that are placed at the human–seat interface. The recordings are evaluated for acceleration power within one-third octave bands of the aforementioned frequency band, and results for each band are weighted based upon a set of coefficients that have been fitted based upon experimental studies and consensus expert opinion. Presently, the maximum allowed limit is 1.15 m/s², with single action level set at 0.5 m/s².

Overuse Syndromes

Overuse, repetitive stress, cumulative trauma, and other terms have been assigned to a vast array of musculoskeletal disorders that produce injury and pain at or near articulations. Some would classify whole-body overexertion injuries as cumulative trauma disorders (CTDs) or overuse syndromes. As with falls, injury investigators must determine whether their musculoskeletal injury outcomes are due to excessive whole-body exertions or are a result of other occupational exposures, underlying pathos, etc. While whole-body overexertion risk factors may be shared with many overuse syndromes, overuse and overexertion injuries, as scoped within this chapter, are different species of musculoskeletal injury.

Prior to conducting whole-body overexertion injury studies, it would be helpful to understand overuse syndromes and the potential for confounding results with overuse outcomes (Bernard 1997; Putz-Anderson 1988). Studies should also consider dissenting perspectives prior to development etiological models and study design (Hadler 1987; Hadler 1990; Hadler 1992; Hadler 1993; Hadler 1997).

Physical Work Capacity and Workload Analysis

The rate at which work is performed determines the amount of energy that the body consumes. Knowing the magnitude and duration of the workload one can determine the amount of power that the worker is required to produce both aerobically and anaerobically. Whether one can meet the workload demands, or not, is determined by their aerobic power or maximum physical work capacity for a specified time period. If the aerobic power demands are extreme, which is rare, then risk of cardiac arrhythmia and arrest increases. Typical excessive aerobic power demands result in a decline in muscular capacity to perform physical work, and declines in motivation to continue work at a specified pace for a specified duration – commonly referred to as physical or workplace fatigue.

Systemic Fatigue

When the body is unable to supply sufficient oxygen and other metabolites to contracting muscles to meet their energy demands, muscle metabolism shifts increasingly from aerobic to anaerobic (i.e., glycolysis) metabolism; producing fewer ATP per unit of nutrient substrate, and lactic acid as a byproduct. This shift not only meters the muscle's force production capacity, it creates an oxygen debt that has to be repaid. Oxygen is required to process lactic acid, restore glycogen stores, and repair sarcomeres, and other structures that have undergone structural challenges (Åstrand and Rodahl 1986).

As the muscle begins to fatigue, the human experiences discomfort from increased lactic acid and force-induced ultrastructure changes and direct stimulation of pain fibers. Subsequently, the discomfort is modulated by an inflammatory response that may take hours to days to fully express itself (Chaffin et al. 2006).

Fatigue has been classified as either systemic or localized in nature. The only difference between the classifications is the scope of muscle involvement. Excessive exertions that are constrained to a signal or few localized muscle groups result in Localized Muscle Fatigue (LMF). If the affected muscle groups are widely distributed throughout the body (e.g., manual materials handling), the fatigue is considered systemic. It is important to differentiate the types of fatigue because measurement protocols must differ.

Systemic fatigue will alter blood chemistry as anaerobic metabolism progresses with large numbers of muscle groups. Cardiopulmonary response to depletion of oxygen and build-up of carbon dioxide in a wide distribution of working muscle is profound and reliable. Measurable changes in work behavior may be detected in response to the loss of whole-body strength and mounting discomfort, with concomitant changes in the motivation and affective state of the worker.

This response can be profound and easily detected with systemic fatigue metrics. Extensive bibliographies are provided by others (Åstrand and Rodahl 1986). Systemic metrics are not reliable indicators of LMF because tissue involvement is comparatively small.

Cardiac output is directly linked to the body's capacity to provide nutrients to, and remove metabolic end-products from, working muscles. Cardiac output is defined as the product of cardiac stroke volume, or left-ventricular ejection volume, and contraction or heart rate:

$$\text{Cardiac output} = \text{Stroke Volume} \times \text{Heart Rate} \quad (7.13)$$

Cardiac output demands near a worker's resting levels (e.g., performing seated desk work) are addressed by increasing ventricular ejection or stroke volumes – without increase in heart rate. Use of heart rate as a systemic fatigue or physical workload metric when physical demands are low is fruitless, and promotes Type II errors when addressing fatigue in the office or similar low-effort environments that may produce material levels of LMF and discomfort.

As physical workloads increase beyond resting levels cardiac output demand is addressed chiefly by increased left-ventricular ejection, or stroke, volume. Stroke volume elevations are limited. Workloads that produce systemic fatigue quickly overwhelm initial stroke volume elevation capacity. Stroke volume in the above equation essentially becomes a constant, leaving heart rate the sole driver of subsequent cardiac output.

Strong monotonic and linear relationships exist among physical workload, oxygen consumption, pulmonary ventilation and heart rate, and increases in body core and skin temperature in controlled environmental conditions (Åstrand and Rodahl 1986). The aforementioned physiological metrics are also strongly correlated with psychophysical ratings of physical workloads (Borg 1976; Wiker 1990; Wiker et al. 1989a, b). Thus, selection of any of these metrics is an acceptable surrogate for the measurement of physical workload via measured oxygen consumption rates (Brouha 1964; Brouha and Harrington 1957; Maxfield and Brouha 1963; Monod and Garcin 1996) (Fig. 7.13).

Heart rate is often used to gauge energy consumption. It is strongly correlated with cardiopulmonary system performance. Mask-based measurement of oxygen consumption and carbon dioxide production is intrusive and produces discomfort. Subject cooperation wanes, and focus on the irritation associated with facial pressure created by the mask mounts, after only an hour of wearing such apparatus. Heart rate monitoring is less intrusive and more comfortable.

Heart rate, when compared with a worker's cardiac reserve, indicates physical workload:

$$\text{CR} = (220 - \text{Age}) - \text{HR}_{\text{rest}}, \quad (7.14)$$

where:

- CR = cardiac reserve (heart rate range)
- Age = age of worker in years
- HR_{rest} = resting heart rate

Ergonomic design guidelines for mitigation of systemic fatigue during a typical work shift are not to exceed one-third of the worker's physical work or associated cardiac reserve capacity:

$$P = \frac{\text{HR} - \text{HR}_{\text{rest}}}{\text{CR}} \times 100, \quad (7.15)$$

where:

- P = percent of cardiac reserve or aerobic capacity used
- CR = cardiac reserve (heart rate range)
- HR_{rest} = resting heart rate

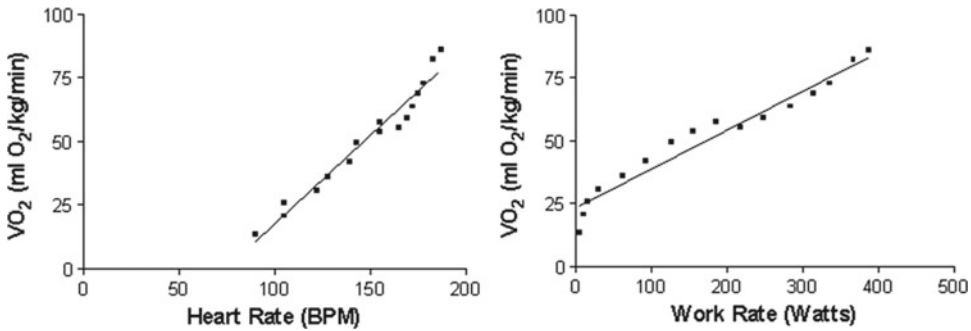


Fig. 7.13 Example of strong relationships found among cardiopulmonary metrics of physical workload. http://www.nismat.org/physcor/max_o2.html

Injury investigators may be best served if heart rate or a global metric of cardiopulmonary demand is combined with a psychometric assessment of global and regional discomfort. Psychometric methods (e.g., discomfort surveys using cross-modal matching methods) will correlate well with heart rate and will point to the muscle groups or distribution of exertions that are responsible for onset of systemic fatigue.

Heart rate cannot be used to detect or gage LMF. Other metrics such as EMG, psychometric tools, tremor, and other metrics must be used. Which metric is best depends upon the locus and magnitude of LMF (Wiker et al. 1989a, b).

Metabolic Energy Prediction Models

Direct measurement of cardiopulmonary demand may not be feasible for various reasons (e.g., lack of cooperation on the part of labor management, historical changes have occurred in the work process, etc.). Metabolic energy expenditure prediction models can be used to estimate the aerobic power demands of the job or activity. The Department of Labor and a number of investigators have performed energy expenditure rate measurements for a wide array of work tasks (e.g., shoveling, carrying, etc.). The tasks may have equations that consider variables that impact the energy expenditure prediction. For example, the energy required to carry a box depends upon the carriage posture, the weight of the load, gait velocity, grade of the walking surface, and distance the load is carried.

Knowing the metrics that have to be evaluated to produce accurate predictions of energy expenditure, one can design their task analysis of the job(s) to insure that such information is obtained. Once the information is obtained, then characteristics of the workforce (gender, body mass, etc.) and task activities (e.g., durations, rest periods, loads, walking velocities, grades, etc.) may be used to predict metabolic energy expenditure for tasks or jobs.

If tasks performed are spanned by industrial tasks studied by Garg et al. (1978), then one can use task characteristics, gender, body mass, duration of work, three general classes of posture to estimate energy expenditure rates. The model was validated with 48 different industrial jobs; producing a correlation between predicted and measured oxygen consumption rates of 0.95 and a low coefficient of variation (i.e., 10.2%):

$$MR = \frac{\sum_{i=1}^n E_{\text{task}_i} + \sum_{i=1}^n E_{\text{posture}_{\text{task}_i}}}{\sum_{i=1}^n \text{Time}_{\text{task}_i}}, \quad (7.16)$$

where:

MR	=	metabolic rate (kcal/min)
E_{task}	=	kcal performing the i th task
E_{posture}	=	kcal associated with either standing, sitting or standing bent
Time _{task}	=	time

The advantages of predicting metabolic energy expenditure are: (a) it only requires a task analysis to gather the predictor metrics, (b) it saves time and costs, (c) points tasks and task characteristics that may be driving excessive metabolic energy expenditure, and (d) allows one to evaluate the impact of the modified job upon metabolic energy expenditure and risk of systemic fatigue. Model limitations are that: (a) many industrial tasks are not addressed (e.g., cart push and pull, climbing/descending ladders, etc.), and (b) highly unstructured work can produce hundreds of variations of tasks that require analysis.

Systemic fatigue impedes physical work production and increases the length of time needed to recover from oxygen debt created by greater anaerobic metabolic activity. If a fatigued worker finds a means to short-cut workload at the expense of safety, fatigue may promote unsafe behaviors.

Often injury investigators are focused so heavily upon pathophysiological outcomes that they fail to consider the social impact upon the workforce. Workers who are taxed too heavily at work must eat more and often sleep through nonoccupational periods of their life, (e.g., spend weekends resting to recover for the next week of exertions). This outcome has both economic and social consequences for the affected workforce.

Localized Muscle Fatigue

Excessive exertions in small groups of muscles produces LMF, acute or chronic strain, without signs of systemic fatigue. Muscle contraction intensities and rates can be sufficient to overwhelm perfusion-based supplies of nutrients and removal of waste products. During contractions, intramuscular pressures increase and, thereby, impede muscle perfusion. As contractions wane, perfusion increases. Depending upon the contraction intensity and contraction:relaxation ratios (i.e., work-rest cycles), LMF may or may not occur.

LMF often leads to small adjustments in postures to alter patterns and intensities of muscle recruitment to “rest” working tissues and, thereby, stave off onset of LMF. Injury investigators should evaluate the potential for this type of behavior before making selections for LMF metrics.

Historically, LMF has been detected by declines in muscle force production (strength) capacity (Rohmert 1973a; Rohmert 1973b), increased extremity tremor (Hefter et al. 1987; Wiker et al. 1989a, b; Young and Hagbarth 1980), elevations in moving windows of EMG ensemble root mean square (RMS) or negative shifts in spectral density functions (Bigland-Ritchie 1981a; Bigland-Ritchie 1981b; Bigland-Ritchie et al. 1981; Mathiassen 1993; Wiker et al. 1989a, b), and onset of reports of perceived loss of muscle contractility or discomfort (Corlett and Bishop 1976; Wiker et al. 1989a, b). Each metric has its strengths and weaknesses. Generally, specific measures must be chosen for specific exposures and study objectives – often requiring careful thought and rationale for selection.

Above 20% MVC, endurance and strength begin to decline with sustained exertions – if exertion:rest durations are sufficient. If there is sufficient rest between exertions, then perfusion may be adequate to wash away indexes of LMF. Until exertion intensities produce reductions in perfusion rates of 50% or greater, EMG metrics of LMF do not appear to be sensitive or reliable. If 50% MVC or more impedes perfusion rates, then use of RMS values of myoelectric ensembles is likely to be more statistically powerful than monitoring negative spectral density (Wiker et al. 1989a, b).

Fig. 7.14 A representative WBGT monitor



There is little disagreement that LMF and associated discomfort are indicators of potential ultra-structural musculoskeletal damage. At what point exertions become injurious is uncertain.

Thermal Stress

Physical workloads determine metabolic heat production and internal heat burden. The direction and rate of heat flow between the worker and their environment determines the type and severity of thermal stress. If heat is pulled from the body due to much colder environs, the worker risks hypothermia and other forms of cold injury. If the environment is sufficiently hot to force heat flow into the body, then risk of heat injuries (e.g., syncope, rash, burns, heat exhaustion, and stroke) become material (Malchaire 1994). The heat balance equation characterizes heat flow due to internal metabolism, conduction, convection, radiation, and evaporation:

$$H = M \pm K \pm C \pm R - E, \quad (7.17)$$

where:

- H = body heat content
- M = metabolic heat production
- K = conduction gain or loss
- C = convection gain or loss
- R = radiation gain or loss
- E = evaporative loss

The heat balance equation is most directly evaluated using a Wet-Bulb-Globe-Thermometer (WBGT) measurement system; a representative apparatus is shown in Fig. 7.14.

The WBGT monitor can switch from indoor readings, without solar radiation, providing:

$$\text{WBGT} = 0.7\text{NWB} + 0.3\text{GT}, \quad (7.18)$$

To outdoor readings with solar radiation:

$$\text{WBGT} = 0.7\text{NWB} + 0.2\text{GT} + 0.1\text{DB}, \quad (7.19)$$

where:

- WBGT = wet bulb globe temperature index
- NWB = natural wet-bulb temperature
- DB = dry-bulb (air) temperature
- GT = globe thermometer temperature

The average WBGT for a shift, including break and rest interval WBGT exposures, is determined using the following:

$$\text{WBGT}_{\text{Average}} = \frac{\sum_{i=1}^n \text{WBGT}_i \times t_i}{\sum_{i=1}^n t_i}, \quad (7.20)$$

where:

- WBGT_i = WBGT for the time interval (t_i)
- $\text{WBGT}_{\text{Average}}$ = the average WBGT over the work period

If the worker is storing heat, then NIOSH recommends reducing the metabolic heat production, or physical workload, convective, conductive, or radiation gains to return heat storage content to an acceptable equilibrium.

Example Analysis

Often, startup companies initiate operations without adequate health and safety staff, and design work without conforming with ergonomic design principles. In this example, a startup tasked workers with feeding live stock. The workers offloaded forty 25 kg 24×16×5 in. plastic bags of meal, that were palletized and delivered in a standard shipping container, onto the tailgate of a pickup truck bed. The bags were then palletized in the truck bed. The truck was driven 10–15 min to a site where the bags were offloaded from the truck, carried to the vessel, and palletized on the dock adjacent the vessel. The bags were then transferred from the dock to the vessel and palletized on the forward deck of the vessel. The workers drove the boats to floating cages, lifted bags of meal into a pumped-deployment hose and took turns diving with scuba gear to place feed hose close to the fish and to determine when the fish were satiated. Feedings reoccurred throughout a shift. After the last feeding, the workers returned the vessels to the harbor marina and drove the truck to the container site – ending a typical shift.

Nearly half of the workers had experienced back injuries, all complained of excessive physical fatigue (psychophysical ratings of eight of ten using a modified Borg Scale) and all reported a

lifestyle of going to sleep immediately after dinner and spending most weekends resting to recover for the next week of work. All workers were very lean with no history of thermal stress injuries.

No manual material handling health and safety training was provided. Task training was provided by on-the-job supervision. Examination of sensory perceptual, mental workload, precision motor demands indicated demands were low and within design limits. Bags were frequently found breeched within the container where floors and the bags were coated with an oily and slippery fish-meal; thus risk of slipping and loss of grasp would have to be considered. Workers, who were all young adult men, wore bathing suit/shorts, t-shirts, and flip-flops while performing bag-handling tasks. The WBGT in the work area ranged from 25 to over 30°C throughout the work year.

An operational analysis was performed to determine the sequencing, timing, and human–task–equipment interface demands. Lifting tasks were recorded on video, determined to be quasi-static in nature, and lighter bag surrogate was used to eliminate risk of injury during measurements. The University of Michigan Static Strength Prediction Model was used to evaluate predicted lumbar disk compression and population static strength demands. The Metabolic Energy Prediction Model (Garg et al. 1978) was used to evaluate metabolic energy expenditures. A WBGT assessment was made to determine if metabolic energy expenditures presented risk of thermal injury for the range of WBGT exposures. The injury risk focus was on bag handling, so additional analyses associated with diving operations were excluded from analyses for this example. Table 7.2 summarizes the operational analysis findings for the fish-meal handling tasks. Informal methods were included in the analyses.

Tasks 1–7 required about 1 h and 45 min to handle transport and movement of up to about a ton of fish food onto the vessel. The bags were lifted and lowered three times (300 bag lifts and lowers in 105 min) for bow loaded paradigms. Each worker handled 4–5 tons of lifting during a shift.




Task 9 required workers to lift a bag and pour and blend the fish meal with piped water and use a venturi effect to pull the fish meal into the water stream passing via hose down to the fish in the submerged cage. One of the workers would then become a diver, descend to a submerged cage, feed the fish and then would have to ascend back to the boat pulling the hose up with them. That process took typically 1 h to complete. Divers were in the water for 2 h per shift. All work was completed typically in 10–12 h. The work schedule was typically 4 days on, and 3 days off and varied with 2 work: 1 rest or 2 work: 2 rest day cycles.

Biomechanical Analysis

A total of 126 biomechanical analyses were performed using the University of Michigan Static Strength Prediction Model (3DSSP) to obtain estimated lumbar disk (L5/S1) compression and joint load:strength moment ratios to determine the proportion of the working population that would have sufficient strength to perform the assigned tasks using slow and controlled exertions. Each activity recorded was recorded with video, examined and postures which presented the greatest load moments acting upon the body were selected for analysis to reduce analytical costs. Lumbar disk compressions were compared against the NIOSH Action Limit (3,400 N) and Maximum Permissible Limit (6,400 N) lines to define hazard levels.




Results are summarized for lumbar compression analyses in the following figure. The hip proved to be taxed most from a strength demands perspective. All joints demonstrated behavior similar to that shown in the strength plot for the hip (Figs. 7.15 and 7.16).

Table 7.2 Activity description of workers handling bags of fish meal

Task	Description	Interfaces	Representative photographs
1	Offload fish-meal bags from container onto pickup bed	40 bags offloaded ^a . BA ^b , MEP ^c , COF ^d , WBG ^e	
2	Palletize bags in pickup	40 bags palletized ^f . BA, MEP, COF, WBG ^g	
3	Drive to marina No accident history	10–15 min	
4	Offload bags for carry to dock adjacent to boat	40 bags offloaded. BA, MEP, COF, WBG ^h	



(continued)

Table 7.2 (continued)

Task	Description	Interfaces	Representative photographs
5	Transfer bags to worker on dock to avoid excessive carry distance	40 bags transferred. BA, MEP, COF, WBGT	
6	Transfer bags from dock to bow of vessel	40 bags transferred. BA, MEP, COF, WBGT	
7	Palletize bags in bow of vessel	40 bags. BA, MEP, COF, WBGT	

(continued)

Table 7.2 (continued)

Task	Description	Interfaces	Representative photographs
8	Drive the boat 45 min to get to the cages for the fish and they would typically arrive at the cages about 9:00 a.m.	WBV [§]	
9	Cut, lift, and pour meal into feeder pump	40 bags. BA, MEP, COF, WBGT	
10	Transfer bag to shoulder of coworker who stands on floating pipe and pours or spreads contents onto cage water surface	Few bags and only when small fish are present BA, MEP, COF, WBGT. Transfer strongly mediated by momentum by the transferring worker-dynamic biomechanical analysis is required	
11	Scuba diving cyclically for 30 min to feed fish	Heat balance	
12	Surface feed small fish by emptying bag from shoulder into surface water of cage		

(continued)

Table 7.2 (continued)

Task	Description	Interfaces	Representative photographs
13	Drive the boat 45 min to harbor, wash down and secure vessel, drive truck to container area		

^aOily fish-meal leaked out of broken bags-making walking surfaces and plastic bags slippery. All bag handling and walking surfaces will be considered for COF confounders

^bBA biomechanical analysis to determine strength demands and lumbar disk compression exposures

^cPerform metabolic energy expenditure analysis

^dPerform COF analysis to determine if slips are occurring and increasing risk of low-back injury

^eWBGT analysis to determine if thermal injury risk exists

^fOily fish-meal leaked out of broken bags-making walking surfaces and plastic bags slippery. All bag handling and walking surfaces will be considered for COF confounders

^gExamine WBV exposures using ISO 2631

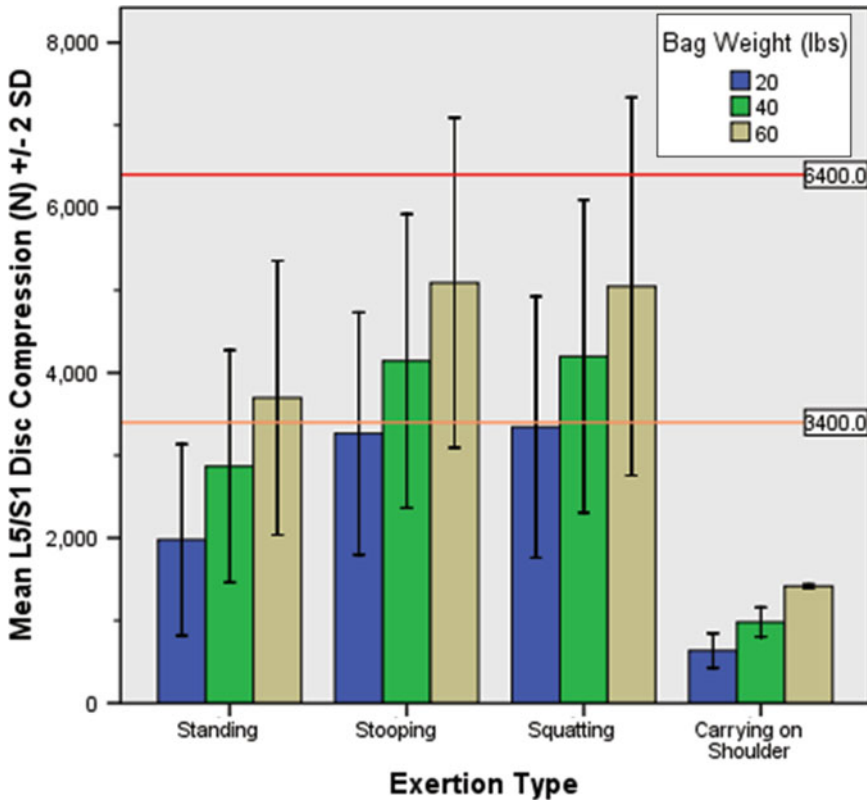


Fig. 7.15 Mean (± 2 SD) lumbar compression forces for types of whole-body exertions across all bag handling and transfer activities

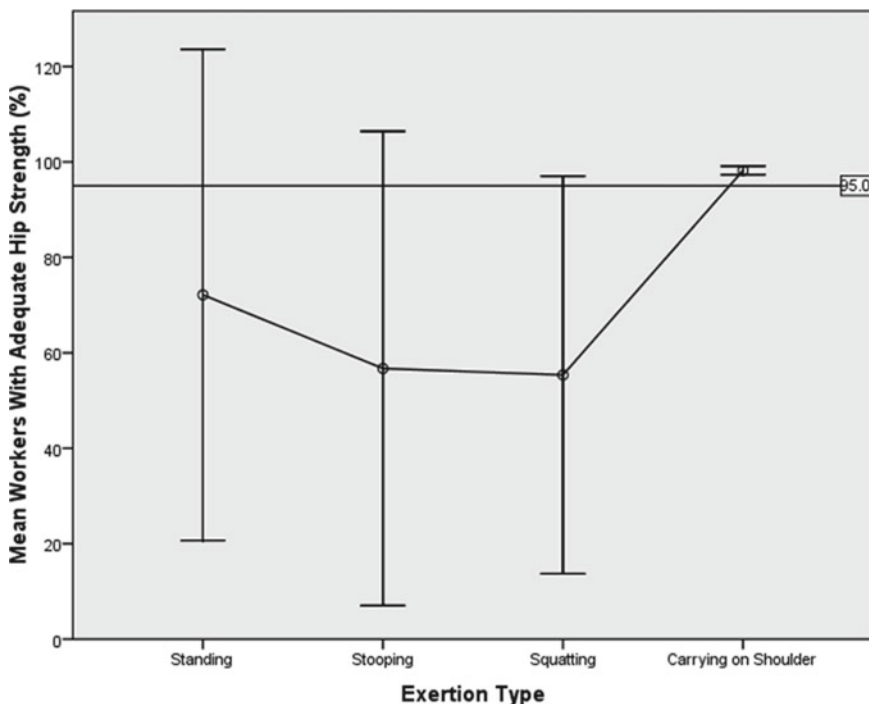


Fig. 7.16 Mean (± 2 SD) worker population hip extension strength capabilities for observed bag handling exertion types

Metabolic Energy Expenditure Analysis

Metabolic energy expenditure prediction (Garg et al. 1978) analysis was performed on tasks that included the numbers of bags handled, distances carried, walk back distances, lift and lower geometries. To handle day-to-day variations in the number of bags handled a Monte Carlo simulation analysis using 25,000 random variations in numbers of bags lifted, lowered, and carried, provided averages and variances for metabolic energy expenditure using triangular, normal, and uniform distributions for the aforementioned parameters.

For an average industrial male, their maximum aerobic power capacity for an eight-hour shift is approximately 15 kcal/min. A Borg scale rating of eight of ten (max), provided by workers, was consistent with a job energy expenditure of 10.8 kcal/min (i.e., 80% of the difference between maximum aerobic capacity and basal metabolic rate 15–1.5 kcal/min) (Fig. 7.17).

Thermal Stress Analysis

The metabolic energy expenditure of 10.3 kcal/min was compared against NIOSH WBGT exposure limits in the following plot. The results show that workers are at risk of a thermal stress injury. Mediating risk factors were that heavy lifting work was performed early in the mornings, workers wore limited clothing, and work at sea involved significant conductive heat loss while scuba diving (Fig. 7.18).

Worker Characteristics		%		Kcal	
Gender (1=Male, 0=Female)	1.00				
Body Mass (Kg)	75.00				
Cycle Time (min)	480.00				
Percent of Time Spent					
Sitting Posture	13.8			114.2	
Standing Posture	43.1			372.4	
Standing, Bent Posture	43.1			434.5	

Task	Repetition Per Cycle	Push or Pull Force (Kgs)	Initial Lift Height (m)	Ending Lift Height (m)	Load Lifted (kg)	Walk Speed (m/s)	Walking Grade (%)	Horizontal Movement (m)	Duration (min)	Task Kcal
Lift: Squat Out of Container	22.16	9.78	0.62	1.00	25.00				7.61	
Lift: Stoop Out of Container	33.61	31.94	0.72	1.00	25.00				4.19	
Carrying - Load at Waist/Thighs at Container	33.61	31.94	0.80	1.00	25.00	0.68	0.00		3.32	542.34
Walking	22.16	9.78	0.47	1.00	25.00	0.92	0.00		2.44	173.33
Lower: Stoop onto Tailgate	22.16	9.78	0.56	1.00	25.00				0.12	11.10
Lift: Squat Out of Truck	31.94	22.16	0.33	1.00	25.00	0.66	0.00		3.23	495.09
Carrying - Load at Waist/Thighs at Container	9.78	22.16	0.36	1.00	25.00				14.47	5.72
Lift: Squat From Dock	22.16	9.78	0.53	1.00	25.00				9.64	9.64
Lift: Squat From Boat Deck to Feeder	9.78	31.94	0.38	1.00	25.00				5.63	5.63
Lift: Stoop From Boat Deck to Feeder									120.63	2986.11
Holding Bag to Feed Meal Tube										0.00
Other Activity										0.00

Posture Component (Kcal):	921.11
Task Component (Kcal):	4260.20
TOTAL JOB ENERGY DEMANDS (Kcal):	5181.31
JOB DURATION (Min):	480.00
JOB ENERGY EXPENDITURE RATE (KCAL/MIN):	10.79

Fig. 7.17 Summary of task and metabolic analysis for fish meal bag lifting, lowering, and carrying tasks

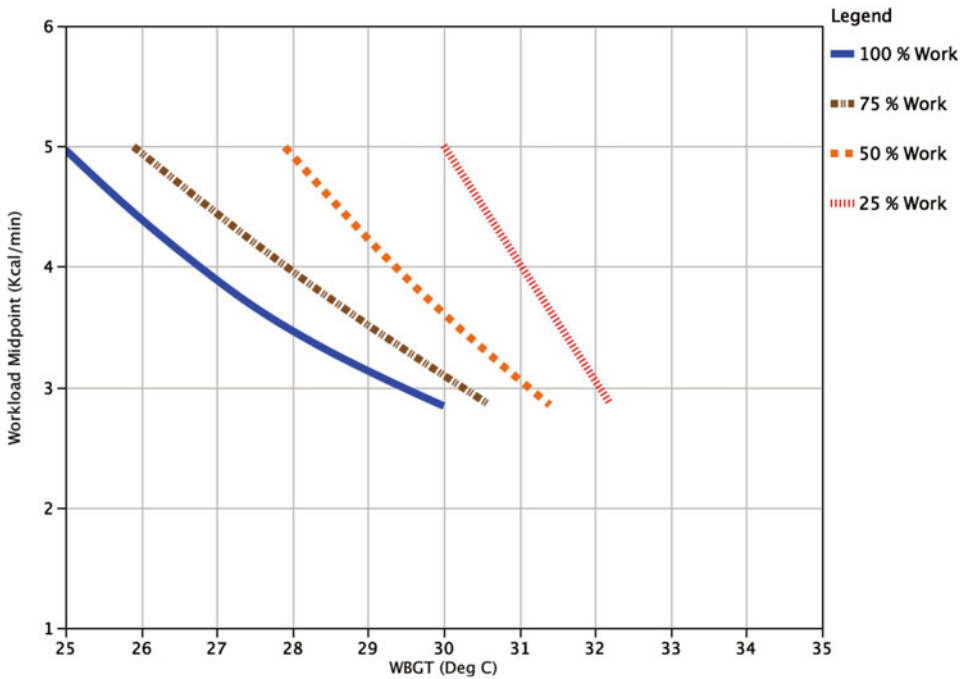


Fig. 7.18 Relationship between environmental conditions, metabolic energy expenditure, and work:rest ratios recommended by NIOSH for prevention of thermal injury

Intervention

Given the material risk of overexertion injury and strain based upon static strength and lumbar disk compression analysis, excessive metabolic energy expenditures, and material risk of thermal stress, no dynamic biomechanical analyses or COF tests were performed. Injury incidence, psychophysical ratings of strain, and strategies used to cope with excessive metabolic strain (e.g., Task 5) were consistent with the type and magnitudes of overexertion stressors.

The company changed the job by purchasing 1 ton bags of fish meal, placing the container adjacent to the marina's boat ramp, and off-loading, transferring and lowering the single bag into the bow of the vessel at the boat ramp. This eliminated human lifting, lower, carries and transfers of bags. The pump was modified to allow deployment of fish meal on the surface for small fish, and pumping meal to the larger fish at deeper depths. Injury incidence returned to idiopathic levels, work shifts shortened, metabolic energy demands were reduced to levels below 5 kcal/min, thermal stress risks were reduced to safe levels, and strength demands for all joints returned to 95 percentile accommodation limits (Fig. 7.19).

Conclusion

Poor design of machines and tools, tasks, and working environments challenge human performance capacities and tissue tolerances; resulting in greater risk for experiencing accidents, injuries, or illnesses. Errors in design may affect recognition of hazards, overload the cognitive capacity of the



Fig. 7.19 View of fork-truck now used to lift one-ton bags of fish meal for transport to and placement into boats

human and force errors, create excessive motor demands, and expose one to excessive biomechanical, metabolic, and thermoregulatory stressors. As design flaw(s) increase in breadth or depth, their impact upon human performance, safety and health becomes far more interactive and distributed in nature. The inherent complexity of their impact may require injury investigators to “peel away” at the problem through a cycle of series of studies and administrative or design interventions.

Fortunately, injurious and unsafe designs have consistently demonstrated performance, quality and cost saving dividends when design flaws have been removed or mitigated. Such dividends should be capitalized upon and be given careful consideration when shaping injury study designs. Understanding the efficacy of design or administrative policy modifications in reducing incidence and severity of injury shares importance with understanding the mechanism or etiology of the injury or illness of immediate concern.

This chapter focused upon a small set of whole-body overexertion injuries that create the largest share of industrial injuries. However, the approach advocated provides a general template for the study and mitigation of a broader range of injuries that are induced by poor ergonomic design.

References

- (1966). *Perception and psychophysics*. Austin, TX: Psychonomic Society.
- Åstrand, P., & Rodahl, K. (1986). *Textbook of work physiology: Physiological bases of exercise*. New York: McGraw Hill.
- Bahr, R., Reeser, J. C., Fédération Internationale de Volleyball. (2003). Injuries among world-class professional beach volleyball players. The fédération internationale de volleyball beach volleyball injury study. *The American Journal of Sports Medicine*, 31(1), 119–25.
- Baumert, M., Brechtel, L., Lock, J., Hermsdorf, M., Wolff, R., Baier, V., et al. (2006). Heart rate variability, blood pressure variability, and baroreflex sensitivity in overtrained athletes. *Clinical Journal of Sport Medicine*, 16(5), 412–7.
- Berg, K., & Norman, K. E. (1996). Functional assessment of balance and gait. *Clinics in Geriatric Medicine*, 12(4), 705–23.

- Bernard, B. P. (1997). Musculoskeletal disorders and workplace factors: A critical review of epidemiologic evidence for work-related musculoskeletal disorders of the neck, upper extremity, and low back. Washington, DC: NIOSH.
- Bigland-Ritchie, B. (1981a). EMG/force relations and fatigue of human voluntary contractions. *Exercise and Sport Sciences Reviews*, 9, 75–117.
- Bigland-Ritchie, B. (1981b). EMG and fatigue of human voluntary and stimulated contractions. *Ciba Foundation Symposium*, 82, 130–56.
- Bigland-Ritchie, B., Donovan, E. F., & Roussos, C. S. (1981). Conduction velocity and EMG power spectrum changes in fatigue of sustained maximal efforts. *Journal of Applied Physiology: Respiratory, Environmental and Exercise Physiology*, 51(5), 1300–5.
- Birlik, G. (2009). Occupational exposure to whole body vibration-train drivers. *Industrial Health*, 47(1), 5–10.
- Blood, R. P., Ploger, J. D., & Johnson, P. W. (2010). Whole body vibration exposures in forklift operators: Comparison of a mechanical and air suspension seat. *Ergonomics*, 53(11), 1385–94.
- Booher, H. R., & Knovel. (2003). *Handbook of human systems integration*. Hoboken, NJ: Wiley-Interscience.
- Borg, G. (1976). Simple rating methods for estimation of perceived exertion. In G. Borg (Ed.), *Physical Work and Effort*, pp. 39–47, Oxford: Pergamon Press.
- Bovenzi, M., & Zadini, A. (1992). Self-reported low back symptoms in urban bus drivers exposed to whole-body vibration. *Spine*, 17(9), 1048–59.
- Bower, G. H. (1977). *Human memory: Basic processes: Selected reprints with new commentaries, from the psychology of learning and motivation*. New York: Academic Press.
- Brouha, L. (1964). Physiological aspects of work measurement. *Occupational Health Review*, 16, 3–7.
- Brouha, L., & Harrington, M. E. (1957). Heart rate and blood pressure reactions of men and women during and after muscular exercise. *The Journal-Lancet*, 77(3), 79–80.
- Budgett, R. (1990). Overtraining syndrome. *British Journal of Sports Medicine*, 24(4), 231–6.
- Cermak, L. S., & Craik, F. I. M. (1979). *Levels of processing in human memory*. Hillsdale, N.J. New York: Lawrence Erlbaum Associates; distributed by Halsted Press Division, Wiley.
- Chaffin, D. B. (1969). A computerized biomechanical model—development of and use in studying gross body actions* 1. *Journal of Biomechanics*, 2(4), 429–441.
- Chaffin, D. B., Andersson, G., & Martin, B. J. (2006). *Occupational biomechanics*. Wiley-Interscience.
- Cham, R., & Redfern, M. S. (2001). Lower extremity corrective reactions to slip events. *Journal of Biomechanics*, 34(11), 1439–45.
- Chambers, A. J., & Cham, R. (2007). Slip-related muscle activation patterns in the stance leg during walking. *Gait and Posture*, 25(4), 565–72.
- Chambers, H. G., & Sutherland, D. H. (2002). A practical guide to gait analysis. *The Journal of the American Academy of Orthopaedic Surgeons*, 10(3), 222–31.
- Chapanis, A. (1965). *Research techniques in human engineering*. Baltimore: The Johns Hopkins University Press.
- Cobb, W. S., Burns, J. M., Kercher, K. W., Matthews, B. D., James Norton, H., & Todd Heniford, B. (2005). Normal intraabdominal pressure in healthy adults. *The Journal of Surgical Research*, 129(2), 231–5.
- Contini, R., Drillis, R. J., & Bluestein, M. (1963). Determination of body segment parameters. *Human Factors*, 5, 493–504.
- Corlett, E. N., & Bishop, R. P. (1976). A technique for measuring postural discomfort. *Ergonomics*, 19, 175–82.
- Craik, F. I. M., & Salthouse, T. A. (2007). *The handbook of aging and cognition*. New York, NY: Psychology Press.
- Davis, S. F. (2003). *Handbook of research methods in experimental psychology*. Malden, MA, Oxford: Blackwell Pub.
- Davis, P. R., Stubbs, D. A., & Ridd, J. E. (1977). Radio pills: Their use in monitoring back stress. *Journal of Medical Engineering and Technology*, 1(4), 209–12.
- De Greene, K. B., & Alluisi, E. A. (1970). *Systems psychology*. New York: McGraw-Hill.
- Dempster, W. T., & Gaughran, G. R. L. (1967). Properties of body segments based on size and weight. *American Journal of Anatomy*, 120(1), 33–54.
- Dennis, H., Dowling, J., & Ryan, R. F. (1975). *Abdominal hernias*. New York: Appleton-Century-Crofts.
- Durso, F. T., & Nickerson, R. S. (2007). *Handbook of applied cognition*. Chichester; New York: Wiley.
- Englander, F., Hodson, T. J., & Terregrossa, R. A. (1996). Economic dimensions of slip and fall injuries. *Journal of Forensic Sciences*, 41(5), 733–46.
- Estes, W. K. (1975). *Handbook of learning and cognitive processes*. Hillsdale, N.J. New York: L. Erlbaum Associates; distributed by the Halsted Press Division of Wiley.
- Fechner, G. T., Adler, H. E., Howes, D. H., & Boring, E. G. (1966). *Elements of psychophysics*. New York: Holt Rinehart and Winston.
- Fisher, B. O. (1967). *Analysis of spinal stresses during lifting*. Unpublished Masters Thesis, University of Michigan, Department of Industrial and Operations Engineering.
- Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–91.

- Fitts, P. M. (1958). Engineering psychology. *Annual Review of Psychology*, 9, 267–94.
- Fitts, P. M., & Deininger, R. L. (1954). S-R compatibility: Correspondence among paired elements within stimulus and response codes. *Journal of Experimental Psychology*, 48(6), 483–92.
- Fitts, P. M., & Peterson, J. R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology*, 67, 103–12.
- Fitts, P. M., & Radford, B. K. (1966). Information capacity of discrete motor responses under different cognitive sets. *Journal of Experimental Psychology*, 71(4), 475–82.
- Fitts, P. M., & Seeger, C. M. (1953). S-R compatibility: Spatial characteristics of stimulus and response codes. *Journal of Experimental Psychology*, 46(3), 199–210.
- Fitts, P. M., Weinstein, M., Rappaport, M., Anderson, N., & Leonard, J. A. (1956). Stimulus correlates of visual pattern recognition: A probability approach. *Journal of Experimental Psychology*, 51(1), 1–11.
- Freivalds, A., Chaffin, D. B., Garg, A., & Lee, K. S. (1984). A dynamic biomechanical evaluation of lifting maximum acceptable loads. *Journal of Biomechanics*, 17(4), 251–62.
- Futatsuka, M., Maeda, S., Inaoka, T., Nagano, M., Shono, M., & Miyakita, T. (1998). Whole-body vibration and health effects in the agricultural machinery drivers. *Industrial Health*, 36(2), 127–32.
- Gaillard, A. W. (1993). Comparing the concepts of mental load and stress. *Ergonomics*, 36(9), 991–1005.
- Gallagher, S., Hamrick, C. A., Love, A. C., & Marras, W. S. (1994). Dynamic biomechanical modelling of symmetric and asymmetric lifting tasks in restricted postures. *Ergonomics*, 37(8), 1289–310.
- Gao, C., & Abeysekera, J. (2004). A systems perspective of slip and fall accidents on icy and snowy surfaces. *Ergonomics*, 47(5), 573–98.
- Gardiner, J. M. (1976). *Readings in human memory*. London: Methuen.
- Garg, A., Chaffin, D. B., & Herrin, G. D. (1978). Prediction of metabolic rates for manual materials handling jobs. *American Industrial Hygiene Association Journal*, 39(8), 661–74.
- Gescheider, G. A. (1976). *Psychophysics: Method and theory*. Hillsdale, N.J. New York: L. Erlbaum Associates; distributed by Halsted Press.
- Gescheider, G. A. (1984). *Psychophysics: Method, theory, and application*. Hillsdale, NJ: L. Erlbaum.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.
- Goglia, V., & Grbac, I. (2005). Whole-Body vibration transmitted to the framesaw operator. *Applied Ergonomics*, 36(1), 43–8.
- Goldish, G. D., Quast, J. E., Blow, J. J., & Kuskowski, M. A. (1994). Postural effects on intra-abdominal pressure during valsalva maneuver. *Archives of Physical Medicine and Rehabilitation*, 75(3), 324–7.
- Granata, K. P., Marras, W. S., & Davis, K. G. (1997). Biomechanical assessment of lifting dynamics, muscle activity and spinal loads while using three different styles of lifting belt. *Clinical Biomechanics (Bristol, Avon)*, 12(2), 107–115.
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: R. E. Krieger Pub. Co.
- Greenwood, J., & Parsons, M. (2000). A guide to the use of focus groups in health care research: Part 2. *Contemporary Nurse*, 9(2), 181–91.
- Griffin, M. J. (1978). The evaluation of vehicle vibration and seats. *Applied Ergonomics*, 9(1), 15–21.
- Griffin, M. J. (1990). *Human vibration handbook*. New York, NY: Academic Press.
- Gronqvist, R., Abeysekera, J., Gard, G., Hsiang, S. M., Leamon, T. B., Newman, D. J., et al. (2001). Human-centred approaches in slipperiness measurement. *Ergonomics*, 44(13), 1167–99.
- Hadler, N. M. (1987). Regional musculoskeletal diseases of the low back. Cumulative trauma versus single incident. *Clinical Orthopaedics and Related Research*, 221, 33–41.
- Hadler, N. M. (1990). Cumulative trauma disorders. An iatrogenic concept. *Journal of Occupational Medicine: Official Publication of the Industrial Medical Association*, 32(1), 38–41.
- Hadler, N. M. (1992). Arm pain in the workplace. A small area analysis. *Journal of Occupational Medicine: Official Publication of the Industrial Medical Association*, 34(2), 113–9.
- Hadler, N. M. (1993). Arm pain in the work place. *Bulletin on the Rheumatic Diseases*, 42(8), 6–8.
- Hadler, N. M. (1997). Repetitive upper-extremity motions in the workplace are not hazardous. *The Journal of Hand Surgery*, 22(1), 19–29.
- Hancock, P. A. (1987). *Human factors psychology*. New York: Elsevier Science Publishing Co.
- Hancock, P. A. (1999). *Human performance and ergonomics*. San Diego, Calif, London: Academic.
- Hancock, P. A., & Caird, J. K. (1993). Experimental evaluation of a model of mental workload. *Human Factors*, 35(3), 413–29.
- Hancock, J. C., & Wintz, P. A. (1966). *Signal detection theory*. New York: McGraw-Hill.
- Hancock, P. A., Wulf, G., Thom, D., & Fassnacht, P. (1990). Driver workload during differing driving maneuvers. *Accident; Analysis and Prevention*, 22(3), 281–90.
- Heffer, H., Hömberg, V., Reiners, K., & Freund, H. J. (1987). Stability of frequency during long-term recordings of hand tremor. *Electroencephalography and Clinical Neurophysiology*, 67(5), 439–46.
- Helstrom, C. W. (1960). *Statistical theory of signal detection*. New York: Pergamon Press.

- Hemborg, B., Moritz, U., Hamberg, J., Löwing, H., & Akesson, I. (1983). Intraabdominal pressure and trunk muscle activity during lifting-effect of abdominal muscle training in healthy subjects. *Scandinavian Journal of Rehabilitation Medicine*, 15(4), 183–96.
- Hockey, G. R., Briner, R. B., Tattersall, A. J., & Wiethoff, M. (1989). Assessing the impact of computer workload on operator stress: The role of system controllability. *Ergonomics*, 32(11), 1401–18.
- Hockey, G. R., & Sauer, J. (1996). Cognitive fatigue and complex decision making under prolonged isolation and confinement. *Advances in Space Biology and Medicine*, 5, 309–30.
- Horrey, W. J., & Simons, D. J. (2007). Examining cognitive interference and adaptive safety behaviours in tactical vehicle control. *Ergonomics*, 50(8), 1340–50.
- Inamasu, J., & Guiot, B. H. (2007). Thoracolumbar junction injuries after motor vehicle collision: Are there differences in restrained and nonrestrained front seat occupants? *Journal of Neurosurgery Spine*, 7(3), 311–4.
- Inamasu, J., & Guiot, B. H. (2009). Thoracolumbar junction injuries after rollover crashes: Difference between belted and unbelted front seat occupants. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, 18(10), 1464–8.
- Kibler, W. B., Chandler, T. J., & Stracener, E. S. (1992). Musculoskeletal adaptations and injuries due to overtraining. *Exercise and Sport Sciences Reviews*, 20, 99–126.
- Kingma, I., Faber, G. S., Suwarganda, E. K., Bruijnen, T. B., Peters, R. J., & van Dieen, J. H. (2006). Effect of a stiff lifting belt on spine compression during lifting. *Spine*, 31(22), E833–9.
- Kuipers, H. (1998). Training and overtraining: An introduction. *Medicine and Science in Sports and Exercise*, 30(7), 1137–9.
- Lamberts, K., & Goldstone, R. L. (2005). *Handbook of cognition*. Thousand Oaks, CA, London: SAGE.
- Lavender, S. A., Li, Y. C., Andersson, G. B., & Natarajan, R. N. (1999). The effects of lifting speed on the peak external forward bending, lateral bending, and twisting spine moments. *Ergonomics*, 42(1), 111–25.
- Lee, Y. H., & Chen, Y. L. (2000). Regressionally determined vertebral inclination angles of the lumbar spine in static lifts. *Clinical Biomechanics (Bristol, Avon)*, 15(9), 672–7.
- Lee, P. J., & Granata, K. P. (2006). Interface stability influences torso muscle recruitment and spinal load during pushing tasks. *Ergonomics*, 49(3), 235–48.
- Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Hum Factors*, 49(3), 376–99.
- Lord, M., Reynolds, D. P., & Hughes, J. R. (1986). Foot pressure measurement: A review of clinical findings. *Journal of Biomedical Engineering*, 8(4), 283–94.
- Luginbuhl, R. C., Jackson, L. L., Castillo, D. N., & Loring, K. A. (2008). Heat-related deaths among crop workers – United States, 1992–2006. *Morbidity and Mortality Weekly Report*, 57(24), 649–653.
- Madden, J. P. L. (1989). *Abdominal wall hernias*. Philadelphia: Saunders.
- Mairiaux, P., Davis, P. R., Stubbs, D. A., & Baty, D. (1984). Relation between intra-abdominal pressure and lumbar moments when lifting weights in the erect posture. *Ergonomics*, 27(8), 883–94.
- Malchaire, J. B. (1994). *Heat stress evaluation*. Boca Raton: Lewis Publishers.
- Manzey, D., Lorenz, B., & Poljakov, V. (1998). Mental performance in extreme environments: Results from a performance monitoring study during a 438-day spaceflight. *Ergonomics*, 41(4), 537–59.
- Marras, W. S., Granta, K. P., & Davis, K. G. (1999). Variability in spine loading model performance. *Clinical Biomechanics (Bristol, Avon)*, 14(8), 505–14.
- Marras, W. S., Lavender, S. A., Leurgans, S. E., Fathallah, F. A., Ferguson, S. A., Allread, W. G., et al. (1995). Biomechanical risk factors for occupationally related low back disorders. *Ergonomics*, 38(2), 377–410.
- Marras, W. S., & Mirka, G. A. (1990). Muscle activities during asymmetric trunk angular accelerations. *Journal of Orthopaedic Research*, 8(6), 824–32.
- Mathiassen, S. E. (1993). The influence of exercise/rest schedule on the physiological and psychophysical response to isometric shoulder-neck exercise. *European Journal of Applied Physiology and Occupational Physiology*, 67(6), 528–39.
- Maxfield, M. E., & Brouha, L. (1963). Validity of heart rate as an indicator of cardiac strain. *Journal of Applied Physiology*, 18, 1099–104.
- McNicol, D. (1972). *A primer of signal detection theory*. London: Allen and Unwin.
- Menant, J. C., Steele, J. R., Menz, H. B., Munro, B. J., & Lord, S. R. (2008). Optimizing footwear for older people at risk of falls. *Journal of Rehabilitation Research and Development*, 45(8), 1167–81.
- Mirka, G. A., & Marras, W. S. (1993). A stochastic model of trunk muscle coactivation during trunk bending. *Spine*, 18(11), 1396–409.
- Monod, H., & Garcin, M. (1996). Use of physiological criteria for improving physical work conditions. *Journal of Human Ergology*, 25(1), 29–38.
- Morris, J. N., Heady, J. A., Raffle, P. A., Roberts, C. G., & Parks, J. W. (1953). Coronary heart-disease and physical activity of work. *Lancet*, 265(6795), 1053.

- Mowbray, H. M., & Gebhard, J. W. (1958). *Man's senses as informational channels*. Silver Spring, MD: The Johns Hopkins University, Applied Physics Laboratory.
- National Center for Health Statistics (NCHS), National Vital Statistics System (2007).
- Niebel, B. W., & Freivalds, A. (2002). *Methods, standards and work design*. New York: McGraw-Hill, Inc.
- NIOSH (1981). *Work practices guide for manual lifting*. Washington, D.C.: GPO.
- Nordhoff, L. S. (2005). *Motor vehicle collision injuries: Biomechanics, diagnosis, and management*. Sudbury, MA: Jones and Bartlett.
- Norman, D. A. (1988). *The psychology of everyday things*. New York: Basic Books.
- O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, C. Kauffmann, & J. Thomas (Eds.), *Handbook of perception and human performance* (pp. 42/1–42/49). New York: Wiley & Sons Inc.
- Ozkaya, N., Willems, B., & Goldsheyder, D. (1994). Whole-Body vibration exposure: A comprehensive field study. *American Industrial Hygiene Association Journal*, 55(12), 1164–71.
- Paffenbarger, R. S., Jr., Laughlin, M. E., Gima, A. S., & Black, R. A. (1970). Work activity of longshoremen as related to death from coronary heart disease and stroke. *The New England Journal of Medicine*, 282(20), 1109–14.
- Plagenhoef, S., Curtis, D., & Musante, L. (1971). *Patterns of human motion. A cinematographic analysis*. Upper Saddle River, NJ, USA: Prentice-Hall.
- Poor, H. V. (1994). *An introduction to signal detection and estimation*. New York: Springer.
- Pope, M. H., Goh, K. L., & Magnusson, M. L. (2002). Spine ergonomics. *Annual Review of Biomedical Engineering*, 4, 49–68.
- Pope, M. H., Wilder, D. G., & Magnusson, M. L. (1999). A review of studies on seated whole body vibration and low back pain. Proceedings of the institution of mechanical engineers. Part H. *Journal of Engineering in Medicine*, 213(6), 435–46.
- Psychonomic Society. (Ed) (1966). *Perception & psychophysics*. Austin: Psychonomic Society.
- Purvis, D., Gonsalves, S., & Deuster, P. A. (2010). Physiological and psychological fatigue in extreme conditions: Overtraining and elite athletes. *PM & R: The Journal of Injury, Function, and Rehabilitation*, 2(5), 442–50.
- Putz-Anderson, V. (1988). *Cumulative trauma disorders: A manual for musculoskeletal diseases of the upper limbs*. New York: Taylor and Francis.
- Redfern, M. S., Cham, R., Gielo-Perczak, K., Gronqvist, R., Hirvonen, M., Lanshammar, H., et al. (2001). Biomechanics of slips. *Ergonomics*, 44(13), 1138–66.
- Rohmert, W. (1973a). Problems in determining rest allowances part 1: Use of modern methods to evaluate stress and strain in static muscular work. *Applied Ergonomics*, 4(2), 91–5.
- Rohmert, W. (1973b). Problems of determination of rest allowances part 2: Determining rest allowances in different human tasks. *Applied Ergonomics*, 4(3), 158–62.
- Schust, M., Kreisel, A., Seidel, H., & Blüthner, R. (2010). Examination of the frequency-weighting curve for accelerations measured on the seat and at the surface supporting the feet during horizontal whole-body vibrations in x- and y-directions. *Industrial Health*, 48(5), 725–42.
- Seidel, H. (2005). On the relationship between whole-body vibration exposure and spinal health risk. *Industrial Health*, 43(3), 361–77.
- Seidler, A., Bolm-Audorff, U., Siol, T., Henkel, N., Fuchs, C., Schug, H., et al. (2003). Occupational risk factors for symptomatic lumbar disc herniation; a case-control study. *Occupational and Environmental Medicine*, 60(11), 821–30.
- Shanks, D. R. (1997). *Human memory: A reader*. New York: St. Martin's Press.
- Shoenberger, R. W. (1979). Psychophysical assessment of angular vibration: Comparison of vertical and roll vibrations. *Aviation, Space, and Environmental Medicine*, 50(7), 688–91.
- Simpson, M. R., & Howard, T. M. (2009). Tendinopathies of the foot and ankle. *American Family Physician*, 80(10), 1107–14.
- Smets, M. P., Eger, T. R., & Grenier, S. G. (2010). Whole-Body vibration experienced by haulage truck operators in surface mining operations: A comparison of various analysis methods utilized in the prediction of health risks. *Applied Ergonomics*, 41(6), 763–70.
- Smith, S. D. (2006). Seat vibration in military propeller aircraft: Characterization, exposure assessment, and mitigation. *Aviation, Space, and Environmental Medicine*, 77(1), 32–40.
- Smith, A. B., Dickerman, R. D., McGuire, C. S., East, J. W., McConathy, W. J., & Pearson, H. F. (1999). Pressure-overload-induced sliding hiatal hernia in power athletes. *Journal of Clinical Gastroenterology*, 28(4), 352–4.
- Smith, J. A., Siegel, J. H., & Siddiqi, S. Q. (2005). Spine and spinal cord injury in motor vehicle crashes: A function of change in velocity and energy dissipation on impact with respect to the direction of crash. *The Journal of Trauma*, 59(1), 117–31.
- Stemper, B. D., & Storvik, S. G. (2010). Incorporation of lower neck shear forces to predict facet joint injury risk in low-speed automotive rear impacts. *Traffic Injury Prevention*, 11(3), 300–8.
- Stevens, S. S. (1951). *Handbook of experimental psychology*. New York: Wiley.

- Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64(3), 153–81.
- Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. New York: Wiley.
- Stokes, A., Wickens, C. D., & Kite, K. (1990). *Display technology: Human factors concepts*. Warrendale, PA: Society of Automotive Engineers.
- Stone, M. H. (1990). Muscle conditioning and muscle injuries. *Medicine and Science in Sports and Exercise*, 22(4), 457–62.
- Stubbs, D. A. (1981). Trunk stresses in construction and other industrial workers. *Spine*, 6(1), 83–9.
- Stubbs, D. A. (1985). Human constraints on manual working capacity: Effects of age on intratruncal pressure. *Ergonomics*, 28(1), 107–14.
- Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: L. Erlbaum Associates.
- Teepie, E., Shalvoy, R. M., & Feller, E. R. (2006). Overtraining in young athletes. *Medicine and Health, Rhode Island*, 89(7), 236–8.
- Tisserand, M. (1985). Progress in the prevention of falls caused by slipping. *Ergonomics*, 28(7), 1027–42.
- Troup, J. D. (1978). Driver's back pain and its prevention. A review of the postural, vibratory and muscular factors, together with the problem of transmitted road-shock. *Applied Ergonomics*, 9(4), 207–14.
- Van Cott, H. P., & Warrick, M. J. (1972). *Man as a system component*. Washington, DC: Government Printing Office.
- Veltman, J. A., & Gaillard, A. W. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5), 656–69.
- Veres, S. P., Robertson, P. A., & Broom, N. D. (2010). The influence of torsion on disc herniation when combined with flexion. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, 19(9), 1468–78.
- Vetter, R. E., & Symonds, M. L. (2010). Correlations between injury, training intensity, and physical and mental exhaustion among college athletes. *Journal of Strength and Conditioning Research/National Strength and Conditioning Association*, 24(3), 587–96.
- Wagner, J. H. (2011). *Hernias*. Hauppauge, NY: Nova Science.
- Welford, A. T. (1968). *Fundamentals of skill*. London: Methuen.
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford; New York: Oxford University Press.
- Wickens, C. D., Gordon, S. E., & Liu, Y. (2004). *An introduction to human factors engineering*. Pearson Prentice Hall.
- Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice Hall.
- Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. *Hum Factors*, 21(5), 575–93.
- Wiker, S. F. (1990). Shoulder postural fatigue and discomfort: No relationship with isometric strength capability in a light-weight manual assembly task. *International Journal of Industrial Ergonomics*, 5, 133–46.
- Wiker, S. F., Chaffin, D. B., & Langolf, G. D. (1989). Shoulder posture and localized muscle fatigue and discomfort. *Ergonomics*, 32(2), 211–37.
- Wiker, S. F., Langolf, G. D., & Chaffin, D. B. (1989). Arm posture and human movement capability. *Human Factors*, 31(4), 421–41.
- Wiker, S. F., & Miller, J. M. (1983). Acceleration exposures in forward seating areas of bowrider recreational boats. *Human Factors*, 25(3), 319–27.
- Wilber, C. A., Holland, G. J., Madison, R. E., & Loy, S. F. (1995). An epidemiological analysis of overuse injuries among recreational cyclists. *International Journal of Sports Medicine*, 16(3), 201–6.
- Winter, D. A. (2009). *Biomechanics and motor control of human movement*. New Jersey: Wiley.
- Wise, J. A., Hopkin, V. D., & Garland, D. J. (2010). *Handbook of aviation human factors*. Boca Raton: CRC Press.
- Yost, W. A., Popper, A. N., & Fay, R. R. (1993). *Human psychophysics*. New York: Springer.
- Young, R. R., & Hagbarth, K. E. (1980). Physiological tremor enhanced by manoeuvres affecting the segmental stretch reflex. *Journal of Neurology, Neurosurgery, and Psychiatry*, 43(3), 248–56.

Chapter 8

Experimental Methods

Jonathan Howland and Damaris J. Rohsenow

Introduction

This chapter uses examples of the authors' research to illustrate the use of experimental methods for testing hypotheses relevant to the evaluation of injury prevention/control intervention and injury causation. It briefly reviews characteristics of some common experimental designs and provides examples of measurement strategies for objective and subjective assessment of injury-related performance. In drawing on our own research we have necessarily omitted much material relevant to this discussion. Such material can be found not only in the injury literature but also in the literatures of other disciplines, such as psychology, human factors, space and aviation medicine, operations research, and transportation research. The aim of this chapter is to raise awareness of the possibilities of experimentation among injury prevention investigators and practitioners whose research has been primarily limited to observational study designs.

Study Designs

A variety of experimental designs have been used to study the effectiveness of interventions for preventing or reducing injuries (clinical trials, treatment trials) and to investigate hypotheses about injury causation (assessment studies, epidemiological studies, experimental laboratory studies). Designs for both treatment and laboratory studies include between-groups randomized controlled trials (RCTs), within-subjects randomized trials, mixed between-groups–within-subjects trials, cluster randomized trials, and quasi-experiments. The following reviews and provides examples for several of these experimental designs.

J. Howland, PhD, MPH, MPA (✉)

Department of Emergency Medicine, Boston Medical Center, Boston University School of Medicine,
One Boston Medical Center Place, Dowling 1 South, Boston, MA 02118, USA
e-mail: jhowl@bu.edu

D.J. Rohsenow, PhD

Center for Alcohol and Addiction Studies, Brown University, Box G-S121-5, Providence, RI 02912, USA
e-mail: damaris_rohsenow@brown.edu

Between-Groups Trials

In a between-groups design, individuals are randomly assigned to two or more experimental conditions, typically intervention or control. All groups are measured on outcomes (dependent variables) at baseline (pretest) and at one or more follow-up points (posttests). The intervention is delivered between pre- and posttest to the experimental group, while the control group receives no intervention, or a placebo or different intervention (attention control) that has no relevance to the outcome of interest, or (in treatment trials) an active treatment of established efficacy that is routinely used. Assuming that randomization has been effective in creating groups that are comparable on dependent variables and other characteristics (potential confounders or moderators) when measured at pretest, posttest measures can be compared for statistical significance. If pretest differences exist, these can be controlled statistically using covariance analysis as long as these are not inherently different between the two groups being studied (e.g., since depressed and nondepressed people inherently differ in anxiety levels, it is not appropriate to covary anxiety levels [Miller and Chapman 2001]).

The nature of the control group used is of crucial importance so that any difference between groups can be attributed only to the intervention being tested. In treatment trials, it is usually important that the two treatments be matched for amount of time and number of sessions, and credibility of the control treatment; otherwise, results could be due to differences in patients' expectancies rather than due to the differences in content of the treatments (a serious problem with having a no-intervention control or waiting list control group). Common valid control groups in treatment trials include treatment as usual, minimal treatment, standard active treatment, placebo treatment (credible but not containing the active ingredients needed), or a dismantled version of the experimental treatment. The exception in matching amount of time is when the purpose is to compare a brief intervention to a standard length intervention, to compare a considerably enhanced treatment to a usual treatment (as in the example below), or when usual treatment is only minimal, such as brief advice and/or a handout listing treatment agencies (e.g., Monti et al. 1999). Implementation fidelity procedures must be put into effect to ensure that the treatments are delivered in a specific manner and do not "drift" over the course of the study, such as by using manuals and reviewing recordings of a sample of sessions for adherence to the manual. The researcher's own staff should deliver any treatment-as-usual since there is little control over agency staff in how they deliver their treatment.

Control groups for experimental lab studies also need careful design to avoid confounds. For example, many studies of alcohol administration have used placebo beverages that participants can easily tell do not contain alcohol (see review by Rohsenow and Marlatt 1981). The only credible alcohol placebos found have been a 1:5 ratio of vodka to chilled tonic vs. tonic, or beer vs. nonalcoholic beer. Order effects must be accounted for by counterbalancing the order of the experimental vs. control condition across participants and including order in the statistical analyses. The exceptions are (1) when a condition is being compared to baseline values that inherently need to come before the condition or (2) when carryover effects are predicted, such as when alcoholics are asked to hold and look at a glass of alcoholic beverage as opposed to a glass of water, and their reactions persist long after the stimulus is removed (e.g., Rohsenow et al. 2000). Time of day and day of week need thought since, for example, giving alcohol in the morning vs. evening can affect outcomes due to people learning to compensate for when they usually drink.

Example

Spirito et al. (2011) conducted a randomized controlled between-groups trial in an emergency department (ED) of an ED-based brief alcohol intervention for adolescent patients who presented at a level 1 trauma center with injury following an alcohol-related event. The authors hypothesized that brief individual motivational interview (IMI) plus family motivational interview (family checkup [FCU]) would reduce alcohol use and alcohol-related problems among these adolescents more effectively than an IMI only. Adolescents, 13–17 years old ($N=125$), were eligible if they had a positive blood or breath alcohol at admission. Two-group randomized design with three follow-up points was used. Parents had to agree to participate without knowing treatment assignment in advance so that groups would not differ on parental motivation. Both conditions resulted in a reduction in all drinking outcomes at all follow-ups (all p 's < 0.001) with strongest effects at 3 and 6 months. Adding the FCU to the IMI resulted in better outcome than IMI alone on high volume drinking days at 3-month follow-up, 14.6% vs. 32.1%, $p < 0.05$, OR 2.79, 95% CI, 0.99–7.75. Investigators concluded that motivational interventions have a positive effect on drinking outcomes in the short term following an alcohol-related ED visit for adolescent drinkers; and, adding the FCU to an IMI resulted in somewhat better effects on high volume drinking at short-term follow-up than an IMI alone, but the considerably higher cost of FCU needs to be weighed against the modest improvements in outcome.

Crossover Trials

In a randomized within-subjects crossover design, participants are compared to themselves on an outcome following exposure to two or more experimental conditions or an experimental vs. control condition. An advantage of this design is that variance is reduced, relative to a between-groups design, since participant characteristics are held constant and only the experimental conditions are changed. This enhances statistical power resulting in the need for fewer participants. The order of experimental condition is counterbalanced such that participants randomly receive the treatment or the control condition (placebo) at the first experimental session and receive the alternate experimental condition at the second session. Crossover designs can only be used when the effects of the treatment or experimental condition are transient. Thus, this approach is not appropriate for interventions that involve changing knowledge, attitudes, behaviors, or proficiencies in injury-related performance. The disadvantages include fewer people being willing to complete the additional sessions (especially in a medication trial involving a regimen lasting for a number of weeks or when each condition requires many visits) and difficulty in maintaining participant blinding to their experimental condition. For example, since government-supplied marijuana and placebo marijuana differ in appearance, a crossover design may be a poor choice in studying marijuana's effects on risk taking.

Example

Transdermal scopolamine is commonly used by mariners to prevent or treat seasickness. Howland et al. (2008) conducted a study to test the effects of the transdermal scopolamine patch on merchant mariners' navigation and collision avoidance skills using a maritime training simulator to assess performance under simulated rough-weather conditions. A randomized

(continued)

Example (continued)

double-blind crossover study assessed 32 Swedish maritime cadets under transdermal scopolamine and placebo conditions on simulated navigation and ship handling performance, sleepiness, and subjective measures of fitness and performance. There were no significant differences on occupational outcomes by medication condition, but sustained reaction time was significantly increased under transdermal scopolamine relative to placebo. The investigators concluded that the transdermal scopolamine does not impair simulated ship handling although it could impair other abilities relevant to safety.

Mixed Between-Groups/Within-Subjects Trials

There are times when it is desirable to test two types or levels of exposure, in which case a between-groups design is embedded within a within-subjects design. This can be a way to reduce attrition due to excessive length of time in the study for a fully within-subjects design. For example, if one wanted to test neurocognitive decrements at two dosages of a common medication (e.g., a sedating cold medication) as compared to a placebo, one could assign people to placebo, low or high dose (called dose A or B) in a three-group between-subjects' design. However, this requires about 50% more subjects than even a fully between-subjects 2×2 design due to differences in power (Cohen 1988). However, a stronger 2×2 design is one where you randomize people to dose A or B (the between-groups component) but all participants would be compared to themselves in both dosed and placebo conditions (the within-subjects component). In analysis, the effects of the medication regardless of dose level, the differential effect of each dosage vs. placebo, and the effect of each dose compared to each other can all be calculated and tested.

Example

Most alcoholic beverages contain small amounts of chemicals other than ethanol as a by-product of the materials used in the fermenting process (e.g., grains and wood casks). These congeners are complex organic molecules with toxic effects including acetone, acetaldehyde, fusel oil, tannins, and furfural, with bourbon having 37 times the amount of congeners as vodka. Rohsenow et al. (2010) studied the effects of heavy drinking with high and low congener beverages on next-day neurocognitive performance to determine whether, holding blood alcohol constant, congener content correlated with neurocognitive functions that could effect risk for injury. Young adult heavy drinkers ($N=93$) received bourbon or vodka mixed with caffeine-free cola (to help disguise the color and taste) sufficient to raise their breath alcohol level to a mean of 0.11 g%, with random assignment to bourbon or vodka one night, and with matched placebo (the cola with a little bourbon or vodka floated on top for taste) another night, 1 week later. The type of alcoholic beverage and order were randomized. After an 8-h opportunity to sleep (under observed laboratory conditions), self-report and neurocognitive measures were assessed the next morning. After alcohol of either type, significant performance deficits were evident in attention/reaction time. No effect of beverage congeners was found except on hangover severity, with participants feeling worse after bourbon. This finding suggests that residual alcohol effects on next-day neurocognitive function are not mitigated by low congener content alcohol beverages.

Cluster Randomized Trials

Depending on the nature of the intervention, randomized trials that randomize at the level of the individual cannot be used because contamination, the inadvertent transference of the intervention from treatment to control group, threatens the study's validity. For example, individual-level randomization in a school-, hospital-, or workplace-based intervention trial that involves changes in knowledge, attitudes, and behaviors would be compromised because the content of the intervention could, through communication or observation, be transmitted from the intervention group to the control group. This secondary exposure of the control group thus exposes both groups to the intervention potentially masking real intervention effects. In this circumstance, a cluster randomized trial can be used such that participants are randomized at the level of the group (e.g., schools), but because many groups are involved, the random assignment tends to eliminate baseline differences when individuals are aggregated in their experimental groups. Statistical analyses that account for nested effects may be needed unless a large number of different settings are involved.

Example

Tennstedt et al. (1998) used a cluster randomized design to assess the effectiveness of a community-based group intervention designed to reduce fear of falling (a risk factor for falls) and increase mobility and psychosocial status among older adults. A sample of 434 persons aged 60+ years, who reported fear of falling and associated activity restriction, was recruited from senior housing sites in the Boston Metropolitan Area. The unit of randomization was the senior housing site. Forty housing sites participated and pair matched on the basis of the number of dwelling units and percent minority residents, with one site in each pair randomly assigned to the intervention group and the other to a placebo attention control group. Data were collected at baseline, and at 6-week, 6-, and 12-month follow-ups; analysis was conducted on the basis of intention-to-treat. Compared with contact control subjects, intervention subjects reported increased levels of intended activity ($p < .05$) and greater mobility ($p < .05$) immediately after the intervention. Effects at 12 months included improved social function ($p < .05$) and mobility range ($p < .05$). The intervention had immediate but modest beneficial effects that diminished over time in the setting with no booster intervention. Investigators concluded that fear of falling can be reduced among older adults, with benefits for mobility and psychosocial status.

Quasi-experiments

Quasi-experiments can be used in circumstances when resources required for a cluster randomized trial are unavailable or, as in the case of community-based interventions, are not practical. In these cases a small number of population units is used. In some studies, experimental condition is not assigned at random, such as when certain communities choose to adopt a program and they are compared to non-adopters. When allocation to experimental status is random (such as certain hospitals receiving the training) there are not enough units to ensure comparability of populations on baseline characteristics. In these cases, studies mitigate effects of group nonequivalence by using nested statistical analysis that can control for known initial group differences. Because experimental conditions cannot be truly

equivalent in this design, unknown and therefore unmeasured differences cannot be accounted for, thus making the results of quasi-experiments less compelling than those of randomized trials.

Example

Hingson et al. (1996) conducted a quasi-experimental study to assess whether a community program that organized multiple city departments and private citizens could reduce alcohol-impaired driving, related driving risks, and traffic deaths and injuries. Trends in fatal crashes and injuries per 100 crashes were compared in six Saving Lives Program cities to the rest of Massachusetts. In annual roadside surveys conducted at randomly selected locations, safety belt use among occupants of 54,577 vehicles and travel speeds of 118,442 vehicles were observed. Four statewide telephone surveys ($n = 15,188$) monitored self-reported driving after drinking. In program cities relative to the rest of Massachusetts during the 5 program years in comparison with the previous 5 years, fatal crashes declined 25%, from 178 to 120, and fatal crashes involving alcohol decreased 42%, from 69 to 36. Visible injuries per 100 crashes declined 5%, from 21.1 to 16.6. The proportions of vehicles observed speeding and teenagers who drove after drinking were cut in half. Investigators concluded that community-based interventions can reduce local traffic injuries and fatalities.

Measurement

Experimental methods can be used to identify and quantify risk factors for injury. This is possible only when exposure to the hypothesized risk factor can be administered with relative safety and the outcome can likewise be measured in a way that does not place participants at risk. Examples of exposures include alcohol and other drugs and medications or induced conditions such as fatigue that can be administered in controlled laboratory settings. Depending on the nature of the exposure, the between-groups and within-subjects study designs described above can be used to experimentally test hypotheses raised by observational studies or survey research.

Simulators for Measuring Safety-Related Performance

Decrements in the performance of certain activities of daily living (e.g., driving) or occupational tasks (e.g., piloting a commercial aircraft or ship) can place individuals, clients (e.g., airline passengers), or coworkers at risk for injury. Simulators developed for training or assessing proficiency can provide safe and practical means for assessing the impacts of various exposures on safety-related performance. As opposed to psychological tests (discussed below) that break performance into discrete neurocognitive functions (e.g., attention/reaction time), simulation creates environments that approximate actual operation conditions. A simulator is a device that generates simulation (National Research Council) and a scenario is a simulation that represents a specific set of circumstances. It is likely that simulation will be used increasingly in experimental research due to its increased availability and decreased cost. Technology rapidly increases the capacity and sophistication of simulators, while at the same time reducing costs; globalization increases industry's demand for productivity and thus demand for effective and efficient training methods. Perhaps simulation's most significant advantage for research is safety. Simulation allows participants to perform under a range of circumstances, from routine to emergent, without risk of injury or property loss. When measured under actual operating conditions, performance assessment must stop when risk becomes unacceptable. This may be at a point well before a given scenario becomes catastrophic, thus leaving an experimental blind spot at

precisely the point where it is important to assess performance. Simulation, however, has no ceiling with regard to how far a scenario can go in terms of risk. As a consequence, the experimental participant has full command of the situation, without the presence of a backup person prepared to take over if real-world circumstances become too dangerous.

Simulators are often equipped to quantify and record performance data in real time along different dimensions of performance. Thus, objective measures are produced in a form that lends itself to quantitative analysis. Indeed, this attribute is one that differentiates a simulator from a computer game. This is important because self-reports of performance may be affected by the exposure or by recall bias. Moreover, some exposure situations may partially impair cognition, such that the task can be performed overall but only at the expense of attention to potentially critical subtasks, which might themselves generate risk for injury or operation failure, as was found in a study on residual alcohol effects with aircraft simulators (Yesavage and Leirer 1986).

Simulation is also useful in a research context in that it allows replication of experiments. This can yield confirmation of findings. It can also yield precise comparisons of the risk of different types of exposure (e.g., alcohol vs. fatigue).

Some types of exposure, or certain dose levels of substances, may not impair performance of well-learned routine tasks but may impair performance under high-risk rare situations. These situations are by definition unlikely to occur in real-world training environments. Simulation, however, can generate rare events and measure performance across a range of possible scenarios, as is described in an example below. Similarly, simulation can generate situations that require multitasking, wherein performance of a routine task is burdened by the need to respond to distracting stimuli, the intensity, and frequency of which can be varied.

It should be noted that while simulation may be generated by devices that create a totally artificial environment (e.g., flight training simulators), it can also be generated by an ordinary workplace device (e.g. word processor) that is equipped to measure performance or staged with props and “actors” as, for example, when a disaster is enacted to train first responders.

The advantage of simulation is that it allows for measurement that in the real world might be dangerous or too intrusive to be practical. The disadvantage is that no matter how sophisticated the simulation, much of what affects performance in the real world is lost. As in most experimental situations, participants in simulators are usually aware that they are being observed and this awareness may provide motivation to perform beyond what would occur under nonexperimental circumstances (testing effects). Alternatively, without real-world consequences, simulation may reduce performance, although our experience with simulation suggests that on the whole participants are highly motivated to perform well.

Simulation is high in face validity, the superficial appearance to test-takers and others that a measure measures what it is supposed to measure, a quality important for the acceptability of the results (Anastasi 1988). As opposed to neurocognitive measures, which assess components of real-world tasks, simulation attempts to measure performance on the whole task. Face validity can be important for influencing the public or policy makers about the relationship between an exposure and performance effects. Face validity is not a true form of validity, therefore internal validity, the extent to which a measure actually measures the concept it is supposed to measure, is more important. A simulation is internally valid if it assesses performance as that performance would occur in the real world. To some degree, internal validity is a function of the simulator’s technical sophistication. For example, the difference between a Federal Aviation Agency certified flight simulator and a computer game that mimics flying is a function of the degree to which the behavior of the plane is mathematically modeled to respond to pilot behaviors as it would in the real world; time in a certified flight simulator is considered equivalent to time in the actual plane whereas time with a computer game time is not. Many simulations have not been validated for the simple reason that the validation process could expose persons and property to harm. Moreover, even if validation could be performed, simulators are usually developed for training and assessment purposes and not necessarily for experimental measurement. Thus, for many simulators, high face validity is sufficient.

Fig. 8.1 Current federal regulations for alcohol use

CURRENT FEDERAL REGULATIONS FOR ALCOHOL USE

	Aircraft	Truck	Merchant Vessel	Train	Nuclear Power Plant
BAC Legal Limit	.04 g% BAC	Zero BAC (defined as .02 g%)	.04 g% BAC	.04 g% BAC	.04 g% BAC
Pre-Duty Alcohol Free Period	8 Hours	4 Hours	4 Hours	4 Hours	5 Hours

The authors have used simulation to test the value of federal rules that govern the use of alcohol in conjunction with regulated safety-sensitive jobs. These regulations specify a per se intoxication level, typically 0.04 g%, above which on-duty personnel can be sanctioned and the “bottle-to-throttle” period, the time (which varies by occupation) before duty during which alcohol cannot be consumed (see Fig. 8.1). In separate studies, we tested each of these regulatory dimensions.

Examples

Using a between-groups study design, Howland et al. (2000) examined the acute effects of low-level alcohol on simulated operation of a commercial vessel’s power plant (main engines and electrical generating systems). Participants were third and fourth year maritime engineer cadets ($N=19$). This study used a randomized between-groups design in which alcohol administration (vodka and tonic vs. tonic only) was fully crossed with the expectancy that alcohol was administered (told receiving alcohol vs. told receiving tonic only). The target BAC was 0.04 g%, the legal limit for operation of a commercial vessel under current federal law. The simulator used was a NorControl diesel engine simulator (Kongsberg NorControl Simulation AS, Bekkajordet 8 A, P.O. Box 1039, N-3194 Horten, Norway). Two scenarios were developed involving initial normal operation, development of an operating problem, and system component failure alarm. The participant was required to identify the problem, mitigate the problem, and initiate an engine reignition procedure. Performance was measured in terms of time (seconds) between the alarm sounding and the engine restarting. The order of alcohol-placebo administration was balanced across participants, as was the order of the two scenarios. On Day 1, participants received placebo (tonic) and both performed on the simulator 1/2 h after receiving beverage. On Day 2, a week later, one half were told they were getting alcohol and the other only tonic, then half of each of these expectancy groups received vodka and tonic in sufficient quantity (determined by weight and sex of the participant) to raise BAC to 0.04 g% while the rest received placebo, tonic alone. Then, both groups again performed on the simulator 1/2 h after receiving beverage. Under the placebo–placebo condition, the time increased relative to the baseline day by 5%, while under the placebo–alcohol condition, the time increased by 200% ($p < .05$), with a large effect size (Cohen’s $f=0.71$; Cohen 1988). Neither expectancy nor the interaction of beverage by expectancy was significant. Investigators concluded that low-dose alcohol exposure can cause risk for decrements in occupational performance (Fig. 8.2).

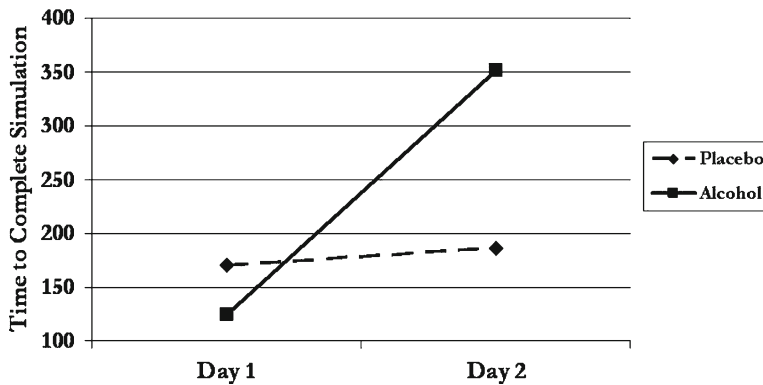


Fig. 8.2 Mean score for diesel simulator performance by alcohol exposure (0.04 g% BAC). $N=18$ cadets, $p<0.05$, effect size ($f=0.71$)

Example

To test whether an 8-h “bottle-to-throttle” period was sufficient to eliminate performance decrements the morning after intoxication at 0.11–0.12 g%, Rohsenow et al. (2006) enrolled maritime engineer cadets at the same academy, but different from those who participated in the low-BAC studies. To measure performance, they used the diesel simulator and scenarios comparable to those described above in the acute alcohol study. A between-groups design was used; 61 cadets were given placebo (nonalcoholic) beer the first evening and randomized to receive placebo or real beer the second evening. The mean BAC was 0.115 g% after alcohol administration. The next morning, after testing at zero BAC, cadets were assessed with self-reports and a power plant simulator that they were already trained to use. Beverage condition did not affect power plant simulator performance (effect size near zero) although cadets who consumed alcohol rated their performance more impaired than those in the placebo condition ($p<0.02$, medium effect $f=0.31$). Investigators concluded that this level of drinking therefore might not impair next-day routine occupational performance by trained young ship engineers despite subjective ill effects.

Measuring Neurocognitive Performance

In contrast to simulation that assesses performance on whole tasks, neurocognitive tests can measure discrete neurobehavioral functions that may have implications for risk of injury to self or others. Relative to simulation, these tests may have known psychometric properties, including validity and reliability, and may have established norms based on large population samples or subsamples (e.g., age and/or gender cohorts). Moreover, neurocognitive tests can potentially explain performance decrements by revealing specific functional impairments. A number of computer- and web-based

neurocognitive batteries are available (see Kelly et al. 2007 for review). The authors have used two in studies of the residual effects of alcohol intoxication on next-day performance: Neurobehavioral Evaluation System 3 (NES3) and the Psychomotor Vigilance Task (PVT).

The NES3 (White et al. 2003) was designed to screen for nervous system damage due to toxic environmental agents. Thus, it is relevant to the measurement of the effects of alcohol and other psychoactive drugs. The specific battery of tests, many of which are computer-based adaptations of preexisting clinical instruments, was derived by a consensus panel convened by the World Health Organization (WHO) and the National Institute for Occupational Safety and Health (NIOSH). Twenty individual tests assess functioning in the following categories: motor ability; focused attention; selective attention; acquisition; memory; and other cognitive performance, including pattern matching, grammatical/logical reasoning, arithmetic computation, and vocabulary. The NES3 is administered using a desktop computer with a touch screen monitor. The computer provides verbal instructions (via headphones) and responses are made using the touch screen monitor, computer keyboard, or joystick, depending on the test.

PVT (Dinges and Powell 1985) is a handheld visual sustained attention/reaction time test (Ambulatory Monitoring, Inc, Ardsley, NY). Participants press a button with their preferred hand on the unit as quickly as possible in response to numbers scrolling on the LCD screen with a 3–7 s interstimulus interval. Response time is counted in milliseconds. A solid-state storage unit collects data for downloading to a PC. The primary outcome variable is median reaction time.

Example

Howland et al. (2010) used a placebo-controlled crossover design with randomly assigned order of conditions to assess the effects of binge drinking on students' next-day neurocognitive and academic test-taking performance. Participants were randomized to either high-alcohol beer (mean = .12 g% breath alcohol concentration [BrAC]) or placebo beer on the first night and then received the other beverage a week later. The next day, participants were assessed on test taking, neurocognitive performance (including attention/reaction time), and mood state. Participants were 193 college students (≥ 21 years) recruited from Greater Boston. The Graduate Record Exams © (GREs) and a quiz on a lecture presented the previous day measured test-taking performance; the NES3 and the PVT measured neurocognitive performance; and the Profile of Mood States measured mood.

Test-taking performance was not affected the morning after alcohol administration, but mood state and attention/reaction time were. Of the NES3 measures, visual-spatial function (Visual Span Test) was significantly different by beverage. Memory (Pattern Memory Test) showed significant gender by beverage interaction ($p < 0.04$), with greater impairment after alcohol for women than men. The morning after beverage administration, median attention/reaction time scores, as measured by the PVT, were significantly longer under alcohol condition, relative to placebo condition ($p < 0.01$). The authors conclude that while academic test taking was not affected, the other outcomes could be relevant to traffic safety and occupational performance in tasks that require rapid decisions during sustained attention.

Behavioral Economics

In recent years, behavioral economics has emerged as a means for testing, measuring, and understanding the influence of social, cognitive, and emotional factors on decision making. Experimental methods from this field can be applied to research questions relevant to the study of injury causation. For example, recent survey studies comparing frequency of risk-taking behaviors among college students who consume caffeinated alcoholic beverages (CABs) vs. those who consume non-CABs suggest that the addition of caffeine to alcohol increases risky behaviors. Thombs et al. (2010) found that bar patrons who consumed CABs had a threefold risk of leaving the bar highly intoxicated ($\text{BrAC} > 0.08 \text{ g/\%}$), compared to those who consumed alcohol without caffeine, and a fourfold risk of intending to drive after leaving the bar. O'Brien et al. (2008) found that students who consumed CABs, relative to those who consumed alcohol without caffeine, were more likely to experience a variety of drinking-related negative consequences, including approximately double the risk of experiencing or committing sexual assault, riding with an intoxicated driver, having an alcohol-related accident, or requiring medical treatment. One hypothesis suggested by these findings is that the caffeine masks alcohol-related sensations of sedation and thus induces continued drinking among individuals who, if drinking non-caffeinated alcohol, might moderate their consumption. To assess experimentally whether caffeine added to alcohol increases demand for alcohol, a behavioral economics approach can be used.

Example

Differential demand for caffeinated vs. non-CAB can be assessed at a given breath alcohol by using the “Alcohol Purchase Task – State Version” (APT-S) (MacKillop and Monti 2007). The APT-S assesses the relative value of money vs. alcohol consumption among drinkers and is designed to assess demand for more alcohol in the moment after some drinking has occurred. The measure can be given before drinking and after drinking. The measure asks: “Imagine that you are permitted to drink alcohol RIGHT NOW. How many alcoholic drinks would you consume if they cost the following amounts of money? The available drinks are a standard size domestic beer (12 oz.), wine (5 oz.), shots of hard liquor (1.5 oz.), or mixed drinks containing one shot of liquor. Please assume that you would consume every drink you request; that is, you cannot stockpile drinks for a later date or bring drinks home with you.” The APT-S uses 25 prices, ranging from no cost (\$0) to \$100 per drink, with greater price density at lower price. This measure is likely to be sensitive to effects of caffeine vs. no caffeine on demand to continue drinking. The comparison between alcohol and placebo alcohol will allow a manipulation check on the measure, since such effects have previously been demonstrated. The measures will be scored for: intensity of demand (i.e., consumption under minimal cost), breakpoint (i.e., the price that suppresses drinking to zero), P_{max} (i.e., the price that reflects the transition from inelastic to elastic demand, another measure of price sensitivity), and O_{max} (i.e., maximum output, or expenditure on alcohol). O_{max} , and breakpoint are most sensitive to state-dependent changes in motivation.

Measuring Sleep Quality

Fatigue due to lack of sleep or poor quality sleep can be a cause of injury, such as injury caused by traffic crashes. Fatigue may also play a mediating role in injury causation as, for example, when an

exposure causes sleep disturbance that in turn causes fatigue and consequent risk for injury. In laboratory settings, polysomnography (used to diagnose sleep disorders) can be used to objectively measure sleep quality, and self-report measures (Rohsenow et al. 2006) can assess perceived sleep quality or fatigue.

The sleep recording montage includes a number of measures: continuous electroencephalogram (EEG) from electrodes affixed to the scalp; bipolar electrocardiogram (ECG) from electrodes taped on the right shoulder and left side; thoracic and abdominal excursions measured by inductive plethysmography bands; airflow detected by a nasal–oral thermocouple; snoring detected by a snoring microphone; blood oxygen saturation measured by finger pulse oximetry; and nocturnal limb movements assessed using bilateral EMG electrodes placed over the anterior tibialis.

Primary measures of sleep quality are sleep continuity and sleep architecture (the pattern of REM and non-REM sleep). Sleep continuity measures include sleep onset latency (time from “lights out” to the first epoch of three consecutive epochs of any stage of sleep), latency to consolidated sleep (time from “lights out” to the first 10 consecutive minutes of sleep), total sleep time (the total minutes of REM and non-REM sleep within time between sleep onset and the last epoch of sleep), sleep efficiency (total sleep time/time between “lights out” and “lights on”), arousal index (number of EEG-defined REM and non-REM arousals per total sleep period), and the amount of wakefulness after sleep onset (total time awake from sleep onset to “lights on”). Sleep architecture measures include minutes and percentages of Stages 1, 2, slow wave sleep (SWS Stages 3 and 4), and REM, latency to SWS (time from initial sleep onset to first 30-s epoch of Stage 3 sleep), REM sleep latency (time from sleep onset latency to first 30-s epoch of REM sleep), and REM density (frequency of eye movements per epoch in total sleep period) calculated during the total sleep period.

Research assistants can be trained in the application of electrodes and respiratory equipments; however, a registered polysomnographic technologist is required to monitor sleep recording in real time and process/interpret recordings. Participants must be screened for sleep disorders through history and a preexperimental night, which serves to screen for sleeping problems as well as accommodate to the sleep monitoring equipment and procedures. In addition, participants must adhere to a sleep regimen for several days before experimental sessions. Compliance can be validated by use of an actigraph (which records motion), a sleep diary, and bedtime and wake-up calls to a time/date stamped phone answering system.

Example

In the aforementioned mixed within-subjects and between-groups study (Rohsenow et al. 2010) that compared the residual effects of high and low congener alcohol on next-day neurocognitive performance, the investigators monitored participants' sleep quality. Polysomnography was used during an 8-h sleep opportunity to document the effects of intoxication at BrAC 0.11 g%, vs. placebo, on sleep and to examine whether alcohol-related sleep disturbance mediated next-day neurocognitive impairment. Alcohol decreased sleep efficiency and rapid eye movement sleep, and increased wake time and next-day sleepiness, with bourbon and vodka having equivalent effects. Alcohol effects on sleep did not mediate next-day decrements on tests requiring both sustained attention time and speed. Investigators concluded that sleep disturbance was not the mechanism by which heavy drinking impaired performance the following day.

Measuring Subjective States

Measuring subjective states can also be important to understanding injury causality. The authors have developed a number of scales for measuring alcohol and other drug effects on participant perceptions of their own level of intoxication, their performance on a particular task, and their fitness to perform functions.

Examples

Self-estimates of intoxication: In a randomized between-groups trial, Howland et al. (2011) compared the effects of intoxication (mean BrAC: 12.0 g%) with caffeinated to non-caffeinated beer on simulated driving and subjective measures. One hypothesis regarding the effects of CABs is that the caffeine offsets the sedating effects of alcohol and thus may distort sensations of intoxication. This could lead to increased alcohol consumption and thus increased risk for injury to self or others. To determine whether the coadministration of alcohol and caffeine affects perceptions of intoxication, relative to alcohol alone, participants estimated their blood alcohol level following beverage administration and absorption period on a scale from 0 to 0.15 g% using a questionnaire we developed for previous studies, the self-estimate of blood alcohol concentration (SEBAC). There was no difference in the estimate of BrAC between participants who received caffeinated beer vs. non-caffeinated beer (0.11 ± 0.02 g% vs. 0.10 ± 0.02 g%).

Self-rated performance: In an aforementioned study of the acute effects of low-dose alcohol exposure on occupational performance, Howland et al. (2000) administered a self-rated performance scale after completion of simulator scenarios to participants in both the placebo and alcohol (breath alcohol of 0.04 g%) groups. Self-rated performance did not differ between the groups, despite significant actual performance decrements among the group receiving alcohol. This finding suggests that in addition to impairing occupational performance, low doses of alcohol affect perception of performance. This finding has implications for safety because it indicates that the participants were not able to discern their performance decrements and thus may not be prompted to take corrective actions or remove themselves from duty in a real-world occupational situation.

Self-rated fitness-to-drive: To assess state risk-taking propensity the authors have developed a scale measuring willingness to drive, which can be used under various experimental conditions. Rated on 5-point fully anchored scales: “Right now, would your ability to drive a car be worse than usual?” (1, “not affected at all” to 5, “very much worse”), “How likely is it that you would drive a vehicle the way you feel right now?” (1, “definitely would not” to 5, “definitely would.”), and “How likely is it that you would ride with a driver who’d been drinking given the way you feel right now?” (1, “definitely would not” to 5, “definitely would.”). Surveys found different effects for likelihood of driving vs. riding with a drinking driver, so questions on both behaviors are asked. In Rohsenow et al. (2010), although people had slower reaction time the morning after drinking to intoxication, they rated their ability to drive as less impaired compared to after placebo, although they also said they would be less likely to drive when feeling that way. Therefore, people may not be able to detect their own impairment.

Summary

Injury research plays a vital role in informing public and private policy and in guiding individuals' decision making. Thus, it is important that research yields the strongest evidence about the causes and prevention of injury-related morbidity and mortality. As noted above, injury research is often constrained by ethical and practical considerations and consequently relies heavily on observational studies. An advantage of observational research is that exposures and outcomes occur in the real world. The disadvantage is that the real world includes innumerable elements that threaten internal validity. Experimentation, on the other hand, can be conducted in more controlled settings that can increase the validity of causal testing but often at the expense of ecological validity. Accordingly, observational and experimental designs can be complementary, and when findings align their combined evidence can be compelling. Our aim in this chapter is to argue for the application of experimentation when possible and to demonstrate the value of experimental methods as a complement to those traditionally used in injury research.

References

- Anastasi, A. (1988). *Psychological testing* (6th ed., p. 144). New York: Macmillan.
- Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dinges, D. F., & Powell, J. E. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, *17*, 652–655.
- Hingson, R., McGovern, T., Howland, J., Heeren, T., Winter, M. R., & Zakocs, R. (1996). Reducing alcohol-impaired driving in Massachusetts: the Saving Lives Program. *American Journal of Public Health*, *86*(6), 791–797.
- Howland, J., Rohsenow, D. J., Cote, J., Siegel, M., & Mangione, T. W. (2000). Effects of low-dose alcohol exposure on simulated merchant ship power plant operation by maritime cadets. *Addiction*, *95*, 719–726.
- Howland, J., Rohsenow, D., Minsky, S., Snöberg, J., Tagerud, S., Hunt, S., et al. (2008). The effects of transdermal scopolamine on simulated ship navigation and attention/reaction time. *International Journal of Occupational and Environmental Health*, *14*, 250–256.
- Howland, J., Rohsenow, D. J., Greece, J. A., Littlefield, C. A., Almeida, A. B., Heeren, T., et al. (2010). The effects of binge drinking on college students' next-day academic test-taking performance and mood state. *Addiction*, *105*, 655–665.
- Howland, J., Rohsenow, D. J., Arnedt, J. T., Bliss, C. A., Hunt, S. K., Calise, T. V., et al. (2011). The acute effects of caffeinated versus non-caffeinated alcoholic beverage on driving performance and attention/reaction time. *Addiction*, *106*, 335–341.
- Kelly, T. H., Taylor, R. C., Heishman, S. J., & Howland, J. (2007). Performance-based assessment of behavioral impairment in occupational settings. In S. B. Karch (Ed.), *Pharmacokinetics and pharmacodynamics of abused drugs*. Boca Raton, FL: CRC.
- MacKillop, J., & Monti, P. M. (2007). Advances in the scientific study of craving for alcohol and tobacco: from scientific study to clinical practice. In P. M. Miller & D. J. Kavanagh (Eds.), *Translation of addictions sciences into practice* (pp. 187–207). Amsterdam: Elsevier.
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*, 40–48.
- Monti, P. M., Colby, S. M., Barnett, N. P., Spirito, A., Rohsenow, D. J., Myers, M., et al. (1999). Brief intervention for harm-reduction with alcohol-positive older adolescents in a hospital emergency department. *Journal of Consulting and Clinical Psychology*, *67*, 989–994.
- O'Brien, M. C., McCoy, T. P., Rhodes, S. D., Wagoner, A., & Wolfson, M. (2008). Caffeinated cocktails: energy drink consumption, high-risk drinking, and alcohol-related consequences among college students. *Academic Emergency Medicine*, *15*(5), 453–460.
- Rohsenow, D. J., & Marlatt, G. A. (1981). The balanced placebo design: methodological considerations. *Addictive Behaviors*, *6*, 107–122.
- Rohsenow, D. J., Monti, P. M., Hutchison, K. E., Swift, R. M., Colby, S. M., & Kaplan, G. B. (2000). Naltrexone's effects on reactivity to alcohol cues among alcoholic men. *Journal of Abnormal Psychology*, *109*, 738–742.
- Rohsenow, D. J., Howland, J., Minsky, S., & Arnedt, J. T. (2006). Effects of heavy drinking by maritime academy cadets on hangover, perceived sleep, and next day ship power plant operation. *Journal of Studies on Alcohol*, *67*(3), 406–415 [PMID 16608150].

- Rohsenow, D. J., Howland, J., Arnedt, J. T., Almeida, A. B., Greece, J. A., Minsky, S., et al. (2010). Intoxication with bourbon versus vodka: effects of hangover, sleep and next-day neurocognitive performance in young adults. *Alcohol Clinical and Experimental Research*, *34*, 509–518.
- Spirito, A., Sindelar-Manning, H., Colby, S. M., Barnett, N. P., Lewander, W., Rohsenow, D. J., et al. (2011). Motivational interventions for alcohol positive adolescents treated in an emergency department: results of a randomized clinical trial. *Archives of Pediatric and Adolescent Medicine*, *165*(3), 269–274.
- Tennstedt, S., Howland, J., Lachman, M., Peterson, E., Kasten, L., & Jette, A. (1998). A randomized, controlled trial of a group intervention to reduce fear of falling and associated activity restriction in older adults. *Journal of Gerontology Psychological Sciences*, *53B*(6), 384–392.
- Thombs, D. L., O'Mara, R. J., Tsukamoto, M., Rossheim, M. E., Weiler, R. M., Merves, M. L., et al. (2010). Event-level analyses of energy drink consumption and alcohol intoxication in bar patrons. *Addictive Behaviors*, *35*(4), 325–330.
- White, R. F., James, K. E., Vasterling, J. J., Letz, R. E., Marans, K., Delaney, R., et al. (2003). Neuropsychological screening for cognitive impairment using computer-assisted tasks. *Assessment*, *10*, 86–101.
- Yesavage, J., & Leirer, V. (1986). Hangover effects on aircraft pilots 14 hours after alcohol ingestion: a preliminary report. *American Journal of Psychiatry*, *143*, 1546–1550.

Chapter 9

Epidemiologic Methods

Guohua Li and Susan P. Baker

Introduction

Epidemiology is defined as the scientific discipline that studies the distribution and determinants of disease and injury in population groups across time and space. As a basic science of medicine and public health, epidemiology is practiced in virtually every clinical specialty and health-related field. At the core of epidemiology is a systematized set of concepts and methods developed over the past 150 years for understanding illness and health in human populations. These epidemiologic concepts and methods serve a wide array of functions in disease investigations and health studies, including (1) identifying the causes of a given medical condition or health problem; (2) measuring the morbidity and mortality in a defined population; (3) understanding the natural history and prognosis of disease; (4) evaluating the performance of diagnostic, therapeutic, and preventive measures; and (5) providing empiric evidence for the development of public policy to protect and improve health at the population level (Gordis 1996). Although some of the basic epidemiologic concepts, such as endemic, epidemic, and environment, can be traced back to Hippocrates, modern epidemiologic methods did not emerge until the 1840s and did not take root until the 1950s. John Snow, a general practitioner and a pioneer of anesthesia, is widely regarded as one of the founding fathers of modern epidemiology. He is credited with not only identifying contaminated water as the transmission mode of cholera, but also establishing the central paradigm of epidemiology – namely, that the understanding of the causes of disease can be advanced by methodically examining the tempo-spatial patterns of the disease in different population groups (Bhopal 2008).

During the second half of the nineteenth century and the first 4 decades of the twentieth century, epidemiologic methods were primarily used in the studies of infectious diseases, such as tuberculosis, influenza, typhoid, and diphtheria. By the 1940s, most of the major infectious diseases had been effectively controlled in industrialized countries. The public health benefit from the remarkable reduction in infectious diseases, however, was offset, to a considerable degree, by the rapidly increasing morbidity and mortality from injury, particularly injury resulting from motor vehicle crashes (Li et al. 1995). In the USA, by the early 1940s, injury had become the leading cause of death for

G. Li, MD, DrPH (✉)

Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA

Department of Anesthesiology, Columbia University College of Physicians and Surgeons, New York, NY, USA

e-mail: GL2240@mail.cumc.columbia.edu

S.P. Baker, MPH, ScD (Hon.)

Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

e-mail: sbaker@jhsph.edu

children and young adults and the fifth leading cause of death overall (Armstrong et al. 1945). With the shift in mortality patterns, researchers began to increasingly apply epidemiologic methods in studying noncommunicable chronic diseases and injuries. Godfrey (1937), Armstrong et al. (1945), Press (1948), Gordon (1949), and King (1949) were among the first to recognize the utility of epidemiologic methods in “accident” research and propose the public health approach to “accident” prevention. Since then, epidemiologic methods have become a prominent component of the tool box of injury researchers, and injury epidemiology has grown into a well-established specialty. While Gordon (1949) and Gibson (1961) laid the foundation for injury epidemiology by defining the agent of injury, Haddon (1968, 1980) played a pivotal role in establishing injury epidemiology and prevention as a scientific discipline through his contribution to the conceptual frameworks, pioneering research, prolific writing, and effective advocacy. His work has had a lasting influence on injury research and practice and has inspired several generations of injury epidemiologists.

Epidemiologic concepts and methods are described in depth in many books and are widely taught to public health and medical students. The contributions of injury researchers to epidemiologic concepts and methods and the unique features of injury epidemiology, however, are often overlooked in general epidemiology texts. This book contains several chapters discussing different aspects of epidemiologic methods as applied to injury research, including Chaps. 1–4 on injury surveillance, Chaps. 13–19 on trauma management and outcomes, Chaps. 20–26 on injury data analysis, and Chap. 35 on intervention program evaluation. This chapter discusses the basic epidemiologic concepts, methods, and study designs in the context of etiologic studies of injury.

Epidemiologic Framework

Epidemiology as a discipline is eclectic; it draws on theories and methods from both biological and social sciences. Epidemiology has penetrated almost every subject area of public health and medicine and developed into many specialties with regard to the specific health problems of interest, such as infectious disease, cardiovascular disease, cancer, and injury. Accordingly, epidemiology is often classified into specialties such as infectious disease epidemiology, cardiovascular disease epidemiology, cancer epidemiology, injury epidemiology, and psychiatric epidemiology. Another way to classify epidemiology into different specialties is based on the etiologic domains of inquiry. Major specialties under this classification system include environmental epidemiology, social epidemiology, genetic epidemiology, and behavioral epidemiology. Because of the important role environmental and behavioral factors play in injury (see Chaps. 11 and 12), injury epidemiology intersects substantially with environmental epidemiology and behavioral epidemiology. Finally, epidemiology can be classified according to the methodological orientations that underpin all the subject areas and etiologic domains. Therefore, reflective of the design features, epidemiologic studies can be categorized into observational epidemiology, experimental epidemiology, and theoretical epidemiology (Lilienfeld and Lilienfeld 1980).

Epidemiologic Triad

The conceptual model used by epidemiologists for understanding the causes of disease has evolved from studies of infectious diseases and proven readily applicable to studying other health problems. The central premise of this conceptual model is that disease is always caused by the interplay of three component factors: agent, host, and environment. The graphic presentation of these three factors in relation to disease causation is known as the epidemiologic triad.

The agent factor refers to the etiologic element that is necessary for the initiation of the pathological process of a disease. Depending on the medical conditions, etiologic elements may appear in different forms. For infectious diseases, the agents are the microorganisms that cause the infections. For other diseases, the agents may be the conditions or states created by the presence, absence, or imbalance of the etiologic elements. Disease agents are generally composed of three types (1) biological, such as bacteria, viruses, and fungi; (2) chemical, such as toxins, lipids, proteins, and vitamins; and (3) physical, such as radiation, heat, and mechanical forces. For many infectious agents, an animal carrier (e.g., an insect) often serves as a vector for the pathogen to pass from one person to another. The host factor consists of the personal characteristics (e.g., genetic makeup, age, education, occupation, medical history, health status, and risk-taking behavior) that influence the likelihood of being exposed to the disease agent, susceptibility to the disease given exposure, and ability to recover once diseased. In the epidemiologic triad, the environmental factor encompasses all the extrinsic variables that may influence the occurrence and outcome of the disease in a population, including the physical environment (e.g., climate, terrain, air quality, structures, and products), the biological environment (e.g., population density, fauna, and flora), and the socioeconomic environment (e.g., socioeconomic status, policy, law, culture, and access to care).

The Agent Factor

Gordon (1949) was probably the first epidemiologist to systematically apply the epidemiologic triad in examining the causes of injury. Although the common agent for most injuries – energy transferred acutely and excessively – had not been explicitly identified at that time, Gordon (1949) stated aptly that “the agents concerned with injuries and with accidents, like those of disease, are variously of physical, chemical, and biologic nature.” Before the specific agent of injury was clearly delineated by Gibson (1961), researchers mistook vectors for agents. For instance, Gordon (1949) thought that fall injury related to “a faulty ladder, a playful pup, or a misplaced handbag” had different agents (i.e., ladder, pup, and handbag, respectively) and that injuries sustained through cutting, piercing, and collision with regard to a glass door shared the same agent (i.e., the glass door).

King (1949) came close to identifying the agent of injury when he hypothesized that accidents occurred when external stresses exceeded the minimum level of human tolerance. He even outlined three categories of stress: physical (e.g., thermal, mechanical, radiation, and electrical stresses); chemical (e.g., toxins and drugs); and biological (psychological stress, aging and disease, and malnutrition) (King 1949). However, King failed to separate injury from accident or to differentiate host and environmental factors from etiological factors, thus missing the opportunity to make the fundamental discovery of the common agent of injury. The conceptual breakthrough in explicitly identifying the agent of injury occurred in 1961 when James Gibson, an experimental psychologist, concluded that “injuries to a living organism can be produced only by some energy interchange” (Gibson 1961). There are two general mechanisms through which abnormal energy exchange causes injury. The first is that energy is transmitted to the human body in an amount that exceeds the injury threshold. This mechanism accounts for the majority of injuries, including those resulting from motor vehicle crashes, airplane crashes, gun-shot wounds, falls, burns, and electrocutions. The second mechanism does not involve excess energy exchange but the interference with normal energy exchange necessary for maintaining physiological functions. Drowning and suffocation are among the examples of injury resulting from the second mechanism. Although energies involved in injury causation could be mechanical, thermal, radiant, electrical, and chemical, they are physical in essence and thus can be adequately understood and effectively controlled.

Another notable development in the evolution of injury epidemiology was the recognition of vectors (or vehicles) that carry the etiologic agent of injury. Haddon (1963) laid the conceptual foundation for applying the epidemiologic triad to injury research and prevention by clarifying the

confusion about the agents and vectors of injury. He was the first to specifically point out that objects directly involved in injury causation, such as machinery, bullets, and electric lines, are not agents but vectors. In some injury scenarios, the host could also be the vector, as in the case of a person falling or a pedestrian injured by unintentionally bumping into a telephone pole (Haddon 1963). Because the vectors of injury are mostly man-made and amenable to modification, they represent important targets for effective intervention programs to reduce injury occurrences and mitigate injury severity and other consequences.

The Host Factor

In the epidemiologic triad, the host refers to the individual persons or population groups that are susceptible to injury. The host factor includes all the variables characterizing the human subjects under study on either the individual or the population level. Variables that may influence one's susceptibility to injury can be categorized into three groups (1) biological, such as age, sex, race, and genotype; (2) psychosocial, such as personality, education, occupation, marital status, and place of residence; and (3) behavioral, such as risk-taking, alcohol and drug use, seatbelt use, and safety helmet use. Because the host factor is concerned with people, it is also called a personal factor or intrinsic factor. "Human factors," however, is a specially defined term, referring to the engineering specialty that aims to optimize the man-machine interface.

The host factor has long been the focal point of injury research, particularly before the 1960s. Although biological, psychosocial, and behavioral characteristics are important determinants of injury (see Chaps. 11, 12, and 17) and are useful for understanding the epidemiologic patterns of injury and for identifying high-risk groups, they are either immutable or difficult to change. Therefore, research that focuses on the host factor often has limited value for injury control and prevention.

Despite the exhaustive effort to delineate the host factor of injury, two overarching areas of scientific importance remain largely unexplored. The first area concerns the interaction between injury and human genetics. Genetic markers are associated with risk-taking behavior, susceptibility to injury, and survival outcome of injury (Neves et al. 2010; Grigorenko et al. 2010; Dash et al. 2010). Relative to diseases, injury tends to disproportionately affect children and young adults, often resulting in premature death before procreation, and often aggregating in families. It is not uncommon for several family members or an entire family to perish in the same car crash or plane crash. The hypothesis that injury may play a significant role in natural selection and the evolution of human genetics was proposed by Gibson (1961) and articulated by Haddon (1963, 1980). With the completion of the human genome project and the remarkable advances in biotechnology, examining the interface between injury and human genetics has become increasingly practical.

The second important yet understudied area regarding the host factor of injury is the complex relationship between injury and disease. Although it is well recognized that people with preexisting medical conditions, comorbidities, or disabilities are at a significantly increased risk of injury and that certain types of injury (e.g., traumatic brain injury and hip fracture) may drastically increase the victim's susceptibility to specific diseases or worsen the prognosis of preexisting medical conditions, the dynamics of the injury-disease relationship during the life course of human development has not been well understood.

The Environment Factor

The concept of environment in injury epidemiology is expansive, comprising all the elements constituting the context, circumstance, and conditions that may influence, directly or indirectly, the occurrence of injury. The environmental factor can be generally categorized into three groups:

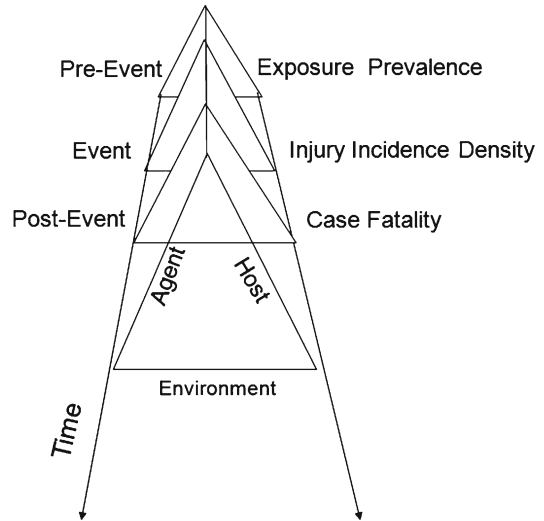
physical, biological, and socioeconomic (Gordon 1949). The physical environment includes both natural features, such as climate, weather, and terrain, and man-made structures and conditions, such as windows, stairs, stoves, lighting, air conditioning, and ventilation. Overall, the physical environment is the most important determinant of injury because it can act as the vector of the etiologic agent, influence the form, duration, and intensity of the energy transferred to the host, and substantially modify the host's susceptibility to injury. The biological environment, such as population density and fauna, may also play a significant role in the causation of traffic injury and other types of injury (Björnstig 1992; Goldstein et al. 2011). The physical environment is especially relevant to injury prevention and control because it can be most easily changed. The socioeconomic environment is composed of a variety of constructs, such as socioeconomic status, income equality, policy, regulation, law, international treaties, culture, and emergency medical services systems. The socioeconomic environment has become increasingly complex and its relation to injury remains an important topic of epidemiologic research.

Like many diseases, injury often shows discernable patterns when examined according to environmental characteristics, such as socioeconomic status and geographic regions. For instance, Gordon (1949) found that the incidence rates of injury and pneumonia in the USA were inversely associated with family income in similar patterns. There is a large body of literature documenting the geographic variations of different types of injury (Baker et al. 1984, 1991; Waller et al. 1989; Kearney and Li 2000). In fact, analysis and interpretation of geographic variations in injury morbidity and mortality have been an essential part of research efforts throughout the history of injury epidemiology (Gordon 1949; Baker et al. 1984; Goldstein et al. 2011). Recent developments in geographic information systems and statistical techniques for spatial data analysis have significantly advanced this area of inquiry (see Chaps. 24 and 25). For instance, researchers had long documented the marked differences in suicide rates across geographic regions in the USA (Baker et al. 1984; Cheng 2010). Early studies, however, were limited to cross-state comparisons of suicide mortality data, and shed little light on the causes of the observed geographic patterns. Lawrynowicz and Baker (2005) showed that suicide rates were higher not only in northern latitudes but also in the most southern latitudes, suggesting the effect of seasonal or diurnal light variation on suicide rates. With more refined geospatial data and analytical techniques, recent studies have pinpointed the altitude gradient in the geographic variations in suicide, leading to the hypothesis that higher suicide rates in mountainous areas are due to mild hypoxia and increased mood disorders (Haws et al. 2009; Kim et al. 2011).

Extension of the Epidemiologic Triad

One of the most important yet underappreciated contributions injury researchers made to the discipline of epidemiology is transforming its conceptual framework from a static model to a dynamic one (Fig. 9.1). Specifically, Haddon (1968, 1972, 1980) added a time dimension to the epidemiologic triad and created a conceptual framework that is much more powerful than the classic one. Time is an ultimate continuous variable. For simplicity, Haddon (1980) divided time into three phases: pre-event, event, and post-event. The pre-event phase refers to those factors that contribute to or prevent the event from happening (e.g., road design and brakes). Event-phase factors determine whether a person is injured when an event occurs (e.g., helmets and airbags). Post-event factors include everything that determines survival and the degree of recovery. Combining the three-phase time dimension with the epidemiologic triad forms a 3 by 3 matrix, known as the Haddon matrix. Although the extended conceptual framework is designed and used primarily for guiding the development of injury control strategies and countermeasures, it provides an enhanced and a coherent theoretical model for understanding the causes of injury and for bridging the divides among different disciplines that concentrate on issues in different cells of the Haddon matrix. Although the enhanced theoretical model

Fig. 9.1 Graphical presentation relating the epidemiologic triad to the Haddon matrix



developed by Haddon is clearly applicable to epidemiologic studies of other health problems, its influence so far has been largely confined to injury research and prevention (see Chap. 29).

In the past 2 decades, injury epidemiologists have made strides in understanding the mathematical foundation of the Haddon matrix and in translating the qualitative conceptual framework into quantitative methods (Li and Baker 1996; Li et al. 1998, 2003; Dellinger et al. 2002; Zwerling et al. 2005; Meuleners et al. 2006; Goldstein et al. 2011). Li and Baker (1996) found that the population-based injury mortality rate is a simple multiplicative function of exposure prevalence, injury incidence density, and case fatality, which correspond to the pre-event, event, and post-event phases in the Haddon matrix (Fig. 9.1). Building on the mathematical function, a decomposition method has been developed for identifying relevant factors contributing to injury mortality (Li et al. 1998; Porta 2008; Goldstein et al. 2011). To facilitate the use of the Haddon matrix by policy makers and injury prevention practitioners, Runyan (1998) added to the conceptual framework a dimension of decision criteria, such as effectiveness, cost, liberty, equity, and feasibility.

Epidemiologic Causation

Methods and conditions for inferring cause-and-effect relations have been the central theme of scientific philosophy for centuries. In epidemiologic investigations of infectious diseases with unknown causes, the most pressing challenge facing researchers is to identify the etiologic agent or the pathogen. The Henle–Koch Postulates, which were formulated to evaluate the causality between an agent and an infectious disease, require that the agent be present in all the patients with the disease, not found in any patient without the disease, capable of reproducing the disease in susceptible human subjects or animals, and present in all the individuals with the experimentally produced disease (Porta 2008). The Henle–Koch model, however, does not apply well to chronic diseases because most of them result from a complex interplay of multiple causal factors and cannot be easily subjected to experimentation. The causes of chronic diseases may be proximal and direct as well as distal and indirect. Consequently, epidemiologists become increasingly reliant on observational data, which often allow only for the identification of statistical associations. Thus, the philosophical

challenge of reasoning about causality is reduced to the practical question of how to determine whether an observed association is of causal nature. In the past 5 decades, epidemiologists have developed a conceptual framework for guiding evidence evaluation to identify causal associations. The foundation of this conceptual framework was laid out by Hill (1965), who proposed “nine different viewpoints” for consideration before interpreting an association as causation:

1. Strength of the association, which is measured by relative risk or absolute risk in those who were exposed to the putative cause compared to those who were not. The evidence for causation increases as the strength of the observed association increases.
2. Consistency of the association, which refers to the degree of reproducibility of the evidence observed in different study population groups, geographic regions, and time periods, and with different study designs. Although inconsistency in findings from different studies does not necessarily rule out a causal association, a high degree of consistency strengthens the evidence for causation.
3. Specificity of the association, which refers to the exclusivity in the association between a putative cause and a disease. If the putative cause is associated with the particular disease only, or if the disease occurs only in those who are exposed to the putative cause, the evidence for causation is strengthened.
4. Temporality, which represents the necessary condition for causation that the exposure to the putative cause must precede the disease.
5. Biological gradient, which refers to the dose–response phenomenon. The presence of a dose–response relationship between the degree of exposure to the putative cause and the disease risk is evidence in favor of causation.
6. Plausibility, which refers to the degree to which the observed association agrees with the contemporary biological science. Biological plausibility provides corroborative evidence for causation.
7. Coherence, which refers to the compatibility between the suspected causation and the general knowledge base. Inferring causality from the observed association should not seriously contradict established scientific theories and knowledge.
8. Experiment, which refers to supportive evidence from experimental regimens. The association between the putative cause and the disease should change logically when the exposure to the putative cause is altered.
9. Analogy, which refers to evidence from established causal relationships that are conceptually similar or biologically analogous to the observed association.

The above framework proposed by Hill is discussed in many subsequent epidemiology texts. Most of the authors omit “coherence” and “analogy” from their discussion because of the perceived irrelevance or redundancy. Although none of the nine elements in the conceptual framework constitutes a sufficient condition for causality, each contributes to the evidence base for the evaluation of causation. For instance, “analogy,” which is conceptually similar to the consideration of precedence in judicial reasoning, may render valuable evidence to the adjudication of the suspected causation, but is widely disregarded by contemporary epidemiologists. There is also a commonly held misperception that the elements in this conceptual framework proposed by Hill (1965) form the criteria of causation. This misperception has led to the questionable practice of using the framework as a checklist.

Hill’s conceptual framework, which has been elaborated by other epidemiology scholars (Rothman 1976; Susser 1991), made a unique contribution to scientific philosophy. Its holistic approach provides a sound and comprehensive structure that incorporates all the relevant evidence into the adjudication of causality, and has served the empirical science of epidemiology extremely well. Implicit in Hill’s conceptual framework is the recognition of the cognitive aspect in causal inference and the limitations inherent in observational data, such as measurement error, confounding, and missing information. While explaining the rationale of his conceptual framework, Hill (1965) stated that “all scientific work is incomplete – whether it be observational or experimental. All scientific work is liable to be upset or modified by advancing knowledge.”

It is noteworthy that causality is defined, inferred, and interpreted differently in different disciplines. The conceptual framework discussed above represents the prevailing view on causation in epidemiology. Recent developments in statistics and computer science have significantly advanced the idea of automating the process of causal inference through direct acyclic graphs, causal networks and models, and mathematical and computational techniques (Pearl 2009). These methodological innovations have been increasingly introduced into epidemiology, resulting in many promising conceptual models and novel approaches to research design and data analysis that may help improve causal thinking and practice. It is worth noting that the automation or mathematicalization of causal inference always relies on some assumption about the nature of the interrelations among the variables under study or the joint distribution of the data that is not testable based on observational data (Pearl 2009). Therefore, statistically- and mathematically-based causal models may play a complementary role in the evaluation of causative relations but are unlikely to substitute for the epidemiologist's content expertise and judicious deliberation called for by the conceptual framework proposed by Hill (1965).

Injury as a Disease

Before 1979, injury was labeled as "accident" in the International Classification of Diseases (ICD) system. Since then, injury has been categorized in the ninth and tenth editions of the ICD system as a miscellaneous group called "injury, poisoning and certain other consequences of external causes." Although both "disease" and "injury" describe adverse health conditions, the current ICD classification seems to imply that the former is due to internal causes intrinsic to the host and the latter due to external causes originating in the environment. This differentiation is obviously inaccurate because the epidemiologic triad applies to injury and disease equally well, and because in some cases the etiologic agents for disease and injury are the same. Baker and Haddon (1974) pointed out that the damage to the spine from an acute, excess transfer of mechanical forces is usually called injury, whereas the damage to the spine from chronic, small-dose mechanical energy exchanges may result in a disease (e.g., herniated disk). Similarly, exposure to ionizing radiation could result in burns (injury) instantly or leukemia (disease) within years.

In spite of the similarities between disease and injury in causes and consequences, there are at least three notable features that distinguish injury from most diseases. First, injury has an extremely short latency (i.e., the time interval between the exposure to the etiologic agent and the manifestation of injury). The latency for infectious diseases is usually called the incubation period, which may range from a few hours to several months. For chronic diseases, it usually takes years for clinical symptoms to emerge after the initial exposure to etiologic agents. The latency for diseases provides a precious time window for early diagnosis and early intervention. For most injuries, the latency is so short that it is virtually nonexistent. Therefore, strategies for early diagnosis and early intervention developed for disease control are largely irrelevant to injury control, with the exception of belated hemorrhage in the brain and other internal organs, in which clinical signs may not manifest until several hours after the initial physical impact.

Second, injury could result from both unintentional and intentional acts. While acquiring a disease on purpose is extremely rare, intentional injury through acts directed to oneself or another is common. Each year, in the USA, intentional injury claims over 50,000 lives in the forms of suicide and homicide and results in about two million emergency department visits in the forms of suicide attempts and assaults (Betz and Li 2005). Although the intentionality of injury is usually dichotomized into unintentional and intentional, in reality the intent for many injuries is not clear-cut and may best be classified as partially intentional and partially unintentional. This continuum of intentionality blurs the artificial boundaries between unintentional and intentional injuries as well as between suicide and homicide (Bills and Li 2005). For instance, a man who is furious with his child might shake him violently, injuring or even killing him, but without the intent to do so. Or, a depressed

woman who did not care whether she lived or died might drive very recklessly. Intentionality introduces many unique factors into injury causation and makes injury a health problem that is more complex and more difficult to understand and control than most diseases.

Finally, unlike many diseases, injury cannot be prevented by vaccination or pharmaceutical prophylaxis, with the exception of taking vitamin D and calcium supplements to reduce the risk of fractures from falls in the osteoporotic elderly. Susceptibility to injury is universal and lasts throughout the lifespan. Past injury does not produce any immunity to the host. On the contrary, a positive history of injury is an indicator of increased propensity to future injury. The human host's universal susceptibility to injury and the inapplicability of vaccine and pharmaceutical prophylaxis to injury prevention make it necessary to focus on countermeasures that operate at the population level, are directed at environmental factors, and require little or no action by individuals (Haddon 1980).

Epidemiologic Designs

Epidemiologic studies can be generally categorized into two groups based on research designs: experimental and observational. The most important attribute that differentiates the two types of research designs is the manner by which study subjects are assigned to different groups. The experimental design requires that study subjects be allocated into treatment and control groups through randomization. The randomized assignment could be performed at the individual level or at the group or community level. The observational design, on the other hand, does not involve randomized assignment of study subjects into different groups. Randomization in the experimental design creates a condition that approximates the counterfactual scenario. Thus, experimental studies are widely regarded as being less susceptible to biases and producing more rigorous evidence for causality than observational studies. Epidemiologic research in recent years, however, has become increasingly reliant on observational designs due to cost and feasibility constraints in the experimental design as well as continuing improvements in observational methodology. Both experimental and observational designs play an integral part in injury epidemiology (Fig. 9.2). The following discussion

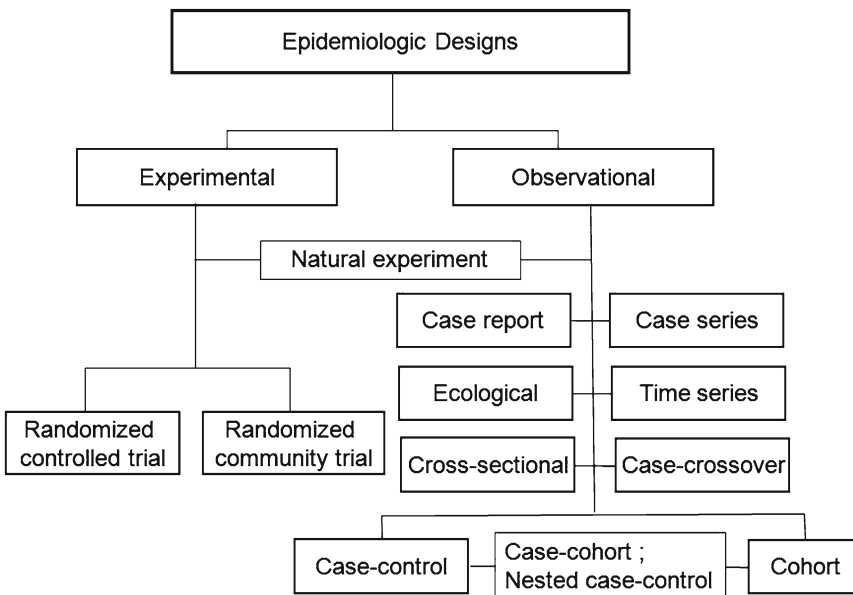


Fig. 9.2 Classification of epidemiologic study designs

focuses on epidemiologic designs that are commonly used in observational studies pertaining to injury causation and prevention. Experimental methods are discussed in detail in Chap. 8.

Case–Control Design

Establishment of the case–control design as a scientifically sound method is arguably the most important development in epidemiology in the twentieth century. The basic tenet of the case–control design is the comparison of exposure prevalence to putative causes between a group of people with the injury (cases) and a group of people without the injury (controls). Because the case status can be determined only after a study subject has sustained the injury and because the search for causality proceeds backward from injury to putative causes, case–control studies are often called retrospective studies. Although case–control studies are retrospective in essence, not all retrospective studies or comparative analyses can be regarded as case–control studies. One of the distinctive features of the case–control design is the special requirement that cases and controls are selected through such a sampling scheme that they are representative of the injured and the non-injured populations, respectively. Specifically, the case–control design requires that the selection of cases and controls be independent of the individual's exposure status to any putative cause. That is, an individual who sustained the injury and who was exposed to a supposed cause has the same chance to be selected as a case as an individual who sustained the injury but was not exposed to the putative cause; and an individual who did not sustain the injury and who was exposed to a supposed cause has the same chance to be selected as a control as an individual who neither sustained the injury nor was exposed to the putative cause. This special requirement is necessary because the scientific foundation of the case–control design that the odds ratio of exposure is equivalent to the odds ratio of injury is based on the assumption that cases and controls are representative of their respective source populations (Mantel and Haenszel 1959). These source populations, particularly the ones that generate the controls, are not always explicitly defined. As a result, the selection of controls is often the most important and most challenging part in a case–control study. It is worth noting that in a case–control study, the investigator cannot directly measure the odds ratio of injury; the scientific foundation underlying the case–control design, however, allows the investigator to substitute the odds ratio of exposure for the odds ratio of injury.

In addition to the assumption regarding the selection of cases and controls, the injury under study is assumed to be a rare occurrence. The assumption of rarity is necessary for the estimated odds ratio to approximate relative risk directly (Cornfield 1951). Subsequent development in research methodology has virtually voided the assumption of rarity in the case–control design as statistical techniques for transforming odds ratios into risk ratios have become readily available (Zhang and Yu 1998; Spiegelman and Hertzmark 2005).

The studies of smoking and lung cancer by Levin et al. (1950), Wynder and Graham (1950), and Doll and Hill (1950) are widely perceived as the first formal applications of the case–control design in epidemiologic research (Morabia 2004). This perception is likely unduly influenced by the historic role these studies played in establishing the causality between smoking and lung cancer. The formal application of the case–control design in injury epidemiology can be traced back to a study of alcohol and driver injury published in the *Journal of American Medical Association* by Holcomb (1938). Use of alcohol by drivers had long been considered a causative factor for injurious motor vehicle crashes. But in the absence of empiric data about the threshold blood alcohol concentration (BAC) above which driving skills are impaired and about the dose–response relationship between BACs and crash risk, it was difficult in the court of laws to convict intoxicated drivers of driving under the influence of alcohol. To shed light on the causal role of alcohol in traffic injury, Holcomb compared alcohol measurement data for 270 drivers who were hospitalized for injury from motor

vehicle crashes and 1,750 drivers who were traveling in the approximate areas where the injurious crashes occurred. Based on this case-control analysis, Holcomb found that overall, 46% of the injured drivers and 12% of the drivers from the general driver population tested positive for alcohol, that the difference in the percentage distribution of BACs in the two groups of drivers increased as BAC increased in a dose-response fashion, and that the percentages of drinking drivers between the two groups were similar at BACs less than 0.05 g/dL, indicating that alcohol was not necessarily a significant causal factor for injurious crashes until BAC exceeded this threshold level. Further analysis of the alcohol data according to time of day, day of week, and driver age and sex confirmed that the difference in BACs between the two groups of drivers persisted (Holcomb 1938).

The case-control design has since been exquisitely refined and elegantly applied by injury epidemiologists. Matching on potential confounding factors between cases and controls is a commonly used technique for bias control. Nobody, however, has used this technique more cleverly and more effectively than Haddon and colleagues. In two landmark studies examining the relationship between BAC and fatal traffic injury (Haddon et al. 1961; McCarroll and Haddon 1962), the investigators incorporated an innovative matching technique into the case-control design. The first study was conducted in pedestrians, in which Haddon and colleagues selected four controls for each case of pedestrian fatalities from passing-by pedestrians who were interviewed at the same site where the injury case took place (Haddon et al. 1961). Furthermore, the site visit for each case was made on a subsequent date around the same time of day and on the same day of the week when the injury case occurred. Following the on-site interview, the investigators tested each control pedestrian for BAC using a breathalyzer. Through this design, Haddon and colleagues were able to match each case and corresponding controls on tempo-spatial confounding factors that might threaten the validity of the study findings. The second study was conducted in drivers, with cases being those who were fatally injured in crashes and controls being those who were selected at random traveling in the same direction of traffic around the same time of day and on the same day of the week as the injury case (McCarroll and Haddon 1962).

Subsequently, the tempo-spatial matching technique pioneered by Haddon and colleagues has been adopted by injury researchers in a variety of studies (Tsai et al. 1995; Li et al. 2001; Smith et al. 2001). This unique matching technique is effective for studying behavioral risk factors as well as other personal risk factors. For instance, in a case-control study conducted in general aviation pilots, Groff and Price (2006) used the tempo-spatial matching technique to examine the associations of pilot age, flight experience, certificate level, instrument rating, and testing performance with crash risk in degraded visibility, in which controls were selected through archived radar data and matched with cases on weather conditions, location, time, and rules of flight. A similar but more rigorous matching technique has been used in studies of combat sports injuries, such as injuries to professional boxers and mixed martial artists (Bledsoe et al. 2005, 2006), in which cases and controls are pair-matched on sex, body weight, and rules, in addition to tempo-spatial factors.

Case-Crossover Design

The case-crossover design is an epidemiologic method especially suited for studying proximal causes, or “triggers,” of acute medical conditions, such as injury, myocardial infarction, and asthma attack. It is a hybrid of the conventional case-control design and crossover design. Its distinct feature is that each case serves as its own control. Therefore, unlike other study designs, the case-crossover design does not require a separate group of study subjects to serve as referents or controls. The essence of the case-crossover design is a pair-wise comparison within the cases of the exposure prevalence between two time windows: the case time window and the control time window. The case time window refers to the defined time period immediately preceding the occurrence of the acute outcome

under study, such as injury. The control time window refers to the same time period experienced by the study subject on a previous date. For enhancing study power and controlling for biases, the investigator may use multiple control time windows from different dates, such as the dates 1 day, 1 week, and 1 month prior to the day of the index injury. The association between a putative cause and the risk of injury can be measured by the estimated odds ratio using the Mantel–Haenszel method for pair-matched case–control data. Use of cases as their own controls in the case–crossover design has two major advantages over other observational methods. First, self-matching strengthens the validity of the study results by eliminating many important confounding factors that are constant or stable within the follow-back period, such as genetic characteristics, age, sex, education, occupation, socioeconomic status, and chronic medical conditions. Second, self-matching increases the efficiency of the study by using the cases themselves as controls, instead of a separate group of study subjects.

The case–crossover design is widely credited to Maclure (1991), who applied the method in studying the transient effects on myocardial infarction onset of episodic exposures, such as physical exertion and anger. Most epidemiologists are unaware that the case–crossover design was originally developed by injury researchers. Wright and Robertson (1976) were among the first to use this method to study hazardous road characteristics as causes of fatal motor vehicle crashes. In this elegant study, Wright and Robertson compared road curvature, superelevation, and gradient measured in a 0.3 km section at 300 crash sites with those measured at 600 control sites (defined as the sites located 1.6 km upstream and 1.6 km downstream from the crash site). They found that crash sites were significantly more likely than control sites to have a road curvature greater than 6° and exhibit a downhill gradient (Wright and Robertson 1976). Wintemute and colleagues (1990) used the same design developed by Wright and Robertson in a study of occupant fatalities resulting from motor vehicle immersions and found that road curvature of 20° or greater was associated with a sixfold increased risk of drowning for crash victims. This finding prompted the researchers to call for the installation of guard rails in highly curved segments of roadways.

The case–crossover design as an effective method for injury research has become increasingly popular in the past 2 decades. Researchers have used this method to assess the causal role of cellular-telephone use, fatigue, alcohol, medication, and environmental hazards in a variety of injuries (Roberts et al. 1995; Vinson et al. 1995; Redelmeier and Tibshirani 1997; Sorock et al. 2004; Edmonds and Vinson 2007; Yang et al. 2011).

Cohort Design

“Cohort” as an epidemiologic concept refers to a group of persons who share certain characteristics or experiences. If the individuals in the designated group were born around the same time, they are usually called a *birth cohort*. The cohort design is a straightforward yet powerful observational research method. Its main feature is follow-up of a designated group of persons over a time period. If the designated group of persons is formed at the beginning of the follow-up (i.e., baseline) and no new study subjects are added to the group during the follow-up, then it is called a *fixed cohort*. On the other hand, if more individuals are intermittently added to the designated group of study subjects during the follow-up, it is called a *dynamic cohort*. Depending on the timing of the inception and follow-up of the designated group, cohort studies can be categorized into three groups (1) prospective cohort, (2) retrospective (or historical) cohort, and (3) ambidirectional cohort. In a prospective cohort study, follow-up proceeds from the present to the future. In a retrospective cohort study, follow-up is carried out for a past time period, usually based on archival data. In an ambidirectional cohort study, the follow-up starts in the past and ends in the future. A longitudinal study is a special form of cohort study in which individual study subjects are assessed periodically multiple times during the follow-up.

The cohort design has several advantages over the case–control design and the case-crossover design. Foremost is its ability to provide data for determining the incidence of injury per person-time of follow-up or per unit of exposure (e.g., mileage of travel and hours of work). The incidence data are crucial for researchers to assess the risk of injury in both absolute and relative measures. Other major advantages of the cohort design include the elimination of recall bias and ambiguity in temporality that often plague the case–control design and hinder causal inference, and the capacity to allow investigators to assess multiple outcomes. The most notable disadvantage of the cohort design is the lack of efficiency because it often requires following a large group of study subjects over an extended time period. Epidemiologic designs that combine the features of case–control studies and cohort studies, such as the nested case–control design and case–cohort design, are computationally more efficient than the cohort design. But, these hybrid methods are no substitute for the cohort design, nor do they exist independent of the cohort studies.

The cohort design has been used by injury epidemiologists in numerous studies (Gardner et al. 1999; Li et al. 2003; Batty et al. 2009). To determine the incidence and risk factors of sports injuries in adolescents, Knowles and colleagues (2006) followed prospectively for 3 years a cohort of 15,038 high school athletes in North Carolina. These study subjects were selected based on a two-stage cluster sampling scheme to represent the state high school athlete population in 12 varsity sports. Baseline data for each study subject were collected using an athlete’s demographic questionnaire. Follow-up data on exposure to sports practices and competitions and related injury were collected using a weekly participation form and an injury report form. A reportable injury was operationally defined as any injury sustained during sports participation that required medical attention or restricted activity. Exposure to sports was measured by “athlete-exposures,” defined as the cumulative count of preseason and regular season practices and competitions each athlete participated in during the follow-up period. Knowles and colleagues (2006) found that the overall incidence of injury for high school athletes was 2.1 per 100 athlete exposures, ranging from 1.0 for baseball players to 3.5 for football players and that players with a prior injury were nearly twice as likely as their counterparts without a history of injury to be injured during the follow-up. In light of their finding, Knowles et al. recommended that injury control strategies for high school athletes should emphasize primary prevention and adequate rehabilitation of the initial injury.

Pollack and colleagues (2007) used a retrospective cohort design to assess the causal relationship between body mass index (BMI) and occupational injury in aluminum manufacturing workers. Their study cohort consisted of 7,690 employees aged 18–65 years in 2002. Baseline BMI was calculated based on body weight and height data recorded during physical examination. These employees were followed through the company’s payroll records and injury surveillance system. With adjustment for demographic characteristics, smoking, job physical demand, and seniority, the investigators found that relative to employees with normal BMI (18.5–24.9 kg/m²), the risk of occupational injury increased by 26% for those who were overweight (BMI = 25.0–29.9 kg/m²), 54% for those who were obese (BMI = 30.0–39.9 kg/m²), and 121% for those who were morbidly obese (BMI ≥ 40.0 kg/m²). The finding underscores the necessity of integrating workplace health promotion programs, such as weight management, into occupational safety strategies.

Natural Experiment

Natural experiment refers to an observational study design in which the assignment of study subjects to different treatment groups or different levels of exposure to a putative cause is made by “natural” circumstances that are not manipulated by the investigator. In many situations, the assignment of exposure levels is also out of the control of the study subjects. Although a natural experiment is not a real controlled experiment, it is widely regarded as a form of research design that

produces stronger evidence for causal inference than do other observational studies. As an observational study design, natural experiment has several notable advantages over alternative designs, including practicality, efficiency, and effectiveness. It is particularly appealing when the randomized controlled trial design is not feasible due to ethical, economic, or other concerns. Since natural experiments are usually conducted in real-world settings, the observed “effects” should be interpreted as the *effectiveness*, rather than the *efficacy* as in randomized controlled trials. As an observational study design with some experimental features, natural experiment is sometimes called quasi experiment.

The term “natural experiment” is believed to be derived from John Snow’s investigation on the cholera outbreak in the Soho District of London in 1854. One of the studies Snow performed to prove that cholera was transmitted through contaminated drinking water was the comparison of mortality rates from cholera between two groups of residents who lived in the same area but used drinking water supplied by two different companies. While one company, Southwark and Vauxhall, obtained water from the heavily contaminated lower section of the Thames River, the other company, Lambeth, took water from the less polluted upper section of the river. By analyzing the population and mortality data, Snow was able to reveal an eightfold increased death rate from cholera for residents served by the Southwark and Vauxhall Company (Lilienfeld and Lilienfeld 1980).

The natural experiment design has long served as a valuable tool for injury research. Robertson and colleagues (1974) used the natural experiment design to assess the effect of television campaigns on seatbelt use. The study was done in a county of approximately 13,800 households that were serviced by a dual cable television system. About 6,400 households that chose to pay for the enhanced signal were on Cable A and about 7,400 households were on Cable B. Both cables provided the same channels and content. Although the assignment of households to one of the two cables was not random, residents living in the households on Cable A were similar to those on Cable B with regard to demographic characteristics and car ownership. Professionally developed messages and images promoting seatbelt use were shown for a 9-month period in Cable A only. Data on driver seatbelt usage and vehicle license plate numbers were collected through field observations in 14 carefully selected sites. Driver home addresses were identified by matching vehicle license plate numbers with the state department of motor vehicles data files. Exposure status to television campaign messages was then determined by linking driver names and addresses to the cable company’s consumer database. Based on over 10,000 observations, Robertson and colleagues found that seatbelt use rates in the two groups of drivers were virtually identical both before and after the television campaigns started. Their study provided the most compelling evidence that changing safety behavior through educational interventions alone represented an ineffective approach to injury control.

The natural experiment design is especially valuable in evaluation research on policy interventions because these interventions are rarely implemented in a randomized manner or in a way that is manipulated by the investigator. Sometimes, implementation of policy interventions may create conditions that approximate controlled experimentation. For instance, in 1995, the federal Department of Transportation initiated a program whereby alcohol testing was mandated for randomly sampled employees who were motor carrier drivers (operators of trucks with a gross vehicle weight rating of more than 26,000 pounds). There was no similar program relevant to car drivers. Thus, fatal multivehicle crashes involving a motor carrier and a passenger car present a unique natural experiment scenario for assessing the effectiveness of the mandatory alcohol testing program in reducing alcohol use by motor carrier drivers. The investigators capitalized on the tempo-spatial-matched multivehicle crash data and found that the mandatory random alcohol testing program resulted in a 23% reduction in alcohol use by drivers of large trucks involved in fatal crashes (Brady et al. 2009).

Future Directions

Epidemiologic methods have played an instrumental role in injury research, particularly in recognizing and defining injury as a public health problem, identifying, and understanding the myriad causal factors of injury, and developing and evaluating intervention programs to mitigate the risk and consequence of injury. At the same time, injury researchers have contributed substantially to the evolution of modern epidemiologic theory and methodology. Transformation of the injury research field to further improve safety and health will require researchers to continue strengthening the scientific foundation of injury epidemiology by embracing advances in bioinformatics, statistics, and other disciplines and by integrating their work into biomechanical engineering, emergency medical services, trauma care, and rehabilitation. New developments in information technology will have a profound impact on future research and practice of injury epidemiology. Data from electronic patient records and many other sources present a double-edged sword to epidemiologists. While the unprecedented amount, diversity, and complexity of data available may allow researchers to tackle questions that were preciously out of reach, making appropriate use of these data is often hampered by increasingly restrictive privacy and ethics rules, questionable data quality, and inadequate research methods. Injury epidemiologists need to develop innovative study designs and analytical methods to capitalize on the rapidly expanding databases to advance the field of injury research. Finally, given the complexity of injury causation and outcomes, it is imperative for epidemiologists to step out of their comfort zones to take an interdisciplinary approach to injury research by collaborating with biomedical scientists, clinicians, engineers, sociologists, and other professionals. An example of the benefit of such collaboration is illustrated by the research approach of Winston and colleagues (Winston et al. 1996, 2002; Durbin et al. 2003), where specialists in epidemiology, medicine, and biomechanics have combined forces to determine how children are killed in crashes and how their injuries can be prevented. Disciplinary integration is the most promising pathway to advancing the science and practice of injury epidemiology and prevention.

References

- Armstrong, D. B., Bauer, W. W., Dukelow, D. A., et al. (1945). Accident prevention – an essential public health service. *American Journal of Public Health*, 35, 216–218.
- Baker, S. P., & Haddon, W., Jr. (1974). Reducing injuries and their results: The scientific approach. *The Milbank Memorial Fund Quarterly. Health and Society*, 52(4), 377–389.
- Baker, S. P., O’Neill, B., & Karpf, R. S. (1984). *The injury fact book*. Lexington, MA: Lexington Books.
- Baker, S. P., Waller, A., Langlois, J. (1991). Motor vehicle deaths in children: geographic variations. *Accident Analysis and Prevention*, 23(1), 19–28.
- Batty, G. D., Gale, C. R., Tynelius, P., Deary, I. J., & Rasmussen, F. (2009). IQ in early adulthood, socioeconomic position, and unintentional injury mortality by middle age: A cohort study of more than 1 million Swedish men. *American Journal of Epidemiology*, 169(5), 606–615.
- Betz, M. E., & Li, G. (2005). Epidemiologic patterns of injuries treated in ambulatory care settings. *Annals of Emergency Medicine*, 46(6), 544–551.
- Bhopal, R. S. (2008). *Concepts of Epidemiology: Integrating the Ideas, Theories, Principles and Methods of Epidemiology*. 2nd edition. New York: Oxford University Press.
- Bills, C. B., & Li, G. (2005). Correlating homicide and suicide. *International Journal of Epidemiology*, 34(4), 837–845.
- Björnstig, U. (1992). Accidents in the north. Some aspects on snowmobile accidents and moose-car collisions. *Arctic Medical Research*, 51(Suppl 7), 56–58.
- Bledsoe, G. H., Li, G., & Levy, F. (2005). Injury risk in professional boxing. *Southern Medical Journal*, 98(10), 994–998.

- Bledsoe, G. H., Hsu, E. B., Grabowski, J. G., Brill, J. D., & Li, G. (2006). Incidence of injury in professional mixed martial arts competitions. *Journal of Sports Science and Medicine*, 5(CSSI), 136–142.
- Brady, J. E., Baker, S. P., DiMaggio, C., McCarthy, M. L., Rebok, G. W., & Li, G. (2009). Effectiveness of mandatory alcohol testing programs in reducing alcohol involvement in fatal motor carrier crashes. *American Journal of Epidemiology*, 170(6), 775–782.
- Cheng, D. (2010). Higher suicide death rate in Rocky Mountain states and a correlation to altitude. *Wilderness & Environmental Medicine*, 21(2), 177–178.
- Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, 11, 1269–1275.
- Dash, P. K., Zhao, J., Hergenroeder, G., & Moore, A. N. (2010). Biomarkers for the diagnosis, prognosis, and evaluation of treatment efficacy for traumatic brain injury. *Neurotherapeutics*, 7(1), 100–114.
- Dellinger, A. M., Langlois, J. A., Li, G. (2002). Fatal crashes among older drivers: decomposition of rates into contributing factors. *American Journal of Epidemiology*, 155, 234–241.
- Doll, R., & Hill, A. B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, 2, 739–748.
- Durbin, D. R., Elliott, M. R., & Winston, F. K. (2003). Belt-positioning booster seats and reduction in risk of injury among children in vehicle crashes. *JAMA*, 289(21), 2835–2840.
- Edmonds, J. N., & Vinson, D. C. (2007). Three measures of sleep, sleepiness, and sleep deprivation and the risk of injury: A case-control and case-crossover study. *Journal of the American Board of Family Medicine*, 20(1), 16–22.
- Gardner, L. I., Landsittel, D. P., & Nelson, N. A. (1999). Risk factors for back injury in 31,076 retail merchandise store workers. *American Journal of Epidemiology*, 150(8), 825–833.
- Gibson, J. J. (1961). The contribution of experimental psychology to the formulation of the problem of safety—a brief for basic research. In: Jacobs, H.H. *Behavioral Approaches to Accident Research*. New York: Association for the Aid of Crippled Children, pp. 77–89.
- Godfrey, E. S. (1937). Role of the health department in the prevention of accidents. *American Journal of Public Health and the Nation's Health*, 27(2), 152–155.
- Goldstein, G. P., Clark, D. E., Travis, L. L., & Haskins, A. E. (2011). Explaining regional disparities in traffic mortality by decomposing conditional probabilities. *Injury Prevention*, 17(2), 84–90.
- Gordon, J. E. (1949). The epidemiology of accidents. *American Journal of Public Health*, 39, 504–515.
- Gordis, L. (1996). *Epidemiology*. Philadelphia: WB Saunders Company.
- Grigorenko, E. L., De Young, C. G., Eastman, M., Getchell, M., Haeffel, G. J., Klinteberg, B., Koposov, R. A., Orelan, L., Pakstis, A. J., Ponomarev, O. A., Ruchkin, V. V., Singh, J. P., & Yrigollen, C. M. (2010). Aggressive behavior, related conduct problems, and variation in genes affecting dopamine turnover. *Aggressive Behavior*, 36(3), 158–176.
- Groff, L. S., & Price, J. M. (2006). General aviation accidents in degraded visibility: A case control study of 72 accidents. *Aviation, Space, and Environmental Medicine*, 77(10), 1062–1067.
- Haddon W., Jr. (1963). A note concerning accident theory and research with special reference to motor vehicle accidents. *Annals of the New York Academy of Sciences*, 107, 635–646.
- Haddon W., Jr. (1968). The changing approach to the epidemiology, prevention, and amelioration of trauma: the transition to approaches etiologically rather than descriptively based. *American Journal of Public Health and the Nation's Health*, 58(8), 1431–1438.
- Haddon, W., Jr. (1980). Advances in the epidemiology of injuries as a basis for public policy. *Public Health Report*, 95, 411–421.
- Haddon, W., Jr., Valien, P., McCarroll, J. R., & Umberger, C. J. (1961). A controlled investigation of the characteristics of adult pedestrians fatally injured by motor vehicles in Manhattan. *Journal of Chronic Diseases*, 14(6), 655–678.
- Haddon, W., Jr. (1972). A logical framework for categorizing highway safety phenomena and activity. *Journal of Trauma*, 12(3), 193–207.
- Haws, C. A., Gray, D. D., Yurgelun-Todd, D. A., Moskos, M., Meyer, L. J., & Renshaw, P. F. (2009). The possible effect of altitude on regional variation in suicide rates. *Medical Hypotheses*, 73(4), 587–590.
- Hill, A.B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.
- Holcomb, R. L. (1938). Alcohol in relation to traffic accidents. *JAMA*, 111(12), 1076–1085.
- Kearney, P. J., & Li, G. (2000). Geographic variations in crash risk of general aviation and air taxis. *Aviation, Space, and Environmental Medicine*, 71(1), 19–21.
- Knowles, S. B., Marshall, S. W., Bowling, J. M., Loomis, D., Millikan, R., Yang, J., Weaver, N. L., Kalsbeek, W., Mueller, F. O. (2006). A prospective study of injury incidence among North Carolina high school athletes. *American Journal of Epidemiology*, 164(12), 1209–1221.
- Kim, N., Mickelson, J. B., Brenner, B. E., Haws, C. A., Yurgelun-Todd, D. A., & Renshaw, P. F. (2011). Altitude, gun ownership, rural areas, and suicide. *The American Journal of Psychiatry*, 168(1), 49–54.

- King, B. G. (1949). Accident prevention research. *Public Health Reports*, 64(12), 373–382.
- Lawryniewicz, A. E., & Baker, T. D. (2005). Suicide and latitude in Argentina: Durkheim upside-down. *The American Journal of Psychiatry*, 162(5), 1022.
- Levin, M. L., Goldstein, H., & Gerhardt, P. R. (1950). Cancer and tobacco smoking: A preliminary report. *Journal of the American Medical Association*, 143, 336–338.
- Li, G., & Baker, S. P. (1996). Exploring the male-female discrepancy in death rates from bicycling injury: The decomposition method. *Accident Analysis and Prevention*, 28(4), 537–540.
- Li, G., Baker, S. P., & Frattaroli, S. (1995). Epidemiology and prevention of traffic-related injuries among adolescents. *Adolescent Medicine: State of the Art Reviews*, 6(2), 135–151.
- Li, G., Baker, S.P., Langlois, J.A., Kelen, G.D. (1998). Are female drivers safer? An application of the decomposition method. *Epidemiology*, 8(9), 379–384.
- Li, G., Baker, S. P., Smialek, J. E., & Soderstrom, C. A. (2001). Use of alcohol as a risk factor for bicycling injury. *JAMA*, 285(7), 893–896.
- Li, G., Baker, S. P., Grabowski, J. G., Qiang, Y., McCarthy, M. L., & Rebok, G. W. (2003). Age, flight experience, and risk of crash involvement in a cohort of professional pilots. *American Journal of Epidemiology*, 157(10), 874–880.
- Lilienfeld, A. M., & Lilienfeld, D. E. (1980). *Foundations of epidemiology* (2nd ed.). New York, NY: Oxford University Press.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Maclure, M. (1991). The case-crossover design: a method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, 133(2), 144–153.
- McCarroll, J. R., & Haddon, W., Jr. (1962). A controlled study of fatal automobile accidents in New York City. *Journal of Chronic Diseases*, 15(6), 811–826.
- Meuleners, L.B., Harding, A., Lee, A.H., Legge, M. (2006). Fragility and crash over-representation among older drivers in Western Australia. *Accident Analysis and Prevention*, 38(5), 1006–1010.
- Morabia, A. (Ed.). (2004). *History of epidemiologic methods and concepts*. Boston, MA: Birkhauser Verlag.
- Neves, F. S., Malloy-Diniz, L. F., Romano-Silva, M. A., Aguiar, G. C., de Matos, L. O., & Correa, H. (2010). Is the serotonin transporter polymorphism (5-HTTLPR) a potential marker for suicidal behavior in bipolar disorder patients? *Journal of Affective Disorders*, 125(1–3), 98–102.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd edition. New York: Cambridge University Press.
- Pollack, K.M., Sorock, G.S., Slade, M.D., Cantley, L., Sircar, K., Taiwo, O., Cullen, M.R. (2007). Association between body mass index and acute traumatic workplace injury in hourly manufacturing employees. *American Journal of Epidemiology*, 166(2), 204–211.
- Porta, M. (2008). *A dictionary of epidemiology* (5th ed.). New York, NY: Oxford University Press.
- Press, E. (1948). Epidemiological approach to accident prevention. *American Journal of Public Health*, 38, 1442–1445.
- Redelmeier, D. A., & Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. *The New England Journal of Medicine*, 336(7), 453–458.
- Roberts, I., Marshall, R., & Lee-Joe, T. (1995). The urban traffic environment and the risk of child pedestrian injury: A case-crossover approach. *Epidemiology*, 6(2), 169–171.
- Robertson, L. S., Kelley, A. B., O’Neill, B., Wixom, C. W., Eiswirth, R. S., & Haddon, W., Jr. (1974). A controlled study of the effect of television messages on safety belt use. *American Journal of Public Health*, 64(11), 1071–1080.
- Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, 104(6), 587–592.
- Runyan, C. W. (1998). Using the Haddon matrix: Introducing the third dimension. *Injury Prevention*, 4(4), 302–307.
- Smith, G. S., Keyl, P. M., Hadley, J. A., Bartley, C. L., Foss, R. D., Tolbert, W. G., & McKnight, J. (2001). Drinking and recreational boating fatalities: A population-based case-control study. *JAMA*, 286(23), 2974–2980.
- Sorock, G. S., Lombardi, D. A., Hauser, R., Eisen, E. A., Herrick, R. F., & Mittleman, M. A. (2004). A case-crossover study of transient risk factors for occupational acute hand injury. *Occupational and Environmental Medicine*, 61(4), 305–311.
- Spiegelman, D., & Hertzmark, E. (2005). Easy SAS calculations for risk or prevalence ratios and differences. *American Journal of Epidemiology*, 162(3), 199–200.
- Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology*, 133(7), 635–648.
- Tsai, Y. J., Wang, J. D., & Huang, W. F. (1995). Case-control study of the effectiveness of different types of helmets for the prevention of head injuries among motorcycle riders in Taipei, Taiwan. *American Journal of Epidemiology*, 142(9), 974–981.
- Vinson, D. C., Mabe, N., Leonard, L. L., Alexander, J., Becker, J., Boyer, J., & Moll, J. (1995). Alcohol and injury. A case-crossover study. *Archives of Family Medicine*, 4(6), 505–511.

- Waller, A. E., Baker, S. P., Szocka, A. (1989). Childhood injury deaths: national analysis and geographic variations. *American Journal of Public Health*, 79(3), 310–315.
- Winston, F. K., Schwarz, D. F., & Baker, S. P. (1996). Biomechanical epidemiology: A new approach to injury control research. *The Journal of Trauma*, 40, 820–824.
- Winston, F. K., Kallan, M. J., Elliott, M. R., Menon, R. A., & Durbin, D. R. (2002). Risk of injury to child passengers in compact extended-cab pickup trucks. *JAMA*, 287(9), 1147–1152.
- Wintemute, G. J., Kraus, J. F., Teret, S. P., Wright, M. A. (1990). Death resulting from motor vehicle immersions: the nature of the injuries, personal and environmental contributing factors, and potential interventions. *American Journal of Public Health*, 80(9), 1068–1070.
- Wright, P. H., Robertson, L. S. (1976). Priorities for roadside hazard modification. *Traffic Engineering*, 46(8), 24–30.
- Wynder, E., & Graham, E. (1950). Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma: A study of six hundred and eighty-four proved cases. *Journal of the American Medical Association*, 143, 329–336.
- Yang, Y. H., Lai, J. N., Lee, C. H., Wang, J. D., & Chen, P. C. (2011). Increased risk of hospitalization related to motor vehicle accidents among people taking zolpidem: A case-crossover study. *Journal of Epidemiology*, 21(1), 37–43.
- Zhang, J., & Yu, K. F. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*, 280(19), 1690–1691.
- Zwerling, C., Peek-Asa, C., Whitten, P. S., Choi, S. W., Sprince, N. L., Jones, M. P. (2005). Fatal motor vehicle crashes in rural and urban areas: decomposing rates into contributing factors. *Injury Prevention*, 11(1), 24–28.

Chapter 10

Qualitative Methods

Shannon Frattaroli

Introduction

Qualitative research methods have long been used in the social sciences. Until recently, researchers have made limited use of qualitative methods in the field of injury prevention and control. Through advancements in quantitative study design, secondary data systems, and statistical methods, researchers have advanced the field in important ways, as the numerous examples throughout this book demonstrate. However, there are limits to what can be measured and counted. Increased use of qualitative methods by injury researchers has the potential to grow the field in different ways than are possible through quantitative methods alone and to provide new insights into how injuries occur, how they can be prevented, and how their consequences can be minimized when they do occur.

In order to encourage injury prevention researchers to consider qualitative methods when designing studies and to promote understanding of the method and its relevance to the field, this chapter provides a general introduction to qualitative methods. For the purposes of this chapter, the phrase “qualitative research methods” refers to research that uses only qualitative data and is consistent with the methods literature on qualitative research traditions. Examples of research that includes qualitative data as part of a quantitative study abound. For example, including an open-ended question that asks respondents about smoke alarm maintenance as part of a large, population-based survey is very different from a study that uses focus groups to understand how people think about fire, how they assess their own risk of being injured in a fire, and the factors related to their decision to buy, install, and maintain a smoke alarm. In both scenarios, the research will yield qualitative data; however, the study aims for each design are very different, as are the processes of selecting the sample, developing the questions, and analyzing data.

This chapter considers research that uses qualitative data to conduct qualitative studies that seek to understand words and their meanings in relation to research questions as opposed to research that includes text data and an analysis approach that ultimately yields a count of those words. While qualitative survey questions and narrative analyses are valid applications of qualitative data for research, they are beyond the scope of this chapter.

Qualitative research methods are an engaging and rewarding style of research that is distinct from the quantitative paradigm that currently defines most injury research. In order to appreciate the approach, this chapter begins with an overview of the methods and a description of when qualitative research methods work best.

S. Frattaroli, PhD, MPH (✉)

Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
e-mail: sfrattar@jhsph.edu

Describing Qualitative Research Methods

At the most basic level, qualitative research methods provide a systematic way to collect, analyze, and present nonnumeric data about a particular subject. Qualitative data originate from a variety of sources: from spoken words through interviews and focus groups; from written words in the form of reports, memos, and meeting minutes; from words as field notes that capture observed events; and from words communicated through different media. Qualitative data can also include photographs and video, as well as artwork.

The reliance on words and other forms of expression for data is a fundamental characteristic of qualitative research methods. Central to qualitative research methods is the value of the emic perspective or an emphasis on understanding the experiences and perspectives of study participants. Individuals' views are both valid and valuable, and the study participants' reality is what qualitative researchers seek. With experiences and perspectives at the center of qualitative inquiry, study subjects are experts in the research process and have a role beyond mere sources of data. A basic tenet of qualitative research methods is that the researcher is an active learner in the process, which is a distinct role when compared to other forms of inquiry where the researcher approaches data as an expert testing narrowly constructed hypotheses and providing an etic interpretation of the results.

The process of gathering data is *iterative* when using qualitative methods. Initial data inform the researcher's understanding of the research questions and lead to an in-depth and nuanced understanding of the topic under study. This insight provided by the data often influences the data sources ultimately included in the data collection process and the researcher's approach to analysis. The iterative style results in a method that is *flexible* and *dynamic*. Researchers who use qualitative methods are accustomed to adjusting their data collection tools and sampling plans based on what they are learning in the field.

The dynamic nature of the method reflects the importance of *context* to qualitative methods, and the recognition that context is part of what explains the phenomena under study. The settings in which injuries happen and interventions take place are not something to be controlled through the study design or during analysis; rather, they are part of the answers to the research questions. Including context as part of data collection is consistent with the *inductive* approach of most qualitative research. Research that is based on observations and notes patterns, which in turn form the basis of theories, is a common approach when using qualitative methods. (Deductive theory-testing research may also use qualitative methods, although this approach is less common.) Inductive theory-building goals lead to results that are generalizable at the theoretical level. Unlike most injury research, population-level generalizability is not the goal in qualitative studies.

The above description of qualitative research methods represents a compilation of characteristics that will be considered more or less complete by others who use qualitative research methods. Several qualitative methodologists have written books that offer their own descriptions of the methods (Creswell 1998; Richards and Morse 2007; Yin 2009).

When to Use Qualitative Research Methods

Qualitative research methods are well suited to certain types of study aims. For example, when a topic needs to be explored because it lacks a developed literature or the existing literature is conflicting, qualitative methods can provide insight and clarity. Qualitative methods are a good fit for research that aims to understand and present a detailed view of a topic or when context is important to addressing the study aims. But the most important criterion for designing a study that uses

qualitative research methods is the nature of the research question (Creswell 1998) or what Richards and Morse (2007) term “methodological congruence.”

As the above description of qualitative research methods illustrates, the methods are not universally applicable. When designing a study, there is not a point at which the researcher makes a decision as to whether to employ a qualitative or quantitative methodology; rather, that decision is made when the aims and corresponding research questions are formulated. It is the questions underlying a research project that drive how well a project will be addressed by a methodology. Research that seeks to answer “how” and “what” questions often fit well with qualitative research methods (Creswell 1998). These questions tend to be more exploratory and less oriented toward assessing the effects of particular variables.

Qualitative Research Methods in Injury Research

The number of publications in the injury prevention literature that use qualitative research methods appears to be increasing. A search of the journal *Injury Prevention* provides evidence in support of this observation. During the 15-year period from 1996 to 2010, articles identified by a search of the words “qualitative methods” and confirmed to utilize qualitative methods by a review of the abstracts increased, as depicted in Fig. 10.1.

Many injury topics are included in the 56 papers identified, including unintentional (Green and Hart 1998) and intentional injuries (Johnson et al. 2004; Barkin et al. 1999), descriptive studies (Ashby et al. 2007; Christie et al. 2007) and evaluations (Shipton and Stone 2008; Steenbergen et al. 2001), and research about injuries within diverse populations (Rothe et al. 2009; Stallones et al. 2009). There are studies that focus on primary prevention (Gibbs et al. 2005) and on postinjury recovery (Sullivan et al. 2010).

The sources of qualitative data are also quite varied in this sample. Examples of data collected through in-depth interviews (Sullivan et al. 2010; Barkin et al. 1999), focus groups (Green and Hart 1998; Christie et al. 2007), narrative text (Lombardi et al. 2005), and media reviews (Clegg Smith et al. 2007; Pfeffer and Orum 2009; Smith et al. 2009) illustrate some of the different approaches to using qualitative data to inform our understanding of injuries. In addition, there is a qualitative research methods book that uses injury examples (Rothe 2000).

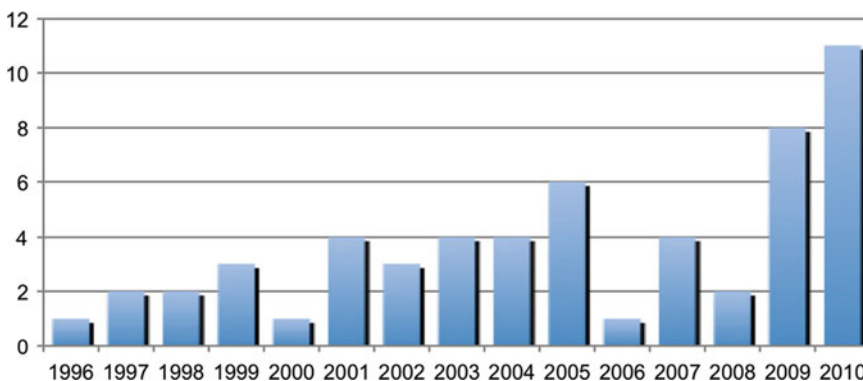


Fig. 10.1 Number of articles using qualitative research methods published in *Injury Prevention*

Tools and Rules in Qualitative Methods

As the review of qualitative papers in *Injury Prevention* revealed, qualitative data come from a variety of sources. Often, qualitative studies involve primary data collection, placing the researcher in the field where he/she is able to more directly witness the experiences of the study participants and better appreciate the perspectives of those providing the data. A discussion of select data collection tools and general rules for analyzing those data follow.

Data Collection

Interviews

The qualitative interview is often used in qualitative studies to gather data from an authoritative source on the study subject (Bernard 2006; Patton 2002; Richards and Morse 2007). Interviewees, also referred to as informants (or “key informants” to designate those interviewees with an extensive knowledge of the study subject), will often have information that is not otherwise available. For example, a New Zealand study used a structured open-ended telephone interview to survey plumbers about barriers to safe water temperatures (Jaye et al. 2001). The results yielded insights from a group uniquely qualified to speak to such barriers, and the interview method provided an effective means for accessing information that exists only with plumbers.

Typically, qualitative interviews are guided by an interview protocol that provides the interviewer with a general framework and key questions. An important characteristic of the qualitative interview is the open-ended nature of the questions and the expectation that some of the informants’ responses will be unanticipated by the interviewer. The qualitative interview is conversational in style, and the questions asked are often designed to encourage the informant to talk. In keeping with the conversational style, most interview protocols do not limit interviewers to the predetermined questions specified in the interview protocol. Follow-up questions that encourage informants to explain and offer examples are an important part of gathering data through this mechanism. Qualitative interviews are often described as in-depth and semistructured, although examples of structured interviews (such as the New Zealand plumber study) and unstructured interviews exist (Bernard 2006).

Interviews may take place in person, over the phone, or through videoconferencing. In-person interviews are generally viewed as preferable since they allow the interviewer to capture data about nonverbal cues and to more effectively build rapport. However, resource limitations may prohibit in-person data collection in some circumstances. Phone interviews can provide a less resource-intensive option, and recent advances in technology make Internet video interviews (through free-ware applications such as Skype) a viable option for conducting interviews.

Regardless of whether interviews take place in person or through some media, the question of whether to record the interview is one that should be considered during the study design, as it will have implications for the consent process, analysis plan, and the budget. An audio recording of interviews can provide a valuable record of the data collected. Any experienced interviewer will attest to the difficulty of capturing the rich detail communicated during an interview through notes and memory alone. In addition, audio recordings facilitate access to verbatim quotations to include in publications. However, with recording comes additional work, for to be useful as an analysis resource, the recording must be transcribed, which is a labor-intensive process. Transcription services exist, but their assistance comes at a price. And there are other potential downsides to recording interviews. Informants may be less forthcoming with information or less at ease if they are being recorded.

Recording devices can, and do, fail and recordings can be inadvertently erased. Note-taking is critical even when the recorder is on. Whether to record interviews is a decision that should be considered in the context of each individual research project. The sensitivity of the topic, available resources, and analysis plan are important considerations when deciding whether to ask informants for permission to record.

Focus Groups

The focus group is another tool for collecting qualitative data. A focus group is a staged gathering of individuals who are brought together to talk with a moderator and one another about a particular topic. Focus groups are used when a study will benefit from the insights gained through communication among the focus group participants (Kitzinger 1995). Focus group participants meet certain criteria related to the study aims. Groups may be homogenous or heterogeneous depending on the needs of the study or may involve a homogenous sample stratified by a particular variable. For example, in order to understand parents' knowledge, attitudes, and beliefs about booster seats, and to identify barriers to their use, Rivara et al. (2001) collected data from three parent focus groups. The study team recruited focus group participants from three different neighborhoods in order to achieve a socioeconomically diverse sample.

While the size of focus groups varies, 8–12 people in a group are generally recommended as this number provides a sufficient number of participants to allow for discussion without becoming unwieldy to manage (Stewart et al. 2007). As with interviews, a moderator uses a guide to direct group discussion of particular questions related to the study subject. Because of the number of participants involved in a focus group, a notetaker may supplement the audio/video recording of the group. Focus groups can provide an efficient means of collecting data (relative to interviews), but planning is needed in order to recruit participants and oversee the logistics associated with coordinating a group.

Observations

Qualitative researchers may observe people in the context of the subject under study or while participating in events related to the study aims. Field notes from observations constitute the resulting data. Such field notes provide the researcher's account of an event or interaction. This direct access complements other sources of data, such as the interview, by allowing the researcher to witness an event or interaction without the filter of a study participant.

Injury researchers have used field observation techniques for decades to collect information that can be observed, such as seat belt use (Wagenaar and Wiviott 1986; Vivoda et al. 2007) and child safety restraint use (Geddis 1979; Ebel et al. 2003). However, such studies ultimately seek to count the observed behaviors and do not invite the contextual detail that characterizes qualitative research. This type of observational study, while useful, is distinct from the way that observation and field notes are used in qualitative research. One example of direct observation in the qualitative genre is a study of a state law authorizing judges to order guns removed when issuing domestic violence protective orders (Frattaroli and Teret 2006). The researchers observed protective order hearings and noted how often and under what circumstances judges ordered guns removed. The resulting field notes complemented data collected from interviews and other documentary evidence.

While primary data collection is prominent in qualitative research, secondary qualitative data sources do exist. Qualitative data in the form of secondary data can take several forms.

Documents

Existing written words and expressions, in the form of reports, meeting minutes, memos, and training manuals, for example, can serve as sources of qualitative data. Yonas et al. (2009) analyzed art produced by youth participating in a community arts program to understand youth perspectives on violence and safety in their communities. To address the study aims, this analysis relied primarily (but not exclusively) on the artwork produced by the youth. Examples of research that uses documentary evidence in combination with other sources can also be found in the literature. An evaluation of a street outreach worker program used program reports, informational materials, and media coverage to supplement informant interview data about the street outreach workers' impact on youth violence (Frattaroli et al. 2010).

Media

As an institution, the media play a significant role in reporting on events and reflecting perspectives on how issues are framed, discussed, and debated in our society. How the media cover issues is a subject of research, and this line of scholarship includes injury examples such as motor vehicle crashes (Rosales and Stallones 2008), residential fires (Clegg Smith et al. 2007), and drinking and driving (Smith et al. 2009).

Sampling Qualitative Data

Who or what and how many comprise a sample when using qualitative research methods is a common question from those new to the method. The answer to the question is often unsatisfying, especially to those familiar with quantitative sampling strategies where there are rules and formulas to guide sampling. In qualitative research, sampling is purposeful: the cases selected, people included, or the documents gathered are determined according to the needs of the study. Sampling is driven by the goal of gathering enough "information-rich" data to address the study aims (Patton 2002). Deciding what should be included in such a sample is often an iterative process guided by sampling strategies, of which there are many. Patton (2002) describes nine purposeful sampling strategies. Sampling may be driven by a decision to achieve a diverse sample (maximum variation sampling), by existing theory (theory-based sampling), or by the sample participants themselves (snowball sampling) (Patton 2002). Snowball, or chain-referral sampling, is a strategy that relies on study participants to expand the sample based on their recommendations of who should be invited to participate in interviews or focus groups (Patton 2002). These strategies make sense, given that the goal of qualitative inquiry is to inform theory and not to reach conclusions that are generalizable to the population.

Data triangulation, or including data from different information sources, is a means of testing the consistency of the data across different data sources and reducing bias that may occur from an over-reliance on one data source, such as interviews (Patton 2002). Assessing whether data triangulation is possible and desirable for a particular study should happen during the study design phase. For some cases, data triangulation is not possible or does not comport with the study aims, such as when a qualitative study seeks to understand whether parents of children who died from an injury would want to collaborate with injury prevention professionals and how best to invite such relationships (Girasek 2005). The information sought resides with parents, cannot be observed, and was not otherwise documented.

In determining sample size, the concept of saturation (also referred to as redundancy) is important. Saturation is the point at which the researcher determines that the new data being collected are not

yielding qualitatively different information, and that additional data are not likely to meaningfully alter the study results (Richards and Morse 2007). Some argue that adhering to the principle of data saturation is unrealistic, given limited resources and the need to forecast how long a research project will take and how much it will cost. As an alternative, an informed estimate based on the literature, the researcher's experience, and an understanding of the information sources available can yield a minimum sample for planning purposes that can be expanded once in the field (Patton 2002).

Data Analysis

Qualitative research methods yield large volumes of data. Managing those data is an important task. Fortunately, the literature provides guidance on how to organize qualitative data (Richards 2005). A good system for managing data will facilitate access to materials, increase the likelihood that all data are included systematically throughout the analysis process, and create an accessible system for other researchers or future use. The importance of having an organized system for managing qualitative data cannot be overemphasized.

Tools for Analysis

Coding, the process of systematically assigning labels to segments of data so that segments with similar content can be connected across data, is a common early step in analyzing qualitative data (Patton 2002; Richards and Morse 2007). However, in order to develop a coding dictionary (a summary of codes, their meanings, and when they should be applied) and begin the coding process, the researcher must be familiar with the data. Familiarity comes from participating in the data collection processes (as an interviewer, observer, or focus group moderator, for example) and from reviewing the data. A researcher who knows the data, and has a good sense of the study aims and research questions driving the project, is well positioned to begin developing a coding dictionary and start coding.

Richards and Morse (2007) identify three types of coding: descriptive, topic, and analytic. The three codes are complementary and can all be applied to the same data since they serve different purposes. *Descriptive coding* is used to capture identifying information about the person, setting, or circumstances. These codes are the most straightforward of the three types and are used to compare how informants' characteristics correspond to aspects of the topic under study (Richards and Morse 2007). For example, if a researcher is interested in understanding from his/her data how parents living in urban areas talk about child pedestrian injury risks and how their views differ from rural parents' perspectives, having codes that identify informants as urban or rural would facilitate access to such responses.

Topic coding is a detailed review of the text that aims to identify categories of content in the data (Richards and Morse 2007). This level of coding involves more thought and interpretation than descriptive coding and engages the researcher in the meanings within rich, detailed text that is quite different from labeling demographic qualities through descriptive coding. A study exploring barriers to smoke alarm maintenance might use a topic code to capture data about the busy, chaotic nature of modern life. The importance of maintaining focus on the research questions driving the analysis is critical when doing topic coding.

The third category of codes is *analytic coding* (Richards and Morse 2007). Analytic codes, as the name suggests, go beyond labeling and grouping to begin the process of interpreting meaning behind coded text. These codes can be used to note when certain concepts are recurring and coming together to provide some insight that will inform the study aims. An analytic code from a study examining the culture of drinking and driving in a community might capture attitudes about business owners' responsibility to promote and assure responsible alcohol service and a role for server training.

Miles and Huberman (1994) have described many analytic tools in detail. The tools provide systematic ways to order and display qualitative data and to help identify and interpret findings. With few exceptions, qualitative studies in the injury prevention literature have relied on coding as the primary analysis strategy. One analytic tool discussed by Miles and Huberman (1994) is the *contact summary sheet*. Researchers use contact summary sheets to identify the salient points conveyed in an interview and, through this process, distill the key findings into a concise written form. These summary sheets can serve to remind researchers of the basic content of the corresponding interview and to keep the analysis focused on the study aims.

With coded data, or data that have been organized using other analytic tools, the researcher is able to identify themes or ideas about the data that cut across the individual data sources. The identification and building of these “common threads that run through the data” is *thematic analysis* (Richards and Morse 2007). Themes that result from this stage of the analysis serve as the core of findings in qualitative research and are used to develop theory and explanations in response to the study aims (Creswell 2003).

Another analytic approach that may be useful to injury researchers with an interest in qualitative methods is pattern matching. As described by Yin (2009), the goal is to identify similar patterns in the data across sources. These patterns provide the basis for understanding the topic under study. A more complex form of pattern matching known as explanation building pulls together patterns into a logical sequence that, taken together, forms an explanation for the question driving the analysis (Yin 2009).

Analyzing qualitative data involves interpretation. When the sources of data are people, they can participate in the analysis process through *member checking*. As the name implies, member checking involves informants in reviewing findings or draft publications for the purpose of providing feedback on the accuracy of those findings (Creswell 2003).

Qualitative Data Analysis Software

Qualitative software is available and commonly used to organize and manage data. These programs typically receive data in the form of text, pictures, or videos and provide a mechanism for the researcher to code the uploaded files electronically. These coded segments of text can be searched and presented together in an electronic report, facilitating the process of identifying and reading coded text across different sources of information. Some software can also allow for double coding of documents and score agreement between two coders. What qualitative software does not do is perform the thematic, pattern matching, or explanation building analysis functions previously described. Such processes require a level of human consideration unmatched by current programming. That limitation aside, the availability of software to facilitate the organization, management, and retrieval of qualitative data marks a significant contribution to qualitative research methods.

Rules of Analysis

Qualitative research methods involve concurrent data collection and analysis. By examining data as it becomes available, the researcher is able to exploit several key characteristics of the method. Qualitative research methods are iterative, and building on the expertise of participants and the information uncovered in the field to refine the data collection process is central to the method. As a study progresses, the ability of the researcher to ask more informed questions during interviews and observe events and interactions with greater purpose occurs by learning from the data. This evolution can only happen if analysis is occurring in parallel with data collection.

The concepts of data saturation and data triangulation were described earlier in this chapter in the context of sample size. However, they are also relevant for analysis. Throughout the analysis process, the researcher must assess whether the data are sufficient and, if not, how they can be expanded to address the study aims. Similarly, assuring there is triangulation of data sources when possible and assessing the quality of the various sources is part of data analysis.

The researcher has a prominent role in qualitative research methods, and there are multiple ways in which his/her decisions influence a study. Who is or is not included in a sample, what questions are asked of informants and how those questions are asked, how codes are applied and interpreted, and the extent to which rival explanations for an emerging theory are explored are some of the ways in which a researcher's decisions shape what data are collected and the approach to understanding those data. The concept of reflexivity is an important reminder to regularly assess these decisions with particular attention to understanding and minimizing researcher bias.

Quality of Qualitative Studies

The quality of research is often judged by two criteria: validity and reliability. Validity refers to the accuracy of the results obtained through the study design and research processes; reliability is the ability to repeat a study using the same procedures and tools to obtain the same result. As applied to qualitative research, these concepts provide useful reminders of the importance of building steps into the research process to increase the rigor and defensibility of research findings. Several of the rules previously discussed directly address validity. Using multiple sources of evidence and triangulating those sources reduces the risk that the findings will be overly influenced by a particular experience or perspective; member checking is a strategy for using informants to review the accuracy of the researcher's interpretation of data (Yin 2009). Additional discussion of validity and the strategies for addressing concerns about validity in qualitative studies is provided in several authoritative texts (Creswell 2003; Silverman 2005; Yin 2009).

Reliability, or the reproducibility of a study, requires that a record of the procedures originally used in designing and fielding the study exist and can provide a guide for replication. In order to increase the reliability of a qualitative study, documenting the study processes and decisions is essential. Various texts suggest different ways of tracking a qualitative research study, including maintaining a case study database (Yin 2009) or an audit trail (Richards and Morse 2007) that includes a record of all the study procedures and tools. However, given that much of the analysis process relies on interpretations of the researcher, the idea that a qualitative study could be replicated precisely is difficult to grasp.

Richards and Morse (2007) suggest focusing on five aspects of the research process to bolster the integrity of a qualitative research project. "Asking the right question, ensuring the appropriate design, making trustworthy data, building solid theory, and verification or completion" incorporate aspects of the research process from beginning to end (Richards and Morse 2007). It is through attention to these components of the research process that qualitative researchers can best assure that their methods are of high quality and the findings that result are defensible.

Traditions of Qualitative Research

"Qualitative research methods" is a phrase that includes the data collection and analysis processes previously described, and applies to several distinct study designs, sometimes referred to as traditions (Creswell 1998). Many qualitative research methods texts describe study design, data

collection, and data analysis in the context of an approach or tradition (Creswell 1998; Strauss 1987; Yin 2009). Creswell (1998) identifies five traditions: biography, phenomenology, grounded theory, ethnography, and case study. The five traditions have their roots in different fields, and each is responsive to a particular type of research goal. For example, the grounded theory approach originated in sociology and is used to “develop a theory grounded in data from the field” (Creswell 1998). Each tradition is associated with certain methods of data collection and analysis that are responsive to the types of questions that define each tradition. Among the five traditions Creswell identifies, two are particularly relevant for injury researchers: case study and ethnography. In addition, concept mapping (Burke et al. 2005) is an emerging approach for the field and will be included in this discussion.

The three traditions (case study, ethnography, and concept mapping) have different strengths and evolved for different purposes. Therefore, depending on the goals of a particular research study as articulated through the research questions, a decision is made about which tradition will provide the best structure for any given study. Yin (2009) defines the case study method as a way to investigate “a contemporary phenomenon in depth, and within its real life context, especially when the boundaries between phenomenon and context are not clearly evident.” He further specifies other characteristics of the method. Case studies “have more variables than data points; rely on multiple sources of evidence and triangulation; and benefit from the prior development of theoretical propositions to guide data collection and analysis” (Yin 2009).

Case studies are useful for studying policies, programs, and organizations and are commonly used in political science (George and Bennett 2004). They can be used to explain, explore, describe, or evaluate the case (or cases) under study (Yin 2009). An examination of prevention initiatives within one state’s fire service used the case study method. Consistent with the method, data collection drew from multiple sources (interviews, documents, and observations) and used explanation building to provide a comprehensive explanation of the fire service’s commitment to prevention (Frattaroli et al. 2011).

Qualitative studies in the tradition of ethnography are used to “describe and interpret a social group or system” (Creswell 1998). Ethnography is a method used by anthropologists, among others, and is characterized by extensive and intensive time in the field. Collecting data through observation is common; key informant interviews are also used in ethnographies. While ethnography is often used to study cultures outside one’s own country, examples from intentional injury prevention demonstrate that there is value in applying the ethnographic method within one’s own country. Urban ethnographer Elijah Anderson has written eloquently about urban life and the role of violence (1999), and his work demonstrates the strengths of qualitative methods and ethnography and how detailed, context-rich data can provide explanations and insights about an issue that has been well measured but not well understood.

Concept mapping developed in the mid-1980s as a tool for developing empirically based theory. It involves a structured process of collecting data from a target population and then involving participants in the analysis and interpretation of their data through a series of six predefined steps. The six steps include *preparation*, or gathering information about the topic from the literature and identifying a population to work with; *generation*, where participants brainstorm ideas about the topic under study; *structuring*, in which participants sort and prioritize the ideas generated; *representation*, inputting data into a computer and using software to map how the data gathered translate into conceptual computer-generated relationships; *interpretation*, receiving and processing feedback from participants about the resulting maps; and *utilization*, where the researchers review the results in relation to the original study aims (Burke et al. 2005). Researchers have used concept mapping to explore intimate partner violence (Burke et al. 2006; O’Campo et al. 2005), youth violence (Snider et al. 2010), and traumatic brain injury (Donnelly et al. 2005).

Future Directions

There is tremendous potential for qualitative research methods to add to the injury research literature and to contribute to the efforts of many to reduce the burden of injury among the world's populations. In considering how qualitative research methods can best contribute to the field, five ideas are immediately apparent. First, more completely incorporating both qualitative research tools (e.g., interviews, focus groups, coding) and the traditions (e.g., case study) into injury research would help to ground the research in the literature in which the methods developed. Greater attention to the methods in this way might help to quiet skeptics and raise the profile of qualitative methods as a legitimate form of inquiry for the field.

Second, in considering the needs of the injury research literature, attention to theory development is a component of the literature that, if strengthened, would certainly benefit the field. Haddon's work has served the field well for many decades and will continue to be foundational in analyzing the phases of injury events and the intervention points associated with each one. However, the science of injury prevention has expanded and evolved since those original contributions, and new theories and frameworks should be part of that expansion. Qualitative research methods are well suited to theory development, and with greater attention to this aspect of the methods, a conceptual breakthrough may result.

Third, the injury research field will benefit immensely by integrating qualitative and quantitative methods. Qualitative methods offer a depth that is not possible through a quantitative approach; quantitative studies provide a certainty about the extent to which the findings apply to a larger population. By bringing the strengths of each method to a particular research topic, a mixed methods approach can achieve richness and depth while also extending findings to a larger population. Increasingly mixed methods are being recognized as a strong research design, and there are excellent resources available to guide researchers through the process of developing a well-designed mixed methods study (Creswell 2003).

Fourth, qualitative research methods can be used to understand partnership and how best to maximize partnerships in translating research into policy and practice. Perhaps more than any other public health problem, injuries require multiple disciplines and effective working relationships between researchers and practitioners for effective interventions to be realized (Margolis and Runyan 1998). The diversity of partners in injury prevention efforts is an additional reason to bring research to understand partnerships. One example from the field provides a reminder of how many interests can be involved. Efforts to address house fires will benefit from frontline fire fighters and from engineers and business people who design and manufacture smoke alarms and sprinkler systems. Cigarettes and unsafe heating sources can cause house fires, and in order to address these sources of risk, aligning with tobacco control groups and utility providers may help to advance the issue. Partnership experiences can be studied, and lessons from those experiences can be learned and generalized into theories that will ultimately guide how best research and practice efforts can work to reduce the burden of injuries.

Finally, qualitative methods can be used to advance translational science. Injury research has a strong tradition of informing practice. Recent calls from leading research institutions for more attention to the science of translating research into practice represent an opportunity for injury prevention researchers (Yonas et al. 2011). Translation as a subject of inquiry is at the early stages of conceptual development, and the exploratory strength of qualitative research methods can contribute to uncovering what factors are associated with successful translation. Importantly, through a more deliberate understanding of how translation occurs and what factors influence the translation process, research findings will move more quickly into practice.

Conclusion

There are numerous applications to injury prevention for the type of research that fits under the umbrella of qualitative research methods. Using qualitative methods to understand the circumstances and culture surrounding an injury problem or to capture the impact of injuries on trauma survivors and their caregivers are two examples that represent the diversity of applications to the existing injury prevention literature. Research that addresses emerging themes within the scientific community, such as translation research, is another category of research that qualitative research methods can be used to explore and develop. There are countless other potential opportunities to advance injury research and practice by expanding the methods used by researchers to more fully incorporate qualitative research methods.

References

- Anderson, E. (1999). *Code of the street: decency, violence, and the moral life of the inner city*. New York: W.W. Norton.
- Ashby, K., Ozanne-Smith, J., & Fox, B. (2007). Investigating the over-representation of older persons in do-it-yourself home maintenance injury and barriers to prevention. *Injury Prevention*. doi:10.1136/ip.2006.012328.
- Barkin, S., Ryan, G., & Gelberg, L. (1999). What pediatricians can do to further youth violence prevention – a qualitative study. *Injury Prevention*. doi:10.1136/ip.5.1.53.
- Bernard, H. R. (2006). *Research methods in anthropology: qualitative and quantitative approaches*. Oxford, UK: AltaMira.
- Burke, J. G., O'Campo, P., Peak, G. L., Gielen, A. C., McDonnell, K. A., & Trochim, W. M. (2005). An introduction to concept mapping as a participatory public health research method. *Qualitative Health Research*, 15(10), 1392–1410.
- Burke, J. G., O'Campo, P., & Peak, G. L. (2006). Neighborhood influences and intimate partner violence: does geographic setting matter? *Journal of Urban Health*, 83(2), 182–194.
- Christie, N., Ward, H., Kimberlee, R., Towner, E., & Sloney, J. (2007). Understanding high traffic injury risks for children in low socioeconomic areas: a qualitative study of parents' views. *Injury Prevention*. doi:10.1136/ip.2007.016659.
- Clegg Smith, K., Cho, J., Gielen, A., & Vernick, J. S. (2007). Newspaper coverage of residential fires: an opportunity for prevention communication. *Injury Prevention*. doi:10.1136/ip.2006.013946.
- Creswell, J. W. (1998). *Qualitative inquiry & research design: choosing among five traditions*. Thousand Oaks, CA: Sage.
- Creswell, J. W. (2003). *Research design: qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.
- Donnelly, J. P., Donnelly, K., & Grohman, K. K. (2005). A multi-perspective concept mapping study of problems associated with traumatic brain injury. *Brain Injury*, 19(13), 1077–1085.
- Ebel, B. E., Koepsell, T. D., Bennett, E. E., & Rivara, F. P. (2003). Too small for a seatbelt: predictors of booster seat use by child passengers. *Pediatrics*, 111, e323–e327.
- Frattaroli, S., & Teret, S. (2006). Understanding and informing policy implementation: a case study of the domestic violence provisions of the Maryland Gun Violence Act. *Evaluation Review*, 30(3), 347–360.
- Frattaroli, S., Pollack, K. M., Jonsberg, K., Croteau, G., Rivera, J., & Mendel, J. S. (2010). Streetworkers, youth violence prevention, and peacemaking in Lowell, Massachusetts: lessons and voices from the community. *Progress in Community Health Partnerships*, 4(3), 171–179.
- Frattaroli, S., Gielen, A. C., Piver-Renna, J., Pollack, K. M., & Ta, V. M. (2011). Fire prevention in Delaware: a case study of fire and life safety initiatives. *Journal of Public Health Management and Practice*, 17(6), 492–498.
- Geddis, D. C. (1979). Children in cars. Results of an observational study in New Zealand. *New Zealand Medical Journal*, 90(649), 468–471.
- George, A. L., & Bennett, A. (2004). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.
- Gibbs, L., Waters, E., Sherrard, J., Ozanne-Smith, J., Robinson, J., Young, S., et al. (2005). Understanding parental motivators and barriers to uptake of child poison safety strategies: a qualitative study. *Injury Prevention*. doi:10.1136/ip.2004.007211.

- Girasek, D. C. (2005). Advice from bereaved parents: on forming partnerships for injury prevention. *Health Promotion Practice, 6*(2), 207–213.
- Green, J., & Hart, L. (1998). Children's views of accident risks and prevention: a qualitative study. *Injury Prevention*. doi:10.1136/ip4.1.14.
- Jaye, C., Simpson, J. C., & Langley, J. D. (2001). Barriers to safe hot tap water: results from a national study of New Zealand plumbers. *Injury Prevention, 7*(4), 302–306.
- Johnson, S. B., Frattaroli, S., Wright, J. L., Pearson-Fields, C. B., & Cheng, T. L. (2004). Urban youths' perspectives on violence and the necessity of fighting. *Injury Prevention*. doi:10.1136/ip.2004.005793.
- Kitzinger, J. (1995). Qualitative research. Introducing focus groups. *British Medical Journal, 311*(7000), 299–302.
- Lombardi, D. A., Pannala, R., Sorock, G. S., Wellman, H., Courtney, T. K., Verma, S., et al. (2005). Welding related occupational eye injuries: a narrative analysis. *Injury Prevention*. doi:10.1136/ip.2004.007088.
- Margolis, L. H., & Runyan, C. W. (1998). Understanding and reducing barriers to collaboration by academics with agencies and community organizations: a commentary. *Injury Prevention*. doi:10.1136/ip.4.2.132.
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. Thousand Oaks, CA: Sage.
- O'Campo, P., Burke, J., Peak, G. L., McDonnell, K. A., & Gielen, A. C. (2005). Uncovering neighborhood influences on intimate partner violence using concept mapping. *Journal of Epidemiology and Community Health, 59*(7), 603–608.
- Patton, M. Q. (2002). *Qualitative research & evaluation methods*. Thousand Oaks, CA: Sage.
- Pfeffer, K., & Orum, J. (2009). Risk and injury portrayal in boys' and girls' favourite television programmes. *Injury Prevention*. doi:10.1136/ip.2008.019539.
- Richards, L. (2005). *Handling qualitative data: a practical guide*. London: Sage.
- Richards, L., & Morse, J. M. (2007). *Read me first for a user's guide to qualitative methods*. Thousand Oaks, CA: Sage.
- Rivara, F. P., Bennett, E., Crispin, B., Kruger, K., Ebel, B., & Sarewitz, A. (2001). Booster seats for child passengers: lessons for increasing their use. *Injury Prevention, 7*(3), 210–213.
- Rosales, M., & Stallones, L. (2008). Coverage of motor vehicle crashes with injuries in U.S newspapers, 1999–2002. *Journal of Safety Research, 39*(5), 477–482.
- Rothe, J. P. (2000). *Undertaking qualitative methods: concepts and cases in injury, health and social life*. Edmonton: The University of Alberta Press.
- Rothe, J. P., Ozeovic, D., & Carroll, L. J. (2009). Innovation in qualitative interviews: "Sharing Circles" in a First Nations community. *Injury Prevention*. doi:10.1136/ip.2008.021261.
- Shipton, D., & Stone, D. H. (2008). The Yorkhill CHIRPP story: a qualitative evaluation of 10 years of injury surveillance at a Scottish children's hospital. *Injury Prevention*. doi:10.1136/ip.2008.018358.
- Silverman, D. (2005). *Doing qualitative research*. Thousand Oaks, CA: Sage.
- Smith, K. C., Twum, D., & Gielen, A. C. (2009). Media coverage of celebrity DUIs: teachable moments or problematic social modeling? *Alcohol and Alcoholism*. doi:10.1093/alcalc/agn006.
- Snider, C. E., Kirst, M., Abubakar, S., Ahmad, F., & Nathens, A. B. (2010). Community-based participatory research: development of an emergency department-based youth violence intervention using concept mapping. *Academic Emergency Medicine, 17*(8), 877–885.
- Stallones, L., Acosta, M. S., Sample, P., Bigelow, P., & Rosales, M. (2009). Perspectives on safety and health among migrant and seasonal farm workers in the United States and Mexico: a qualitative field study. *Journal of Rural Health, 25*(2), 219–225.
- Steenbergen, L. C., Kidd, P. S., Pollack, S., McCoy, C., Pigman, J. G., & Agent, K. R. (2001). Kentucky's graduated driver licensing program for young drivers: barriers to effective local implementation. *Injury Prevention*. doi:10.1136/ip.7.4.286.
- Stewart, D. W., Shamdasani, P. N., & Rook, D. W. (2007). *Focus groups*. Thousand Oaks, CA: Sage.
- Strauss, A. L. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.
- Sullivan, M., Paul, C. E., Herbison, G. P., Tamou, P., Derrett, S., & Crawford, M. (2010). A longitudinal study of the life histories of people with spinal cord injury. *Injury Prevention*. doi:10.1136/ip.2010.028134.
- Vivoda, J. M., Eby, D. W., St Louis, R. M., & Kostyniuk, L. P. (2007). A direct observation of nighttime safety belt use in Indiana. *Journal of Safety Research, 38*(4), 423–429.
- Wagenaar, A. C., & Wiviott, M. B. (1986). Effects of mandatory seat belt use: a series of surveys on compliance in Michigan. *Public Health Reports, 101*(5), 505–513.
- Yin, R. K. (2009). *Case study research: design and methods*. Thousand Oaks, CA: Sage.
- Yonas, M. A., Burke, J. G., Rak, K., Bennett, A., Kelly, V., & Gielen, A. C. (2009). A picture's worth a thousand words: engaging youth in CBPR using the creative arts. *Progress in Community Health Partnerships, 3*(4), 349–358.
- Yonas, M. A., Frattaroli, S., Lillier, K. D., Christiansen, A., Gielen, A. C., Hargarten, S. W., et al. (2011). Moving child and adolescent injury prevention and control into practice: a framework for translation. In: *Injury prevention for children and adolescents: integration of research, practice and advocacy*. American Public Health Association, in press.

Chapter 11

Environmental Determinants

Shanthi Ameratunga and Jamie Hosking

Introduction

In 2004, the World Health Organization (WHO) published a report estimating the proportion of a range of health outcomes that could be attributed to environmental factors, based on both available evidence and expert opinion. Globally, WHO estimated that 40% of road traffic injuries were attributable to environmental factors. Environmental factors were also responsible for 71% of unintentional poisonings, 26–31% of falls, 7% of fire-related injuries, 54–74% of drownings, 30% of other unintentional injuries, 30% of suicides, and 19% of injuries from interpersonal injuries. For the purposes of this exercise, WHO considered “environment” to cover physical, chemical, and biological factors external to the individual, but not social factors (Prüss-Üstün and Corvalán 2006).

While these estimates contain some uncertainty, they are sufficient to indicate that the environment plays a major role in the burden of injury. In addition, individuals are influenced not only by the physical environment, as addressed by the WHO report, but also by factors in the social environment. Following a discussion of the conceptual issues involved in defining environmental influences, this chapter discusses how social and physical environments contribute to injuries and their outcomes and describes how this interaction is depicted in conceptual models in the injury control field. It concludes with some examples of methods and approaches that have been used for researching and communicating findings on the environmental determinants of injury.

Defining Social and Physical Environments

One of the simplest definitions of environmental factors is “all that which is external to the human host” (Prüss-Üstün and Corvalán 2006). Practitioners in the injury prevention field – established on the principles of the epidemiological triad of host, agent/vehicle, and environment – generally distinguish the environment from the more proximal “objects” of agent or vehicle.

The environments of concern can be framed in many ways, from geospatial and physical terms, to broader social, political, economic, cultural, and regulatory terms. While these and many others

S. Ameratunga, MBChB, PhD (✉) • J. Hosking, MBChB, MPH
Section of Epidemiology and Biostatistics, School of Population Health,
University of Auckland, Auckland, New Zealand
e-mail: s.ameratunga@auckland.ac.nz; jamie.hosking@auckland.ac.nz

can be salient contextual factors for particular injury-producing situations, this family of factors is most commonly described in terms of social and physical environments.

According to this distinction, political, economic, cultural, and regulatory factors comprise aspects of the social environment. In some situations, the social environment may effectively be the sum of individual characteristics of a group of individuals. For example, the number of people who choose to drive determines traffic volumes, which influences the risk of collision and injury for each driver. However, other aspects of the environment are quite separate from individual attributes, such as policy and legislation. Thus, collisions and injuries for car drivers are influenced not only by characteristics of other individuals but also by legislation that limits vehicle speeds and by policies that promote alternatives to car use, thereby influencing mode choice and traffic volumes.

Many descriptions in the injury literature make the distinction between social and physical environments self-evident. For example, social networks are clearly a social phenomenon while road structures are physical attributes. However, some concepts combine both physical and social environmental characteristics. For example, the term “workplace” as a description could encompass both attributes relating to its physical location and built environment, as well as to relationships defined by a group of people, processes, and an organizational culture. In addition, attributes of the physical and social environments interact and may be interdependent. For example, high traffic danger in a neighborhood may lead to calls for local traffic calming measures to be instituted. While traffic calming is a modification of the physical environment, it may in turn affect the social environment. That is, by reducing vehicle speeds, traffic calming may help foster social networks and cohesion within a community (Appleyard and Lintell 1972) with increased pedestrian activity being an important pathway (Morrison et al. 2004). However, disadvantaged populations, who are typically at higher risk of road traffic injury, are also less likely to advocate for safety improvements (Roberts 1995), suggesting that traffic calming is more likely to be advocated for and introduced in communities that are already cohesive enough to demand its implementation. Characteristics of the social and physical environments may thus interact and be mutually reinforcing. As noted in the examples above, the outcome could lead to an increase or decrease in the potential for injury.

Although not commonly viewed as such, health-care systems are a component of the social environment. Gabbe et al. studied the effect of trauma registries by comparing trauma mortality in the state of Victoria in Australia, which has an organized regional trauma registry, with mortality in England and Wales, which do not. Mortality was significantly lower in Victoria, indicating that the presence of regionally organized trauma systems can be an important determinant of variations in trauma outcomes (Gabbe et al. 2011). Systematic reviews suggest that trauma systems are associated with trauma mortality rates that are 15% lower or more, although more work is needed to assess the effect of prehospital and postdischarge mortality (Mann et al. 1999).

Similarly, there are large differences in disability outcomes following injury for people living in high-income countries compared with those in less-resourced settings. Allotey et al. (2003) describe the highly disparate experiences and quality of life of paraplegics in Cameroon compared with Australia, differences which are influenced by health-care system resources as well as a multitude of societal factors. This illustrates the importance of considering the interaction of factors that relate to the provision of treatment and rehabilitation services to reduce impairment (often encompassed in the “medical” model of disability) with the social and physical environments that pose barriers that disadvantage people with physical impairments (Shakespeare and Watson 2001).

Social and Physical Environments in Injury Models

This section describes a series of conceptual models that have influenced the practice of injury control and explicitly incorporate the “environment” as a key or central domain.

Table 11.1 Opportunities for environmental interventions to reduce the burden of car crash injuries

Phase	Potential environmental interventions		
	Physical environment	Social environment	
		Laws and regulations	Others
Pre-event	Good road design and lighting	Legislated traffic speed limits	Population levels of car use
Event	Median barriers	Vehicle design regulations	Pricing of child restraints
Post-event	Access for emergency services	Regulation of medical practitioner competence	Presence of organized trauma systems

Haddon Matrix

The Haddon matrix, with its columns representing the epidemiological triad and rows representing pre-event, event, and post-event phases, has played an important part in our understanding of injury and its causes. The matrix helps identify environmental factors before, during, and after the event causing injury. Table 11.1 provides an example of how the environmental portion of this matrix may be used to identify environmental strategies to reduce the burden of car occupant injuries.

Haddon also proposed ten categories of injury countermeasures, which set out a temporally ordered sequence of approaches that focused on controlling, modifying, and interrupting the process of energy transfer from the hazard causing injury. Environmental factors are important aspects of many of these countermeasures. Examples provided by Haddon (1970) included the use of sidewalks and the phasing of pedestrian and vehicular traffic. Haddon (1980) also advocated that “passive” injury countermeasures – those that protect the individual without action on the part of that individual – were likely to be more successful in reducing the risks of injury than “active” countermeasures. This supports an emphasis on environmental measures for injury prevention, since environmental measures are generally “passive” interventions. For example, both provision of sidewalks and child pedestrian skills training are potential interventions for preventing child pedestrian injury, but the former is a “passive” environmental intervention, while the latter is an “active,” individually focused intervention.

Because environments are usually shared by many people (e.g., a roadway), environmental factors also have the potential to influence injury risk for many people simultaneously (Peek-Asa and Zwerling 2003). For example, the density of liquor outlets at a community level is associated with levels of alcohol consumption among community members (Connor et al. 2011; Gruenewald et al. 2002). Alcohol consumption is known to be a powerful risk factor for injury. Thus, regulating liquor outlet density has been identified as a potential strategy for controlling alcohol consumption and preventing injury at the community level. Environmental factors can thus have effects on a relatively large scale, such as at the likely negative effect on injury at the community level.

Systems Approaches and Injury Prevention

Systems approaches to managing human error offer important lessons for injury prevention. Just as the adage “an injury is no accident” (Doegge 1978) reminds prevention advocates that these events are largely predictable and therefore preventable, the systems approach to human error reminds us that errors are not random but rather occur typically as a result of a combination of “active failures” (Reason 2000). In this approach, the unsafe acts of people are recognized to occur within a system

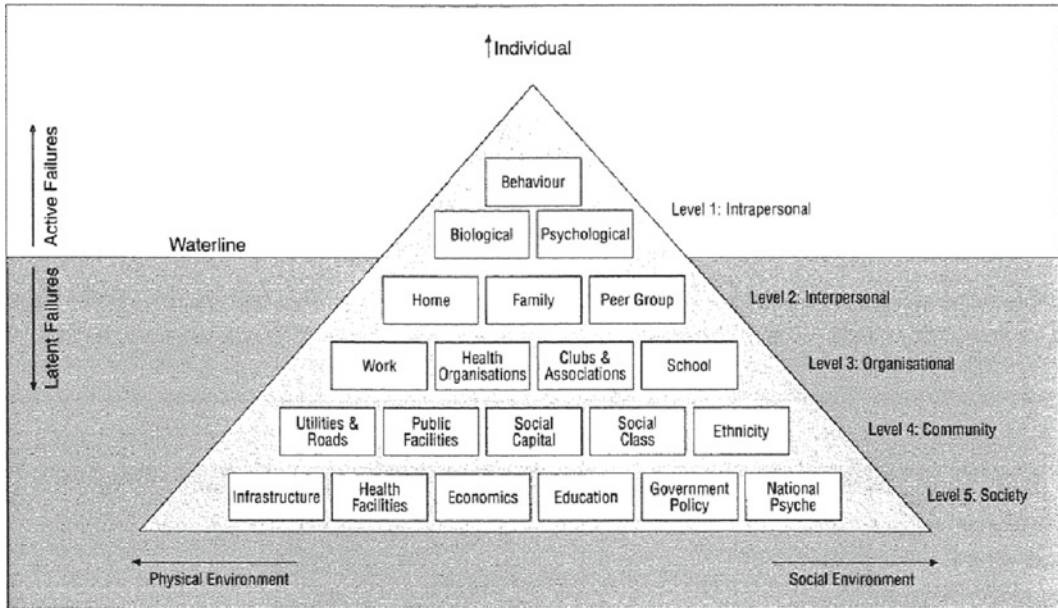


Fig. 11.1 The injury iceberg (*source*: Hanson et al. 2005)

which has “latent conditions” which increase the likelihood of human error or undermine system defenses designed to avoid damage once an error has occurred. This approach contrasts with the “person approach,” in which errors are seen as arising from individual factors such as carelessness, forgetfulness, inattention, or recklessness.

In a parallel vein, the application of the epidemiological triad to injury illustrated that attention to “host” factors (equivalent to the “person approach”) needs to be complemented by attention to aspects of the vehicle causing injury and to the environment. For example, the World Report on Road Traffic Injury Prevention notes that human error in complex traffic systems cannot be entirely eliminated, and therefore, transport systems must be designed to have an inbuilt tolerance of human error (Peden et al. 2004). The systems approach to human error can be considered in many ways to be the counterpart of the environmental approach to injury prevention.

Ecological Models

While the Haddon matrix emphasizes the importance of environmental influences on injury, ecological models give greater attention to the nature and characteristics of the environments involved. This particularly focuses on the physical and social attributes of where people live, including the environmental influences on their behavior. According to this view, individual behavior should not be studied without reference to its context, since individuals form part of a system in which all parts influence each other (Nurse and Edmondson-Jones 2007). While this interdependence is challenging to represent graphically, Hanson et al. have produced an ecological model for injury, which illustrates the different environmental levels that influence injury behaviors and injury risk (Fig. 11.1) (Nurse and Edmondson-Jones 2007). According to this model, the “micro” to “macro” environmental influences range in scale from interpersonal factors to organizational, community, and broader societal factors.

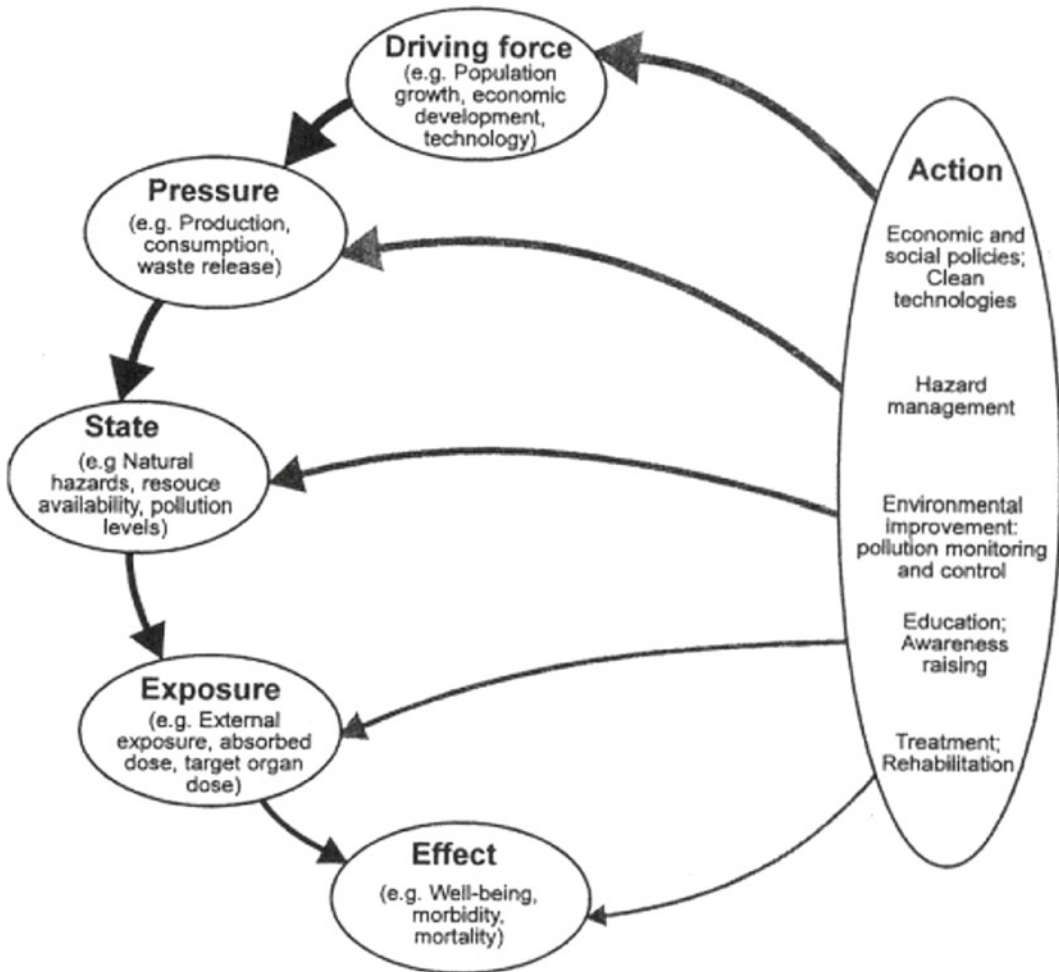


Fig. 11.2 The DPSEEA framework (source: Corvalán et al. 1999)

It is increasingly recognized that individuals who are at high risk of injury in one setting are often at high risk in other settings as well. For example, people who experience workplace injury are also at elevated risk of injury in the home (Smith 2003). In other words, an individual's injury risk in different environments may be correlated, and there is a degree of interdependence between the environmental levels represented in Fig. 11.1.

The DPSEEA framework (Fig. 11.2) represents factors at different levels of the environment (driving forces, pressures, states, and exposures) that lead to health effects. It also illustrates that actions can be taken at each of these levels to avoid adverse health effects (Corvalán et al. 1999). The multiple environmental levels in this framework have similarities to those depicted in the "injury iceberg" model (Fig. 11.1). While this framework has its origins in the wider environmental health field, it was used in the development of an indicator set for road traffic injury. The indicators identified represented the levels of "state" (age of vehicle fleet), "exposure" (distance traveled), and "effect" (mortality rates) (Farchi et al. 2006). The effects of policies that act at the levels of "driving forces" and "pressures" are typically monitored using these more distal indicators.

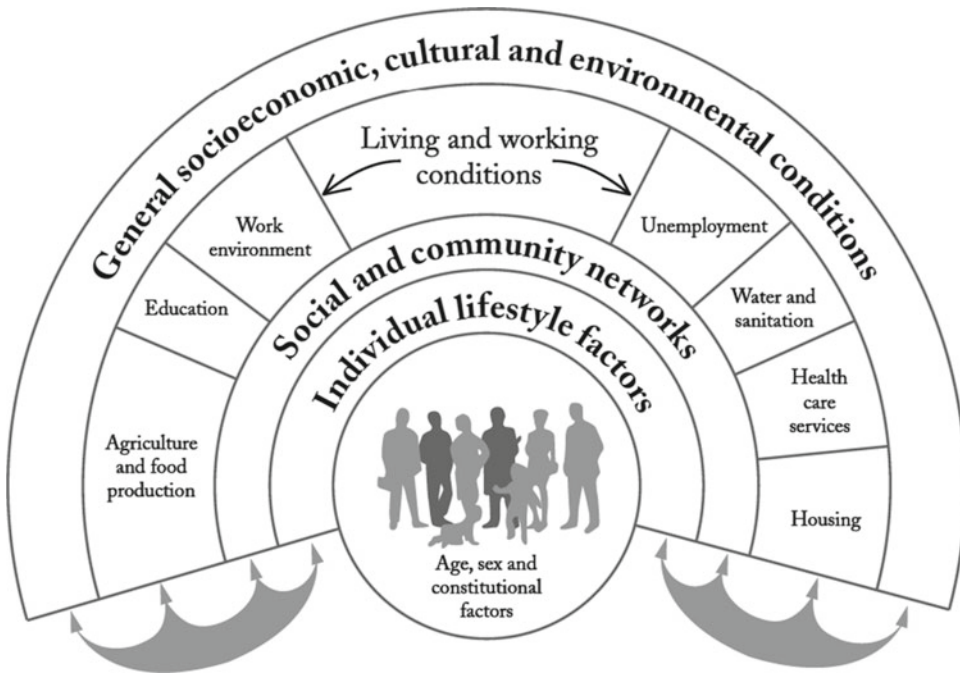


Fig. 11.3 The social determinants of health (source: Dahlgren and Whitehead 1993)

Social Determinants of Health and Injury

It is now accepted that many of the most powerful influences on health arise outside the health sector. While good health care is clearly an important requirement for healthy populations, so too are good education, adequate income, and good working conditions. The WHO Commission on Social Determinants of Health (CSDH) has called for action to improve health by improving daily living conditions – “the circumstances in which people are born, grow, live, work and age” (http://www.who.int/social_determinants/en/). At its heart, this is a call for healthier social environments.

The importance of daily living conditions and social determinants of health is commonly represented in the “rainbow” model of Dahlgren and Whitehead (Fig. 11.3) (Dahlgren and Whitehead 1993, 2007). The multiple levels of social influence on health and well-being are similar to Hanson et al.’s ecological model of injury (Fig. 11.1).

The “rainbow” model is also used to illustrate factors that underlie social inequalities in health, with social disadvantage generally associated with poorer health. Disadvantaged groups include those with lower incomes, lower social class, and lower educational status, with occupation, gender, and race or ethnicity being common dimensions of social disadvantage (CSDH 2008). As with other aspects of health, injury burden is generally higher among disadvantaged groups, though patterns vary somewhat by injury type (Cubbin and Smith 2002). In order to reduce these unfair differences, the CSDH has called on governments to act on the macrolevel social factors underlying social inequalities in health and injury by “tackling the inequitable distribution of power, money and resources” (CSDH 2008). These factors are represented in the outer ring of Fig. 11.3.

A WHO review of injury, social determinants, and equity concluded that addressing social determinants was an important injury prevention strategy. However, while there was some evidence that addressing social determinants could reduce injury, there remain many areas where the evidence is still very limited (Roberts and Meddings 2010).

Life Course Influences

Life course epidemiology is “the study of long-term effects on later health or disease risk of physical or social exposures during gestation, childhood, adolescence, young adulthood and later adult life” (Kuh et al. 2003). The life course approach recognizes that an individual’s current health and well-being are determined not merely by factors that are contemporaneous or in the recent past but by a whole lifetime of exposures, the accumulation of which leads to a person’s current health status. According to this view, it is not just a person’s current environment that influences their health and well-being but the sum of the different environments to which a person has been exposed over their lifetime and how those environments have changed over time (Kuh et al. 2003).

While the life course approach is not commonly applied in injury prevention, examples of long-term and intergenerational influences on injury risk abound (Hosking et al. 2011a). For example, children who are physically punished are at higher risk of being abused by their spouse in adulthood, and they also have a higher risk of abusing their own future children (Gershoff 2002). At a community level, alcohol availability correlates with youth drinking rates (Dent et al. 2005), while age at first drink is associated with subsequent unintentional injury risk (Hingson et al. 2000). Thus, environmental factors can have important influences not only on current injury risk but also long-term future risk.

A further example of the relevance of long-term effects on injury is provided by chronic exposures to hazards. Poisonings represent acute exposures to chemical hazards and are considered a form of injury. In contrast, chronic exposures to chemical hazards are typically considered to lie outside the injury field. Health effects from acute organophosphate exposure, for example, are more likely to be treated as an injury than chronic exposures to the same substance. The lack of recognition of the hazard as a chronic exposure is particularly unfortunate when preventive measures may be similar for both, such as the regulation of hazardous substances. For example, restrictions on the import of pesticides in Sri Lanka were associated with reductions in the suicide rate (Gunnell et al. 2007) but also have the potential to reduce the longer term risks of chronic exposure.

Drawing on several frameworks in the injury and environmental health literature, and the principles of the life course approach, the authors and colleagues have proposed a “lens and telescope” model which integrates an extended temporal dimension (spanning many years or generations) with an ecological perspective, according to which the contexts in which individuals live are critical, as are changes in those contexts over time (Hosking et al. 2011a). This model is designed to serve as a tool to identify injury intervention strategies that can also have cobenefits for other areas of health.

Key Aspects of the Physical Environment

Housing and Home Environment

The home constitutes an important level of the physical environment, with several injury determinants acting at the household level. It is also a leading setting in which injuries can occur, including falls, violence against children and intimate partners, burns, poisonings, and drowning.

Aspects of the home environment determine the risk of injuries in the home. For example, poisonings in children can be prevented through appropriate storage of chemical hazards, including the use of child-resistant containers (Peden et al. 2008; World Health Organization 2007). However, household injuries are also influenced by factors outside the home. For example, community liquor outlet density is associated with alcohol consumption, which raises the risk of several injury types in the home setting, including falls, child abuse, and intimate partner violence.

Home heating methods influence injury through their associated risk of burns. Some heating methods such as open fires are more likely to lead to burns (Peden et al. 2008). In comparison, heating methods such as heat pumps carry little risk of burns. Factors at other environmental levels – for example, policies to address climate change – may also have significant cobenefits including a favorable influence on household injury risk by affecting home heating choices. For example, a switch from an open fire (a very energy-inefficient heating mode) to a heat pump (a highly energy-efficient heating mode) could reduce both emissions and injury risk. Regulating the temperature of household hot water supplies can also help prevent burns (Peden et al. 2008), as well as reducing household energy use, which in turn can reduce energy-related greenhouse gas emissions.

Travel Environment

Road traffic injury is a leading cause of death globally. Both individual factors (e.g., age, sex, seat belt use, choice of travel mode) and vehicle factors (e.g., vehicle speed, safety design features, presence of seat belts, vehicle size and type) are important influences on injury (Peden et al. 2004). Factors in both the physical and social environments also play important roles, often mediated through effects on individual or vehicle factors. Several examples are discussed here.

At the home and neighborhood level, design of the built environment is important. Driveway design influences the likelihood of children being injured in driveways (Shepherd et al. 2010), while roadway design influences vehicle speed and thus both crash probability and severity (Ewing and Dumbaugh 2009). A systematic review of area-wide traffic calming measures, which generally involve modification of the neighborhood physical environment, found significant reductions in injury crashes (Bunn et al. 2003). At the city level, urban density is an important facet of the physical environment. Denser cities, which have greater concentrations of both people and potential destinations, are associated with less motorized travel (Newman and Kenworthy 1999), which in turn is associated with less road traffic injury (Litman and Fitzroy 2010). Aspects of the social environment, including urban planning and land use decisions, can influence the physical environment of cities.

National and regional factors can influence injury, such as through the effect of climate and weather conditions on the choice of travel mode (Humpel et al. 2004). Global climate change could therefore influence mode choice due to climatic alterations. The need for climate mitigation gives rise to policies that encourage the use of low-carbon transport modes, such as walking and cycling, which are also modes with relatively high injury risks for users; public transport is also relatively low-carbon but has very low injury risk for users (Beck et al. 2007). Thus, it is becoming increasingly clear that some transport and urban planning policies may be able to simultaneously mitigate climate change, reduce road traffic injury, and also address other public health problems such as physical inactivity and obesity (Roberts 2010).

Work Environment

The workplace as an important setting for injury is well recognized, and in some cases, the inherent hazards coincide in terms of setting. For example, road traffic injuries on highways account for 20% of US occupational deaths in 2009 (Bureau of Labor Statistics 2009). Other workplace

hazards, such as certain chemical hazards or items of equipment, are specific to a given occupation or industry.

Causes of workplace injuries may be attributed either to an individual or to workplace systems and environments. This is illustrated well by an example from a description of the causes of injuries in textile manufacturing in nineteenth-century France, where accidents (which were at times fatal) were said to be either “the fault of the manufacturer who has neglected to isolate or surround the dangerous parts of machines with a casing or screen” or the fault of workers, “especially children, who neglect to take safety measures” (Barss et al. 1998). While adequate safety training and education – as well as the selection of workers who are physically able to safely carry out the work in question – are important, this example illustrates the powerful but often underemphasized influence of environmental factors external to the individual. As acknowledged by systems approaches to human error, “we cannot change the human condition, but we can change the conditions under which humans work” (Reason 2000).

Environmental modification can be a powerful strategy for workplace injury prevention. Such modifications can address the vehicles through which the energy causing injury is transmitted, as in the case of rollover protection (such as roll bars) on tractors to prevent injuries in the agricultural setting (Rautiainen et al. 2008). Other environmental modifications may introduce or strengthen barriers that protect against injury. For example, the environmentally based approach entitled crime prevention through environmental design (CPTED) aimed to deter robberies and associated exposure to violence for retail staff. In this approach, retail workplaces were designed according to specific goals, including the management of customer visibility and the control of access to the store. Evaluations found substantial decreases in robberies, as well as fewer assaults on employees (Peek-Asa and Zwerling 2003).

Regulations mandating workplace safety requirements are an important way to ensure safe work environments. Regulation can be viewed as a modification of the social environment, and usually applies at the national or state level, though global agreements on workplace safety can also play an important role, especially in encouraging the uptake of appropriate regulations in less-resourced settings. In many low- and middle-income countries, the arrival of new and hazardous technologies can precede the arrival of appropriate safety regulations and can contribute to injury rates that are substantially higher than in high-income countries (Peden et al. 2008).

Children are at particularly high risk of workplace injury, as illustrated by the quotation above. While child labor continues to decline globally, an estimated 215 million children are still affected, with 115 million exposed to hazardous work. While global child labor standards have been developed under the auspices of the International Labour Organization (ILO), a substantial proportion of the world’s children are still not covered by these fundamental protections (International Labour Office 2010).

The urban built environment can also influence injury risk arising from workplace activities. Urban land use and planning instruments can be used to ensure that hazardous industries are separated from population concentrations, such as residential areas, by minimum distances (EPA Victoria 1990). Appropriate siting of hazardous activities can help reduce injuries in the event of catastrophic events such as explosions or unintentional emissions of hazardous substances.

Global Environmental Change

Climate change has been called the “biggest global health threat of the twenty-first century” (Costello et al. 2009) and is one of several planetary environmental limits that have been exceeded (Rockstrom et al. 2009). More extreme weather events are likely consequences of climate change, raising the risk of drownings associated with flooding from heavy precipitation, a risk compounded by rising sea levels. The high winds associated with severe storms are also an injury hazard. Extreme heat events

pose risks of heat stroke and also increase the likelihood of forest fires, with associated risks of burns to those exposed (IPCC 2007).

The pressing need to avoid the risks of climate change means that policy changes are required, particularly in sectors that are major and growing contributors to greenhouse gas emissions, such as the transport sector. In this sector, the imperative to reduce emissions means that measures to reduce private motorized travel (largely car use) will be needed, with many of these car trips replaced by public transport use, walking, and cycling (IPCC 2007). As walkers and cyclists are particularly vulnerable to road traffic injury, the effect of less heavily motorized (and thus safer) roads could be outweighed by the increased number of walkers and cyclists (Woodcock et al. 2009). Given the need to reduce transport emissions, safer walking and cycling environments will be needed to ensure that mode shift does not increase road traffic injury rates.

Key Aspects of the Social Environment

Socioeconomic Status and Ethnicity

The social gradient, according to which people with lower socioeconomic status also have worse health, is prevalent across most domains of health (CSDH 2008), and injury is no exception. Social class as measured by occupation has been associated with mortality from all injury types. Several studies have found higher socioeconomic status to be associated with lower risk of death by homicide. In the case of suicide, however, high rates have been inconsistently associated with socioeconomic status. For unintentional injuries, the relationship between high socioeconomic status and low injury rates is much more consistent (Cubbin and Smith 2002).

Many studies examine the relationship between individual socioeconomic status and injury risk. However, socioeconomic status at the neighborhood level is also associated with injury. For example, several studies of homicide have found higher rates in neighborhoods with lower socioeconomic status (Cubbin and Smith 2002). Associations between neighborhood-level socioeconomic status and health are partly due to the fact that low-socioeconomic-status neighborhoods contain many people with low individual-level socioeconomic status. However, even when individual socioeconomic status is controlled for, people living in low-socioeconomic-status neighborhoods have higher rates of homicide, road traffic injuries, and other unintentional injury. However, the association with suicide is inconsistent (Cubbin et al. 2000).

International studies demonstrate large wealth-based inequalities in disease and injury within and between countries. In both Uganda and Morocco, the under-5 mortality rate for the poorest quintile of the national population is at least 50% higher than in the richest quintile. However, infant mortality in the poorest quintile in Morocco is lower than even the richest quintile in Uganda (CSDH 2008).

Injury rates also vary substantially by racial and ethnic group. Racial and ethnic minorities, especially indigenous groups, experience worse health outcomes in many countries (Bramley et al. 2004). Injury is an important dimension of inequalities in general and is the third-highest contributor to racial disparities in mortality in the USA (Wong et al. 2002). For road traffic injury in the United States, blacks and Native Americans had higher death rates compared with whites, as well as higher rates of risk factors such as nonuse of a seat belt or child restraint (West and Naumann 2011). While racial and ethnic minority groups are often affected by many socioeconomic factors that are barriers to good health, such as low income, lack of health insurance, and lack of access to transport in emergencies, race and ethnicity are also independently associated with health measures such as quality of care (Kelley et al. 2005; Smedley et al. 2002).

As well as experiencing higher injury incidence, racial and ethnic minorities often receive poor-quality care from health services. A range of studies in the USA have found quality of trauma care

to be lower for black patients compared with white patients. Potential contributing factors include low levels of cultural competence among health-care workers, systems that require high levels of health literacy to access good-quality care, and a lack of quality assurance processes, leading to inappropriate variations in care (Hosking et al. 2011b).

While there is extensive evidence that disparities exist across a range of health domains, relatively little is known about the effectiveness of interventions to reduce these disparities. For example, a review of interventions to reduce racial and ethnic disparities in trauma care found no eligible studies (Hosking et al. 2011b). Similarly, in the case of child injury, very few interventions have been evaluated for their impact on disparities (Laflamme et al. 2010). This lack of evidence of which interventions effectively reduce injury disparities is an important research gap.

International Disparities in Injury

The global burden of injury falls largely on low- and middle-income countries. As recognized by the CSDH, a fundamental driver of international differences in health is the unequal distribution of power, money, and resources. Thus, while countries are defined according to geographic boundaries, and countries do differ in physical characteristics such as climate, terrain, and natural resources, the root causes of international disparities are largely social. This is further illustrated by the fact that the imbalance in wealth between rich and poor countries is not static but continues to grow (CSDH 2008).

Hazardous environments account for a large proportion of elevated injury rates in low- and middle-income countries. For children, particular hazards include open fires, unprotected heights, poor-quality building construction, lack of safe storage for chemical hazards, heavy traffic, and a lack of safe play spaces (Bartlett 2002). Once injuries have occurred, access to high-quality health services is also poorer. For example, death rates after hospitalization for burns in Nigeria were 27% in one study, compared with 1% in a similar study in Kuwait, although injury severity could account for some of this difference (CSDH 2008).

The CSDH has identified several policy influences that can contribute to international disparities in health and injury outcomes. For example, “outsourcing” of relatively hazardous jobs from developed to developing countries is a common practice. As poorer countries often have poorer working conditions and less stringent workplace health and safety regulations, this could lead to both more global workplace injuries and also to increased injury disparities between rich and poor countries. “Structural adjustment” policies, which were imposed by global financial institutions particularly upon poorer countries, reduced the role of the public sector, which has an important role in promoting health equity through social protection and other services. The CSDH has called for strengthened global governance and the adoption of health equity as a global goal, as preconditions for addressing global processes that perpetuate and aggravate global health disparities (CSDH 2008).

Family and Community-Level Social Environments

The family environment has an important influence on the well-being of its members. This is particularly obvious for children, but also applies to adults. For example, people with disabilities living in more supportive family environments report higher quality of life (Landolt et al. 2002; Warren et al. 1996). Home visiting interventions can have lasting impacts on injury risk factors such as reducing alcohol use by parents, and even for their children when they reach adolescence (Olds et al. 1998). Home visits are also effective interventions for reducing child maltreatment (Olds et al. 1997; World Health Organization 2007).

Community-level social factors can also interact to have important influences on injury risk. In the case of child abuse, rates are higher in communities with high levels of unemployment and poverty. Conversely, social networks and high levels of social capital at the community level appear protective against child abuse. Communities with higher levels of crime have more youth violence, while high crime rates are also associated with a lack of social capital (Krug et al. 2002).

Neighborhood-level characteristics may also interact with each other to influence injury risk. More “walkable” neighborhoods have higher levels of social capital, whereas streets with high traffic levels have weaker social connections between residents (Appleyard and Lintell 1972; Leyden 2003). Crime is also an important deterrent to walking in communities (Roman and Chalfin 2008). Some of these relationships may involve bidirectional effects and feedback loops. Implications for injury include the effect of community traffic volumes on child pedestrian injury and social capital on child abuse, for example (Krug et al. 2002; Peden et al. 2004).

Policies and Legislation

Policies and legislation occupy part of the “macro” level of the social environment. A government may form specific injury policies, such as developing an injury prevention strategy that sets out its plan for reducing the burden of injury. Such work may also be undertaken at the global level, as illustrated by a WHO report providing advice to national ministries of health describing which potential injury prevention interventions have been shown to be effective or ineffective (World Health Organization 2007). In this case, the health sector has an opportunity to directly intervene to reduce injury. However, policy drivers and options for injury prevention are broad with opportunities for intervention being substantial in many sectors, in and outside health – as is particularly obvious in the context of road traffic injury. For example, laws mandating motorcyclist helmet use have led to reductions in road traffic injury deaths (Peden et al. 2004; World Health Organization 2007). While the pathway of influence may be less direct, increasing fuel taxes in order to reduce greenhouse gas emissions from fuel consumption could also influence road traffic injuries through reductions in car travel (Roberts and Arnold 2007).

Legislation itself is inherently part of the social environment. The target of legislation, however, may be either individual or environmental. Both laws mandating seat belt use and speed limits aim to influence individual behavior to reduce road traffic injuries. However, the environment itself may also be targeted, such as by regulations requiring certain standards in roadway design. Both types of legislation may be useful in different contexts.

Examples of Research Methods Applied to Injury Environments

A range of different methods can be used for research on the environmental determinants of injury. This section provides a few examples of how particular methodological approaches have helped identify important environmental influences on the causes or consequences of injury. More detailed descriptions of research designs and methodologies are covered in other chapters of this book.

Cross-sectional and longitudinal research studies have elucidated significant racial and ethnic differences in the quality of trauma care. While race and ethnicity are individual-level factors, the poorer quality of care experienced by members of some racial and ethnic groups is influenced by environmental factors that are external to the individual who is receiving the care – including characteristics relating to staff and personnel or the organization and processes of the services involved. Studies examining racial and ethnic differences in the quality of trauma care commonly compare

individual-level outcomes, averaged across different racial and ethnic groups (Shafi et al. 2007). However, analysis also requires careful risk adjustment to establish whether racial and ethnic differences in outcome are attributable to baseline differences such as injury severity.

Qualitative research has provided appreciation of some of the factors that may contribute to these observations. A New Zealand project exploring the experiences of families of different ethnic groups when children were admitted to hospital following injury found important difficulties in the domains of patient–provider communication, negotiating the hospital environment, competing demands on families from factors outside the hospital, and issues of cultural competence (Arlidge et al. 2009). These findings were reinforced by a parallel set of interviews with stakeholders involved in injury control, who identified similar issues (Ameratunga et al. 2010). This approach illustrates the value of using information from different levels of the trauma care system to triangulate opportunities to improve trauma care. Such information can provide valuable input into the design of quality improvement interventions, including interventions to reduce ethnic disparities in the quality of care.

Ecological designs have been used to assess the effect of exposures measured at the level of the environment rather than the individual, such as laws and regulations. For example, an ecological design has been used to assess the effect of motorcycle helmet legislation on motorcycle rider death rates in different states (Branas and Knudson 2001). Important methodological challenges in these studies include the need to consider other potential causes that can explain differences between states or countries. Furthermore, associations detected at the ecological level are not directly applicable to the individual level. For example, a study that finds higher injury rates in more deprived neighborhoods should not be interpreted to mean that an individual with low socioeconomic status has an inherently higher injury risk (Robertson 2007). Disentangling the relative effect of individual- and neighborhood-level factors, such as socioeconomic status, requires factors to be measured at both levels and analyzed appropriately. For example, studies of the effects of liquor outlet density may analyze factors at the neighborhood level (e.g., liquor outlet density) as well as factors at individual level (e.g., alcohol consumption) (Gruenewald et al. 2002), and may control for the effects of both individual- and neighborhood-level SES (Connor et al. 2011). Multilevel modeling has been used to simultaneously address the effects of ecological and individual-level factors, such as for the effect of liquor outlet density on drinking and driving (Gruenewald et al. 2002).

A particularly useful tool for studying injury determinants that vary geographically is geographic information systems (GIS). GIS can be used to assess the effects on injury and its risk factors of proximity to certain locations, such as liquor outlets. It can also be used to construct maps that graphically represent differences in injury or the potential for injury due to differential exposure to risk. This is illustrated by a study that used GIS methods to investigate geographical differences in implementation of walking school buses, an intervention that aims to promote safe walking to school through parents chaperoning groups of children. They demonstrated that socioeconomically deprived areas in southern parts of Auckland, New Zealand, had very few walking school buses despite good uptake elsewhere in the city (Fig. 11.4) (Collins and Kearns 2005). Given the well-recognized association between low deprivation neighborhoods and child pedestrian injury, the distribution of the potentially effective intervention has the risk of increasing rather than decreasing existing disparities in this category of injury.

Using a novel approach to investigating driveway run-over injuries, a case–control study design was used to examine the influence of the built environment characteristics by extracting data from Google Earth (Google Inc., CA, USA). This enabled the researchers to identify several environmental characteristics associated with an increased risk of driveway run-overs, including long driveway length and lack of a separate pedestrian pathway for accessing the property (Shepherd et al. 2010).

Many evaluations of environmental interventions for preventing injuries employ observational designs. For example, a review of infrastructure for preventing cyclist injuries and crashes found 23

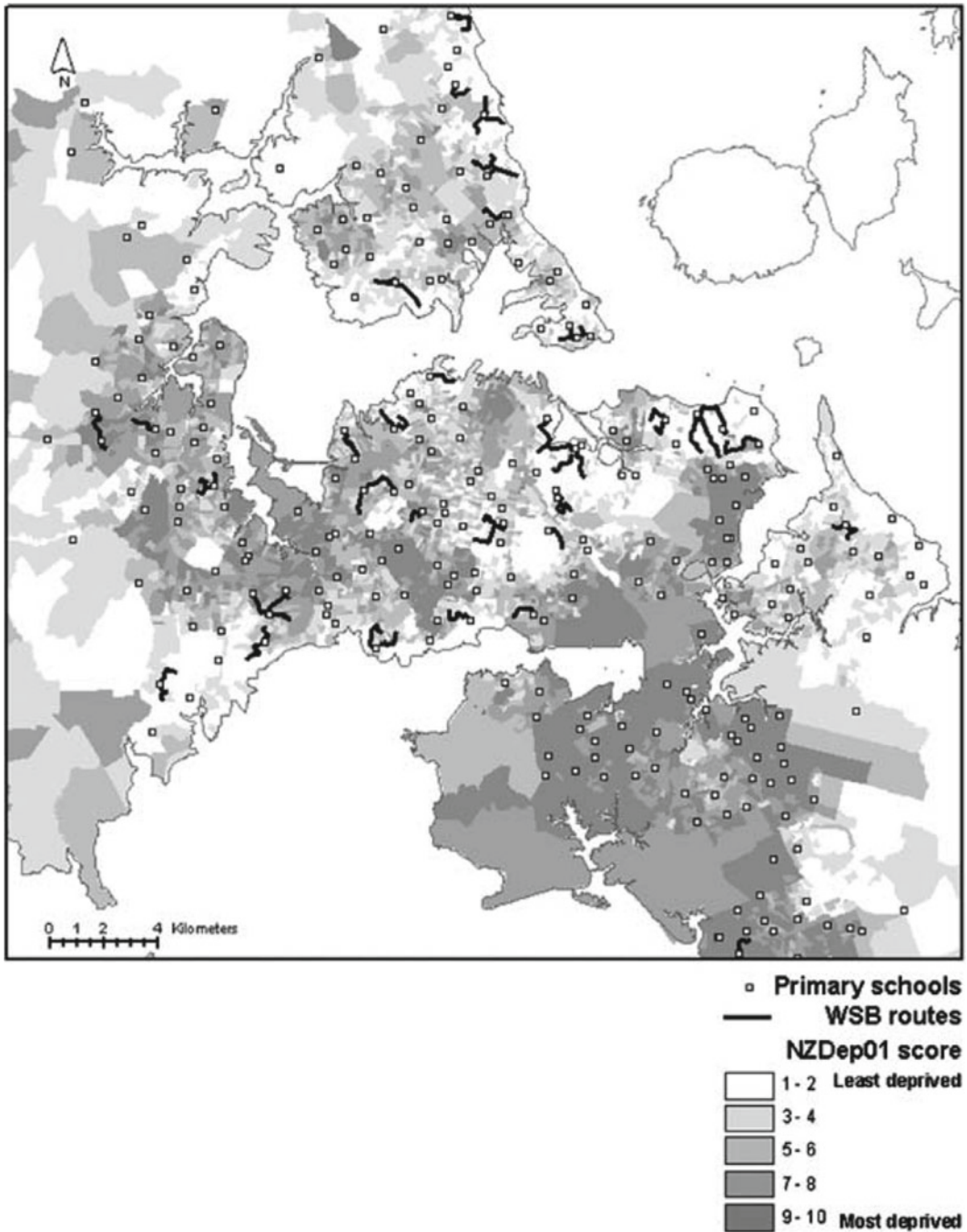


Fig. 11.4 Distribution of walking school bus (WSB) schemes in 2002 mapped against NZ Deprivation Index 2001 quintiles, Auckland metropolitan area (*source: Collins and Kearns 2005*)

studies, all of which were observational in design, including cross-sectional evaluations as well as before–after studies (Reynolds et al. 2009). However, randomized studies offer a stronger basis for making causal inferences. While randomizing individuals is less useful for evaluating environmental interventions, cluster randomized studies have been shown to be feasible for some injury prevention

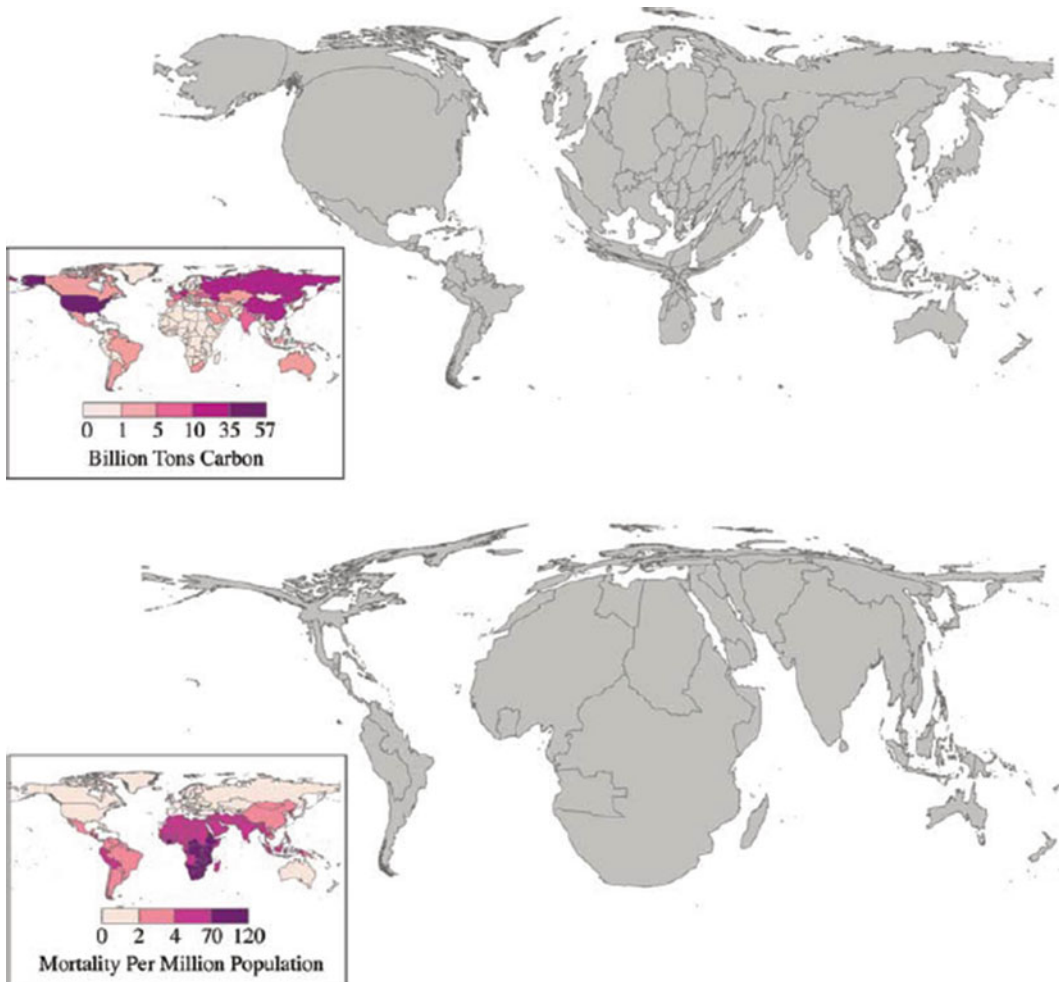


Fig. 11.5 Comparison of undepleted cumulative carbon dioxide (CO₂) emissions (by country) for 1950–2000 versus the regional distribution of four climate-sensitive health effects (malaria, malnutrition, diarrhea, and inland flood-related fatalities) (source: Patz et al. 2007; note: the size of each country is adjusted in proportion to its value for emissions (upper map) and for climate-related mortality (lower map))

interventions, such as provision of smoke alarms (DiGuseppi et al. 2002). Greater use of cluster randomized trials for appropriate injury prevention interventions would help strengthen the evidence base on environmental interventions.

The rationale for addressing the environmental determinants of injury and translation of research evidence to effective policy could be strengthened by utilizing more effective communication and presentation strategies. This has been demonstrated in other domains of health. For example, the Dartmouth Atlas Project in the USA (Dartmouth Atlas Working Group 2011) and the Social and Spatial Inequalities Group in the UK (Shaw et al. 2008; Social and Spatial Inequalities Group 2011) have produced compelling visual representations of geographic inequalities in health and health-care interventions within countries. Similarly, cartograms have been used to illustrate global disparities in injury and other outcomes by presenting world maps that are distorted according to international differences in the variable of interest (Dorling and Barford 2007). Figure 11.5 shows that countries in North America and Europe have very high levels of CO₂ emissions but very low levels of mortality from climate change, whereas for African countries, the reverse is true (Patz et al. 2007). Effectively

communicating research findings on the environmental determinants of injury as well as disparities in the distribution of this burden could stimulate more robust uptake of research findings and galvanize policy action.

Conclusions

The occurrence and consequences of injuries can be influenced by a range of factors in the physical and social environment. These factors operate at a number of levels, from interpersonal and family characteristics to national and global influences. Further research on the environmental determinants of injury will require complementing individually focused studies with other designs that incorporate effects at the ecological level. Addressing these important determinants will require research findings to be effectively communicated, and the development of interventions at the environmental level to complement those focused on other areas such as product designs and individuals. There are significant opportunities for cobenefits for injury prevention in relation to strategies that address climate change.

References

- Allotey, P., Reidpath, D., Kouamé, A., & Cummins, R. (2003). The DALY, context and the determinants of the severity of disease: an exploratory comparison of paraplegia in Australia and Cameroon. *Social Science & Medicine*, *57*, 949–958.
- Ameratunga, S., Abel, S., Tin Tin, S., Asiasiga, L., Milne, S., & Crengle, S. (2010). Children admitted to hospital following unintentional injury: perspectives of health service providers in Aotearoa/New Zealand. *BMC Health Services Research*, *10*, 333.
- Appleyard, D., & Lintell, M. (1972). The environmental quality of city streets: the residents' viewpoint. *Journal of the American Planning Association*, *38*, 84–101.
- Arlidge, B., Abel, S., Asiasiga, L., Milne, S. L., Crengle, S., & Ameratunga, S. N. (2009). Experiences of whanau/families when injured children are admitted to hospital: a multi-ethnic qualitative study from Aotearoa/New Zealand. *Ethnicity & Health*, *14*, 169–183.
- Barss, P., Smith, G. S., Baker, S. P., & Mohan, D. (1998). *Injury prevention: an international perspective. Epidemiology, surveillance, and policy*. New York: Oxford University Press.
- Bartlett, S. N. (2002). The problem of children's injuries in low-income countries: a review. *Health Policy and Planning*, *17*, 1–13.
- Beck, L. F., Dellinger, A. M., & O'Neil, M. E. (2007). Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. *American Journal of Epidemiology*, *166*, 212–218.
- Bramley, D., Hebert, P., Jackson, R., & Chassin, M. (2004). Indigenous disparities in disease-specific mortality, a cross-country comparison: New Zealand, Australia, Canada, and the United States. *The New Zealand Medical Journal*, *117*, U1215.
- Branas, C. C., & Knudson, M. M. (2001). Helmet laws and motorcycle rider death rates. *Accident; Analysis and Prevention*, *33*, 641–648.
- Bunn, F., Collier, T., Frost, C., Ker, K., Roberts, I., & Wentz, R. (2003). Area-wide traffic calming for preventing traffic related injuries. *Cochrane Database of Systematic Reviews*, (1), CD003110.
- Bureau of Labor Statistics. (2009). Census of fatal occupational injuries. United States Department of Labor. <http://www.bls.gov/iif/oshcfoi1.htm>.
- Collins, D. C., & Kearns, R. A. (2005). Geographies of inequality: child pedestrian injury and walking school buses in Auckland, New Zealand. *Social Science & Medicine*, *60*, 61–69.
- Connor, J. L., Kypri, K., Bell, M. L., & Cousins, K. (2011). Alcohol outlet density, levels of drinking and alcohol-related harm in New Zealand: a national study. *Journal of Epidemiology and Community Health*, *65*(10), 841–846.
- Corvalán, C. F., Kjellstrom, T., & Smith, K. R. (1999). Health, environment and sustainable development: identifying links and indicators to promote action. *Epidemiology*, *10*, 656–660.

- Costello, A., Abbas, M., Allen, A., Ball, S., Bell, S., Bellamy, R., et al. (2009). Managing the health effects of climate change: Lancet and University College London Institute for Global Health Commission. *Lancet*, 373, 1693–1733.
- CSDH. (2008). *Closing the gap in a generation: health equity through action on the social determinants of health*. Final report of the Commission on Social Determinants of Health. Geneva: World Health Organization.
- Cubbin, C., & Smith, G. S. (2002). Socioeconomic inequalities in injury: critical issues in design and analysis. *Annual Review of Public Health*, 23, 349–375.
- Cubbin, C., LeClere, F. B., & Smith, G. S. (2000). Socioeconomic status and injury mortality: individual and neighbourhood determinants. *Journal of Epidemiology and Community Health*, 54, 517–524.
- Dahlgren, G., & Whitehead, M. (1993). *Tackling inequalities in health: what can we learn from what has been tried?* Working paper prepared for the King's Fund International Seminar on Tackling Inequalities in Health. Oxford: Ditchley Park.
- Dahlgren, G., & Whitehead, M. (2007). *European strategies for tackling social inequities in health: levelling up Part 2*. Copenhagen: World Health Organization.
- Dartmouth Atlas Working Group. (2011). The Dartmouth Atlas of Health Care. The Trustees of Dartmouth College. <http://www.dartmouthatlas.org/>.
- Dent, C. W., Grube, J. W., & Biglan, A. (2005). Community level alcohol availability and enforcement of possession laws as predictors of youth drinking. *Preventive Medicine*, 40, 355–362.
- DiGiuseppi, C., Roberts, I., Wade, A., Sculpher, M., Edwards, P., Godward, C., et al. (2002). Incidence of fires and related injuries after giving out free smoke alarms: cluster randomised controlled trial. *BMJ*, 325, 995.
- Doegge, T. C. (1978). An injury is no accident. *New England Journal of Medicine*, 298, 509–510.
- Dorling, D., & Barford, A. (2007). Shaping the world to illustrate inequalities in health. *Bulletin of the World Health Organization*, 85, 890–893.
- EPA Victoria. (1990). *Recommended buffer distances for industrial residual air emissions* (EPA Publication No. AQ 2/86). Melbourne: Environment Protection Authority Victoria.
- Ewing, R., & Dumbaugh, E. (2009). The built environment and traffic safety: a review of empirical evidence. *Journal of Planning Literature*, 23, 347–367.
- Farchi, S., Molino, N., Giorgi Rossi, P., Borgia, P., Krzyzanowski, M., Dalbokova, D., & Kim, R. (2006). Defining a common set of indicators to monitor road accidents in the European Union. *BMC Public Health*, 6, 183.
- Gabbe, B. J., Biostat, G. D., Lecky, F. E., Bouamra, O., Woodford, M., Jenks, T., et al. (2011). The effect of an organized trauma system on mortality in major trauma involving serious head injury: a comparison of the United Kingdom and Victoria, Australia. *Annals of Surgery*, 253, 138–143. doi:10.1097/SLA.1090b1013e3181f6685b.
- Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: a meta-analytic and theoretical review. *Psychological Bulletin*, 128, 539–579.
- Gruenewald, P. J., Johnson, F. W., & Treno, A. J. (2002). Outlets, drinking and driving: a multilevel analysis of availability. *Journal of Studies on Alcohol*, 63, 460–468.
- Gunnell, D., Fernando, R., Hewagama, M., Priyangika, W. D., Konradsen, F., & Eddleston, M. (2007). The impact of pesticide regulations on suicide in Sri Lanka. *International Journal of Epidemiology*, 36, 1235–1242.
- Haddon, W., Jr. (1970). On the escape of tigers: an ecologic note. *American Journal of Public Health and the Nation's Health*, 60, 2229–2234.
- Haddon, W., Jr. (1980). Advances in the epidemiology of injuries as a basis for public policy. *Public Health Reports*, 95, 411–421.
- Hanson, D., Hanson, J., Vardon, P., McFarlane, K., Lloyd, J., Muller, R., et al. (2005). The injury iceberg: an ecological approach to planning sustainable community safety interventions. *Health Promotion Journal of Australia*, 16, 5–10.
- Hingson, R. W., Heeren, T., Jamanka, A., & Howland, J. (2000). Age of drinking onset and unintentional injury involvement after drinking. *JAMA*, 284, 1527–1533.
- Hosking, J., Ameratunga, S., Morton, S., & Blank, D. (2011a). A life course approach to injury prevention: a “lens and telescope” conceptual model. *BMC Public Health*, 11, 695.
- Hosking, J. E., Ameratunga, S. N., Bramley, D. M., & Crengle, S. M. (2011b). Reducing ethnic disparities in the quality of trauma care: an important research gap. *Annals of Surgery*, 253, 233–237.
- Humpel, N., Owen, N., Iverson, D., Leslie, E., & Bauman, A. (2004). Perceived environment attributes, residential location, and walking for particular purposes. *American Journal of Preventive Medicine*, 26, 119–125.
- International Labour Office. (2010). *Accelerating action against child labour: global report under the follow-up to the ILO Declaration on Fundamental Principles and Rights at Work*. Geneva: International Labour Office.
- IPCC. (2007). *Climate Change 2007: Synthesis Report*. New York: Cambridge University Press.
- Kelley, E., Moy, E., Stryer, D., Burstin, H., & Clancy, C. (2005). The national healthcare quality and disparities reports: an overview. *Medical Care*, 43, 13–18.
- Krug, E. G., Dahlberg, L. L., Mercy, J. A., Zwi, A. B., & Lozano, R. (Eds.). (2002). *World report on violence and health*. Geneva: World Health Organization.

- Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J., & Power, C. (2003). Life course epidemiology. *Journal of Epidemiology and Community Health, 57*, 778–783.
- Laffamme, L., Hasselberg, M., & Burrows, S. (2010). 20 Years of research on socioeconomic inequality and children's – unintentional injuries understanding the cause-specific evidence at hand. *International Journal of Pediatrics, 2010*, 819687.
- Landolt, M. A., Grubenmann, S., & Meuli, M. (2002). Family impact greatest: predictors of quality of life and psychological adjustment in pediatric burn survivors. *The Journal of Trauma, 53*, 1146–1151.
- Leyden, K. M. (2003). Social capital and the built environment: the importance of walkable neighborhoods. *American Journal of Public Health, 93*, 1546–1551.
- Litman, T., & Fitzroy, S. (2010). *Safe travels: evaluating mobility management traffic safety impacts*. Victoria: Victoria Transport Policy Institute.
- Mann, N. C., Mullins, R. J., MacKenzie, E. J., Jurkovich, G. J., & Mock, C. N. (1999). Systematic review of published evidence regarding trauma system effectiveness. *Journal of Trauma – Injury, Infection and Critical Care, 47*, S25–S33.
- Morrison, D. S., Thomson, H., & Petticrew, M. (2004). Evaluation of the health effects of a neighbourhood traffic calming scheme. *Journal of Epidemiology and Community Health, 58*, 837–840.
- Newman, P., & Kenworthy, J. (1999). *Sustainability and cities: overcoming automobile dependence*. Washington, DC: Island.
- Nurse, J., & Edmondson-Jones, P. (2007). A framework for the delivery of public health: an ecological approach. *Journal of Epidemiology and Community Health, 61*, 555–558.
- Olds, D. L., Eckenrode, J., Henderson, C. R., Jr., Kitzman, H., Powers, J., Cole, R., et al. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect. Fifteen-year follow-up of a randomized trial. *JAMA, 278*, 637–643.
- Olds, D., Henderson, C. R., Jr., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., et al. (1998). Long-term effects of nurse home visitation on children's criminal and antisocial behavior: 15-year follow-up of a randomized controlled trial. *JAMA, 280*, 1238–1244.
- Patz, J. A., Gibbs, H. K., Foley, J. A., Rogers, J. V., & Smith, K. R. (2007). Climate change and global health: quantifying a growing ethical crisis. *EcoHealth, 4*, 397–405.
- Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., & Mathers, C. (2004). *World report on road traffic injury prevention*. Geneva: World Health Organization.
- Peden, M., Oyegbite, K., Ozanne-Smith, J., Hyder, A. A., Branche, C., Fazlur Rahman, A. K. M., et al. (Eds.). (2008). *World report on child injury prevention*. Geneva: World Health Organization.
- Peek-Asa, C., & Zwerling, C. (2003). Role of environmental interventions in injury control and prevention. *Epidemiologic Reviews, 25*, 77–89.
- Prüss-Üstün, A., & Corvalán, C. (2006). *Preventing disease through healthy environments: towards an estimate of the environmental burden of disease*. Geneva: World Health Organization.
- Rautiainen, R. H., Lehtola, M. M., Day, L. M., Schonstein, E., Suutarinen, J., Salminen, S., & Verbeek, J. (2008). Interventions for preventing injuries in the agricultural industry. *Cochrane Database of Systematic Reviews*, (1), CD006398.
- Reason, J. (2000). Human error: models and management. *BMJ, 320*, 768–770.
- Reynolds, C. C., Harris, M. A., Teschke, K., Cripton, P. A., & Winters, M. (2009). The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature. *Environmental Health, 8*, 47.
- Roberts, I. (1995). Who's prepared for advocacy? Another inverse law. *Injury Prevention, 1*, 152–154.
- Roberts, I. (2010). *The energy glut*. London: Zed Books.
- Roberts, I., & Arnold, E. (2007). Policy at the crossroads: climate change and injury control. *Injury Prevention, 13*, 222–223.
- Roberts, H., & Meddings, D. (2010). Violence and unintentional injury: equity and social determinants. In E. Blas & A. S. Kurup (Eds.), *Equity, social determinants and public health programmes*. Geneva: World Health Organization.
- Robertson, L. S. (2007). Research designs and data analysis. In L. S. Robertson (Ed.), *Injury epidemiology: research and control strategies*. New York: Oxford University Press.
- Rockstrom, J., Steffen, W., Noone, K., Persson, A., Chapin, F. S., 3rd, Lambin, E. F., et al. (2009). A safe operating space for humanity. *Nature, 461*, 472–475.
- Roman, C. G., & Chalfin, A. (2008). Fear of walking outdoors: a multilevel ecologic analysis of crime and disorder. *American Journal of Preventive Medicine, 34*, 306–312.
- Shafi, S., de la Plata, C. M., Diaz-Arrastia, R., Bransky, A., Frankel, H., Elliott, A. C., et al. (2007). Ethnic disparities exist in trauma care. *The Journal of Trauma, 63*, 1138–1142.
- Shakespeare, T., & Watson, N. (2001). The social model of disability: an outdated ideology? *Research in Social Science and Disability, 2*, 9–28.

- Shaw, M., Smith, G. D., & Thomas, B. (2008). *The grim reaper's road map: an atlas of mortality in Britain*. Bristol: Policy Press.
- Shepherd, M., Austin, P., & Chambers, J. (2010). Driveway runaway, the influence of the built environment: a case control study. *Journal of Paediatrics and Child Health*, *46*, 760–767.
- Smedley, B. D., Stith, A. Y., Nelson, A. R., & Institute of Medicine (U.S.). Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. (2002). *Unequal treatment: confronting racial and ethnic disparities in health care*. Washington, DC: National Academy Press.
- Smith, G. S. (2003). Injury prevention: blurring the distinctions between home and work. *Injury Prevention*, *9*, 3–5.
- Social and Spatial Inequalities Group. (2011). Social and Spatial Inequalities (SASI). University of Sheffield. <http://www.sasi.group.shef.ac.uk/index.html>.
- Warren, L., Wrigley, J. M., Yoels, W. C., & Fine, P. R. (1996). Factors associated with life satisfaction among a sample of persons with neurotrauma. *Journal of Rehabilitation Research and Development*, *33*, 404–408.
- West, B. A., & Naumann, R. B. (2011). Motor vehicle-related deaths – United States, 2003–2007. *Morbidity and Mortality Weekly Report*, *60*, 52–55.
- Wong, M. D., Shapiro, M. F., Boscardin, W. J., & Ettner, S. L. (2002). Contribution of major diseases to disparities in mortality. *The New England Journal of Medicine*, *347*, 1585–1592.
- Woodcock, J., Edwards, P., Tonne, C., Armstrong, B. G., Ashiru, O., Banister, D., et al. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *Lancet*, *374*, 1930–1943.
- World Health Organization. (2007). *Preventing injuries and violence: a guide for ministries of health*. Geneva: World Health Organization.

Chapter 12

Behavioral Determinants

Deborah C. Girasek*

Introduction

Does human behavior cause most injuries? Casual observers are often quick to point out a mistake or ill-conceived action that the victim took just before being injured. Once an outcome is known, it may be quite easy to identify a risky or “careless” behavior that preceded an injury. Hindsight bias comes into play in these circumstances. Such armchair analyses can be misleading because they do not take into account the number of times that similar behaviors did not lead to injuries. As Haddon (1980) pointed out, humans are but one factor in a complex matrix of injury causation that includes the physical and social environment, and often a commercial product. So, while a victim’s behavior may have been the most proximal event to his/her injury, it is not the only thing we can manipulate to avoid similar occurrences in the future. Modern motor vehicle crash analyses focus less on what caused the crash and more on what caused the crash’s injury outcome (Stigson et al. 2008). When the Swedish Road Administration’s systems model was applied to real-life traffic crashes, they concluded that most crashes occurred when two or three components interacted. Noncompliance with safety criteria that involved the road was associated with a larger proportion of fatal outcomes than violations of their driver or vehicle safety criteria.

This chapter considers behavior from an ecological perspective that includes those at risk, but also other players whose actions can protect the public. The public health approach has proved highly effective, as will be discussed in subsequent chapters that highlight successful preventive interventions. This chapter illustrates how behaviors have been linked to injury, and how to measure those behaviors in a rigorous fashion. Such scientific precision is critical to our understanding of injury causation. Accurate measurement is also important for evaluating interventions that are designed to modify behaviors known to increase or decrease injury risk. We have tried to use diverse examples in the sections that follow. As other authors have noted, however, the scientific literature is heavily weighted toward work that was carried out in high-income settings. Best practices for changing behavior are addressed in Chap. 39.

*Disclaimer

The views expressed by the author are her own. She was not speaking as a representative of any Federal agency.

D.C. Girasek, PhD, MPH (✉)

Department of Preventive Medicine & Biometrics, Uniformed Services University
of the Health Sciences, Bethesda, MD, USA
e-mail: dgirasek@usuhs.edu

Protecting Self

The link between injury and behavior is probably most clear when people engage in activities that are widely recognized as hazardous. Examples are legion and could include smoking in bed, running red lights, or riding a motorcycle without wearing a helmet. While it is possible that individuals who are hurt under such circumstances lack knowledge about the risks they are exposing themselves to, it is more likely that they perceive their likelihood of injury to be so low that the benefits they associate with the behavior take precedence in their decision making. At other times, individuals may not even be conscious of engaging in a “dangerous behavior.” Virginia Tech’s pioneering 100 Car Study, for example, revealed that nearly 80% of the crashes they observed involved driver inattention in the 3 seconds prior to impact (Virginia Tech Transportation Institute 2011). People may also be injured because their cognitive and psychomotor skills are impaired by alcohol or other drugs they have consumed.

When people are not sober, they are more likely to hurt others and to be hurt themselves. Alcohol use has been associated with increased risk of injuries to motor vehicle occupants, bicyclists, pedestrians, and those engaged in recreational activities. Falls, poisonings, aspirations, cold- and fire-related injuries have also been linked to alcohol involvement, as have self-inflicted injuries and those resulting from interpersonal violence (WHO 2004). That is why alcohol use, and its rigorous measurement, is one of the most important behaviors of interest to professionals who carry out injury control research. Drugs other than alcohol can impair human performance, although they have not been studied as extensively. In this category, benzodiazepine use has probably been associated most convincingly with motor vehicle crash risk (Rapoport et al. 2009).

In high income countries, many safety devices are available for protecting individuals engaged in activities that could be injurious. Motorcycle helmets, for example, have been shown to reduce the risk of rider death by 42% and risk of head injury by 69% (Liu et al. 2008). It has been estimated that personal flotation devices (i.e., life jackets) could cut the number of drowning deaths among recreational boaters by 50% (Cummings et al. 2010). The death rate per 100 house fires reported from 2000 to 2004 was twice as high in homes without a working smoke alarm (Aherns 2007).

Passive protection devices, such as vehicle airbags, which operate without the need for human intervention, are generally favored by injury control professionals. It is not always accurate, however, to consider such technologies as being totally divorced from behavior. Engineers had to invent and evaluate them at the start. If regulators have not acted to mandate their exclusive availability, consumers must be persuaded to purchase them over less safe alternatives. And, as our experience with the examples listed in the previous paragraph demonstrates, acquired products are not always used or used effectively.

For these reasons, product marketing and safety education are behaviors that are worthy of consideration by scientists hoping to advance injury prevention. Many studies have focused on health care providers’ ability and likelihood of counseling patients about safety. A review of interventions carried out in clinical settings suggests that provider guidance can increase motor vehicle restraint use, smoke alarm ownership, and maintenance of a safe hot tap water temperature (DiGiuseppi and Roberts 2000). Far fewer investigations have involved retailers and manufacturers. One that was carried out in New Zealand found that adequate advice was provided for confederates purchasing bicycle helmets in less than half of the stores where helmets were sold (Plumridge et al. 1996). A US study demonstrated that the instructions provided by manufacturers of child safety seats are written at too high a reading level for a large proportion of the parents who need to understand them (Wegner and Girasek 2003).

The actions of providing effective education and public willingness to access such information also come into play for risks that do not involve the use or acquisition of safety equipment. The need to avoid hazardous products, for example, must be communicated to consumers (if regulators are

remiss in not removing them from the marketplace). Vitamin D supplementation, which has been demonstrated to reduce the risk of hip fracture in long-term care settings (Sawka et al. 2010), is another example of a risk reduction strategy that is currently dependent upon outreach activities. Training is a related behavior that can enable people to reduce their risk of injury while engaging in potentially hazardous activities. Formal swimming lessons, for example, have been associated with a reduced risk of childhood drowning (Brenner et al. 2009).

Protecting Others

In some respects, it is artificial to separate behaviors that protect self from those that protect others. Substance abuse, for example, puts both the drug user and those around him or her at increased risk of injury. Most risky driving behaviors, including non-use of passenger restraints, (Mayrose et al. 2005) pose a hazard to drivers as well as others in the risk-taker's environment. In this section, though, we will turn our attention to more prosocial behaviors; those that primarily increase other people's safety.

Parents probably come to mind first when an adult's duty to protect is considered. Children lack the developmental capability of protecting themselves from hazards in their environments. Until relatively recently, society's primary response to this challenge was to tell mothers that they should "watch their children." While such recommendations are face valid, they lacked a solid evidence-base. In part that is because they lacked the precision that both scientists and parents required. Tremendous progress has been made in the last decade, however, in operationalizing (i.e., defining for the purposes of measurement) the supervision construct. Supervision is now recognized as multidimensional: encompassing behaviors that indicate attention (watching, listening), proximity (touching, within reach, beyond reach), and continuity (constant, intermittent, not at all) (Morrongiello and Lasenby 2006). This conceptualization has enabled much more sophisticated investigation of supervision, and revealed that while its relationship to injury is clear, it is also complex. For example, parents supervise differently in different rooms of their homes, and different supervisory patterns have been associated with injury for sons versus daughters (Morrongiello and Lasenby 2006). Parenting styles have also been linked with adolescents' self-reports of driving while intoxicated, seatbelt use, speeding, and having been involved in a motor vehicle crash (Ginsburg et al. 2009). For scientists seeking validated measures, reports collected with the Parent Supervision Attributes Profile Questionnaire have been shown to correlate with observed supervision practices and child injury history (Morrongiello and House 2004).

While supervision may be a caretaker's first line of defense when it comes to protecting young children, environmental safety enhancements represent critical backup for busy, fallible adults. The installation of isolation pool fencing versus perimeter fencing, for example, has been shown to reduce toddler drowning by 83% (Thompson and Rivara 1998). Smoke detectors and sprinklers can protect all household residents, when properly installed and maintained.

Caregivers sometimes fail to carry out their protective responsibilities because they are under the influence of drugs, including alcohol. We know that most children who die in motor vehicle crashes that involve a drinking driver, for example, were passengers of the drinking driver (Quinlan et al. 2000). Alcohol consumption, a behavior that comes up throughout this chapter, is also relevant to protecting the larger community. Population-level indicators of drinking, for example, have been associated with dramatic differences in homicide rates (Room et al. 2005).

Members of the public and journalists are often quick to identify individual "villains," after serious injuries occur (Connor and Wesolowski 2004). Far less frequently, however, do they ask policy makers why they failed to enact legislation that could have reduced the likelihood that

dangerous behaviors would occur. This pattern belies the fact that public policy has proven to be one of the safety advocates' most powerful tools. This entire chapter could be devoted to legislative and regulatory interventions that have resulted in dramatic injury reductions. Here are just a few examples. Ten years prior to enactment of the Poisoning Prevention Packaging Act of 1970, which required that aspirin be packaged in child-resistant containers, 144 children died from accidental ingestion of aspirin. By 1988, that number had fallen to 3 (Baker et al. 1992). Motor vehicle passenger restraints – which manufacturers are required to install because of Federal legislation and drivers are required to wear because of state legislation – are estimated to have prevented 12,713 US deaths in 2009 alone (National Highway Traffic Safety Administration 2010). In 1990, the Australian state of Victoria passed a law that required all bicyclists to wear an approved safety helmet. Head injuries among cyclists fell by 70% in Victoria between 1991 and 1992 (Hemenway 2009). After legislation was passed in the UK to restrict the number of paracetamol and salicylate tablets that retail outlets could sell over the counter, suicidal deaths from those agents fell by 22% (Hawton et al. 2004). It has been estimated that a doubling of alcohol taxes would reduce alcohol-related mortality by 35% on average (Wagenaar et al. 2010). This list is not intended to convince readers that safety laws are always effective or desirable. Rather, it is to demonstrate that the actions of legislators and regulators merit the attention of injury control professionals.

An extension of this recommendation is that our work must not ignore the behavior of communicators who make the public aware of what laws have been enacted, or the efforts of civil servants who are charged with enforcing existing laws and regulations. There is considerable evidence that publicity surrounding safety laws increases their effectiveness, likely due to increased perceptions of enforcement (Williams et al. 1987; Williams and Wells 2004; Rogers and Schoenig 1994). Actual levels of enforcement are also related to improved levels of regulatory compliance. In New Hampshire, quarterly checks on retail outlets resulted in a 64% reduction in alcohol sales to underage youth (CDC 2004). The documented effectiveness of photo enforcement devices, which in effect extend the ability of law enforcement agencies to hold violators accountable, also demonstrates this relationship. A rigorous review of 35 studies designed to evaluate speed cameras, for example, found that their installation results in reduced speed and fewer crashes that involve serious injury and death (Wilson et al. 2010). Similarly, the absence of a law or reduction in enforcement may be associated with greater injury. An analysis that explored the impact of a 61% reduction in the number of traffic citations issued in Quebec, for example, concluded that it was associated with 184 additional collisions with injuries (Blais and Gagné 2010).

Research into the enforcement of safety laws and regulations has largely been limited to the traffic safety arena, with some notable exceptions. Ten years after New Zealand passed its Fencing of Swimming Pools (FOSP) Act, Morrison et al. (1999), surveyed authorities and found that they did not know the compliance status of 33% of the pools in their jurisdictions. This is perhaps not surprising since only 9% of respondents had procedures in place for locating and inspecting pools. A more encouraging report published 20 years later found a 65% increase in the proportion of pools that were reported as complying with the Fencing Act, and “a considerable improvement in the enforcement and monitoring activities” of New Zealand’s territorial authorities (Gulliver et al. 2009). It is worth noting that three studies have implicated legislative quality (i.e., language ambiguity) as an impediment to fencing law enforcement (Gulliver et al. 2009; Morrison et al. 1999; van Weerdenburg et al. 2006).

Safety-related laws and regulations are the result of policy makers’ actions, but they also result in changes in the behaviors of at-risk populations. Laws that discourage unsafe behaviors, or encourage protective practices, work by attaching the risk of an undesirable outcome to the behavior being discouraged. Visible enforcement is important because it increases the perceived likelihood of detection and punishment. Laws also codify our common social expectations (Shaw and Ogolla 2006).

Regulations often target the physical environment, by removing or redesigning hazardous products. In such instances, the actions of public administrators are intended to modify the behavior of manufacturers.

Even after safety laws are passed, their implementation should be monitored. A US Government Accountability Office investigation (2008) revealed that high visibility enforcement campaigns designed to reduce the number of people who drive under the influence (DUI) of alcohol are hindered in their effectiveness by shortcomings of the judicial system. Forty-eight percent of prosecutors, for example, reported that they were inadequately trained before beginning to prosecute DUI cases. Similarly, case dismissals have been attributed to inadequate training of arresting officers.

Mothers Against Drunk Driving includes court monitoring among their prevention strategies. They are one of many nonprofit organizations with significant potential to advance the injury prevention agenda. The American Academy of Pediatrics (2001), for example, has taken public stands on important issues like banning the manufacture and sale of mobile infant walkers. Similarly, the National Fire Protection Association (2011) is currently spearheading an initiative that provides advocates for local sprinkler ordinances with important tools and guidance.

Measurement Overview

Self-Reports

The most common method that injury researchers use to measure behaviors is to ask individuals to respond to survey questions. This strategy is perceived as being relatively easy, ethical and inexpensive. Self-reports pose problems, however, if they are assumed to reflect respondent behavior with perfect fidelity. There are many threats to the validity of self-reported behavior measures. The information being sought may never have been known by the survey respondent or the respondent may no longer recall what they did in the past. Self-reports are also subject to social desirability bias. This occurs when respondents report behaviors that they perceive to be more socially acceptable than their actual actions. It is usually obvious to volunteers participating in a safety study which behaviors are likely to be favored by the investigator. Sometimes, the behaviors we study are even illegal, so respondents may fear serious consequences if they report a violation. For certain subgroups or settings, risk behaviors may even be over-reported because they are perceived as enhancing social status (Brener et al. 2003).

Several validation studies suggest that self-reports associated with benign behaviors can be relatively accurate. The sensitivity and positive predictive value of parental reports on home safety practices, for example, have been shown to be quite high (Hatfield et al. 2006; Watson et al. 2003). Similarly, when the prevalence of self-reported safety belt use in 50 states was regressed against the prevalence of observed belt use in 1993, the fit of the resultant model was encouraging (Nelson 1996). In that study, telephone interviewers asked subjects, "How often do you use seat belts when you drive or ride in a car?" Only those respondents who selected the "always" response option were classified as belt users. Under those conditions, aggregated state prevalences calculated from self-reports were only 2% higher than observed use rates. The author referenced both historical and geographic data to support his general conclusion that self-reports of safety behaviors are more likely to be valid when the actual prevalence of the behavior in the population is high.

Validated measures of booster and child seat use have also been published (Uherick et al. 2010). Investigators interested in surveying children about their risk behaviors in relation to falls, burns,

poisoning, motor vehicle crashes, suffocation/choking, drowning, or bike and pedestrian safety may want to consider administering the BACKIE questionnaire, because of its established psychometric properties (Morrongiello et al. 2010).

If survey researchers need to develop their own instrument, there are several steps they can take to reduce sources of measurement error. Cognitive interviewing (CI) is a pretesting method developed by the National Center for Health Statistics' Questionnaire Design Research Laboratory. It is a qualitative technique that calls for subjects to explain how they would respond to draft survey items, based upon their interpretation of each question and its response options. It was cognitive interviewing, for example, that helped the National Highway Traffic Administration determine that 70+ % of survey respondents who report using their seat belts "most of the time" also report that the last time they did not wear their belt was within the previous 7 days (Block 2000). Since the CI process can be relatively simple to carry out (Willis 1994), it is highly recommended to investigators who rely on survey instruments for behavioral measurement.

If your questionnaire asks about behaviors that occur frequently, items should refer respondents to a short, clearly defined timeframe in the immediate past (Kimberlin and Winterstein 2008). It is also best to ask subjects to provide a number, rather than providing them with response options that require subjective interpretation (e.g., often, rarely). Respondents should be required to enter a value, rather than select from a range of values listed on the questionnaire. This is because survey subjects appear to interpret mid-range scales as "average" in the eyes of the investigator and have a tendency to assimilate their answers to that norm (Morsbach and Prinz 2006).

The inclusion of permissive preambles, or supportive/face-saving language should be considered when developing questions that deal with stigmatized behaviors. Such formats have been shown to increase the proportion of respondents who admit to actions that could be perceived as negative (Morsbach and Prinz 2006). An even more powerful method of reducing subjects' tendency to self-censor is to allow them to complete their own questionnaires (i.e., rather than using interviewer-administered survey instruments). Computerized administration of survey instruments has also been shown to enhance respondent candor, in both high- and low-income countries (Brenner et al. 2003; Morsbach and Prinz 2006; Langhuag et al. 2010). Adolescents may be even more sensitive to mode of administration effects (Brenner et al. 2003). Young drivers who used a secure website to access survey questions were shown to report their traffic offenses and crash history with good levels of accuracy (Boufous et al. 2010). Young people are also believed to report more candidly when surveyed at school than at home.

Investigators are encouraged to consider whether the behavior of interest should be conceptualized along a continuum of compliance. For example, for optimal protection bicyclists must purchase approved helmets that fit, and wear them correctly and consistently, on each trip that they take. Homeowners must purchase the correct type and number of smoke detectors, install them in the recommended locations, and maintain them in working order. Under these circumstances, a series of items should be asked about helmet or smoke detector use. This approach will yield ordinal rather than dichotomous data and provide researchers with a much more refined understanding of existing practice gaps.

Alcohol Use

Readers may be pleasantly surprised to learn that alcohol use is not generally perceived as a particularly sensitive topic by survey respondents in the US (Greenfield and Kerr 2008). Decades of research demonstrate that self-reports of alcohol use are generally valid and reliable, assuming that the respondents are alcohol-free at the time of survey administration and they are assured that their responses will be treated confidentially (Connors and Maisto 2003). That does not mean, however,

that assessment of alcohol intake is simple or straightforward. For example, research subjects are often asked to reply in terms of “standard drink” units that may bear little resemblance to the way that alcohol is actually served in social settings (Del Boca and Darkes 2003). In the real world, large variations in drink size and alcohol content are the norm, depending upon the beverage and context in question (Greenfield and Kerr 2008).

The two general approaches that are usually taken to measure alcohol use include:

- Asking people to summarize how much and how often they drink [i.e., quantity/frequency (QF) measures] or
- Asking them to report (retrospectively or prospectively) on the amount of alcohol they consumed each day (i.e., daily estimation procedures) (Del Boca and Darkes 2003).

Both methods have advantages and limitations. Prospective daily estimations may yield the most valid data and provide for more sophisticated analyses, but they are expensive and impose a much greater burden on respondents (Del Boca and Darkes 2003).

When asking subjects to summarize past alcohol consumption, it is critical to provide them with a reference period (Greenfield and Kerr 2008). Most national surveys use 12 months or 30 days. Recall ability is enhanced by using relatively short reference periods (Brener et al. 2003). This is true because of cognitive limitations characteristic of all humans, and the fact that acute and chronic drug use can impair memory. One limitation of measures keyed to the last month, however, is that they may yield data that underrepresent light, infrequent drinkers or those who drink heavy amounts intermittently. The National Institute on Alcohol Abuse and Alcoholism stresses that it is important to understand a respondent’s *pattern* of alcohol use, not just the typical number of drinks they consume each day. To address this, it has been recommended that investigators use graduated frequency (GF) measures. GF measures present response options that pertain to the maximum number of drinks they consumed on any 1 day in the previous year. Using that response as an entry point, subjects are asked how frequently they drank that amount. The same query follows for a series of fixed levels that represent descending quantities (i.e., three to four drinks and one to two drinks) (Greenfield and Kerr 2008). Generally, questions that pertain to current drinking patterns are considered more reliable than those which ask subjects to describe their past drinking habits.

Daily data collection (e.g., diary-keeping) has been used to study alcohol use and parental supervision. There is debate about the degree to which such methods trigger reactivity. Reactivity occurs when the behavior under study changes in reaction to being measured. Some investigators prefer experience sampling methods in which a respondent is paged at random times and asked to report on his or her current behavior (Zimmerman et al. 2006).

Two techniques that have been utilized in an attempt to increase or verify the validity of alcohol-related self-reports are the bogus pipeline and collecting data from “collateral individuals.” Bogus pipeline refers to informing research subjects that they will participate in an objective procedure that can reveal their true behavior, when in fact the technique will do no such thing. The second approach, which is carried out with the respondent’s knowledge and permission, is to see whether informants familiar with the subject corroborate his or her drinking reports. The greatest agreement between subjects and collateral respondents has been observed when the collaterals are spouse/partners and report confidence about their knowledge of the subject’s alcohol consumption (Connors and Maisto 2003). It worth noting that when such reports are discrepant, the subject usually reports more drinking than his or her collateral. The value-added of both these techniques are subject to debate (Brener et al. 2003; Del Boca and Darkes 2003). Like all of the research techniques we are discussing, of course, they should not be employed without the review and approval of a Human Subjects Protection Committee.

The previously noted observation that computer-assisted self-interviewing (CASI) may yield more candid reports, appears to generalize to the study of alcohol consumption. A-CASI, which involves the use of audio-taped questions that are presented over headphones and on a computer

screen, may prove superior to CASI (Del Boca and Darkes 2003). It makes fewer literacy demands upon subjects and allows for graphic displays of beverage containers. For references to multiple, important consensus projects that focused on alcohol measurement, see Greenfield and Kerr (2008). For the exact wording of a minimum set of alcohol consumption questions that are recommended by the NIAA, see <http://www.niaaa.nih.gov/Resources/ResearchResources/TaskForce.htm>

Other Behavioral Measurement Techniques

Direct observation is the gold standard for behavior measurement. It overcomes many of the limitations that are inherent in behavioral self-reports (e.g., recall error and social desirability). Such studies can be challenging to carry out in a rigorous manner, however, and are expensive. They also provide very limited information about the subjects being observed (i.e., unless they can be approached afterwards and persuaded to answer survey questions). We recommend that observational studies employ two coders, so that kappa statistics can be calculated to assess their agreement. Several model protocols for observing road safety-related behaviors (e.g., seat belt use and electronic device use) have been published by the National Traffic Safety Administration. Other field observation studies have involved parents crossing streets with young children and parents supervising their children on playgrounds.

“Naturalistic driving studies” represent an exciting development in the road safety arena (SWOV 2010). These investigations take advantage of technologic advances by equipping vehicles with instrumentation that monitors driver behavior, vehicle maneuvers, and external conditions simultaneously. Large amounts of data are subsequently generated while participants engage in their normal driving activities over extended periods of time.

Many topics of interest to injury prevention researchers do not lend themselves to field observation. For example, it would be unethical to watch small children navigate busy thoroughfares alone or to dose swimmers with ethanol to see how their performance degrades. Other behaviors occur in private, or are unlikely to be performed under conditions of observation (e.g., abusive parenting). Creative investigators have developed promising surrogates for direct observation, often using advanced technology.

A recent special issue of *Accident Analysis & Prevention* featured 25 papers that utilized driving simulators to answer questions related to traffic safety (Boyle and Lee 2010). Such developments allow us to examine how drivers react under controlled circumstances that can be varied to incorporate different weather conditions, road designs, vehicle models, and levels of impairment. Assessments of the validity of results obtained under simulated driving conditions are complex and ongoing. They focus on whether the virtual experience reproduces the physical driving environment accurately, and whether it elicits operator reactions that would occur in the real world (Yan et al. 2008).

Morrongiello and Lasenby (2006) have observed parents interacting with children in a laboratory that has been set up to appear natural (e.g., to resemble a waiting room) and to contain real hazards. Under her “contrived hazard” condition, however, the hazards have been modified to eliminate injury risk. Barton and Schwebel (2007) videotaped children crossing a pretend road that had been constructed perpendicular to a real road. Under varied conditions of supervision, the children were instructed to use the traffic on the real road in judging when to cross the pretend road. Videotape analysis has also been used to study sports injuries, occupational risk exposures, and physician counseling practices.

Biochemical measures have traditionally been favored over self-reports of drug use because of their presumed objectivity and accuracy. Breath samples have been used most often by injury researchers to estimate subject’s recent alcohol consumption. Saliva testing has also proven to yield valid blood alcohol concentration (BAC) estimates (Degutis et al. 2004). Such techniques still pose

limitations, however. The approximated BACs produced by these methods are time sensitive and subject to individual variation (e.g., based upon gender, weight, and food consumption). Exploration into the feasibility of transdermal alcohol sensors is underway.

The National Highway Traffic Safety Association (NHTSA) has developed standardized field sobriety tests for use by law enforcement officials who are trying to assess whether drivers are impaired by alcohol. NHTSA's field tests are under study for the detection of other drug use, as are oral fluids and sweat sampling. For now, however, urinalysis is considered the most advanced means of detecting illegal drug ingestion (Brener et al. 2003). For a technical discussion of alcohol biomarker options, see Peterson (2004/2005).

Archival and Secondary Data

Researchers sometimes substitute proxy measures in lieu of assessing subject behavior directly. Data on helmet sales, for example, might be reviewed before and after a bike safety campaign. Such leaps must be considered carefully, however. Helmet sales might make sense if helmet acquisition was a campaign goal. It would be harder to justify as an indicator of helmet use. State and national data on alcohol sales have been shown to produce much higher per individual estimates than those generated from survey results (Greenfield and Kerr 2008). This may be due to the discrepancies in serving sizes discussed earlier. On a population level, alcohol use is often measured by indicators of average amounts consumed, such as per capita consumption. For acute public health effects such as injuries, however, it may be more important to measure drinking *patterns* (WHO 2004). This is because the former will be heavily influenced by the percentage of persons who drink in the population overall. Also, consuming very high amounts of alcohol episodically is much more dangerous, from the perspective of injury causation, than consuming moderate amounts of alcohol with meals on a daily basis.

Medical chart notes are often regarded as valid indicators of health provider behaviors. It is important to remember, however, that the content of medical records is influenced by institutional policies, provider training and provider preferences, so they may not yield accurate behavioral data (Kimberlin and Winterstein 2008). The use of police data for public health purposes have been shown to be highly problematic (Alsop and Langley 2001; Amoros et al. 2008). In the US, police accident report forms do not adequately mirror the traffic safety laws that are in effect in each state (Brock and Lapidus 2008). Police records have also proven to be unreliable sources of drivers' cell phone use. Phone company billing records have shown promise as a source for verifying drivers' cell phone use (McCartt et al. 2006).

The Centers for Disease Control and Prevention (CDC) collaborates with state health departments to collect telephone survey data on adult risk behaviors on a monthly basis. The Behavioral Risk Factor Surveillance System (BRFSS) is used to produce national, state, county, and city level estimates, as well as interactive maps. While it emphasizes behaviors related to chronic diseases more so than injuries, the BRFSS has included many injury-relevant topics in their questionnaires over the years (e.g., seat belt use, bicycle helmet use, alcohol consumption, firearm storage, sleep patterns, child abuse, disability, and carbon monoxide and smoke detectors). Spanish and English versions of their past surveys are available at the CDC's website (<http://www.cdc.gov/brfss/>). Many of the instrument's measures have been tested for reliability and validity (Nelson et al. 2001). BRFSS data can be accessed as far back as 1984. It has been used to evaluate injury prevention campaigns and to identify subgroups of the population at increased risk for injury.

In 1994 and 2002, the CDC's National Center for Injury Prevention and Control sponsored two national surveys to collect data of specific relevance to injuries. Several published reports have come out of the Injury Control and Risk Survey initiative, and the CDC will share ICARIS data

upon request (cdcinfo@cdc.gov). To view the ICARIS-2 survey instrument and dataset documentation, readers are referred to <http://www.cdc.gov/ncipc/osp/icaris2.htm>

Other organizations that collect data on risk behaviors that are relevant to injury include the AAA Foundation for Traffic Safety, the Federal Bureau of Investigation, the Insurance Institute for Highway Safety, the Inter-University Consortium for Political and Social Research, the National Highway Traffic Safety Administration, the National Institute on Alcohol Abuse and Alcoholism, the Substance Abuse and Mental Health Services Administration, and the World Health Organization.

Conclusion

Behaviors and injuries are linked in myriad ways. Researchers will have limited impact, however, if they confine their study to the behavior of those who are at risk for injury.

Rigorous behavioral measurement is often more challenging than it first appears. If practical or ethical concerns dictate that self-reports be used, a study should incorporate design elements that compensate for their limitations. Investigators should also explore whether measurement tools with known psychometric properties have been published. It is recommended that multiple methods be used to measure important variables. That practice will either increase the study team's confidence in their findings or raise red flags that merit cautious interpretation of results. Finally, to inform future interventions, data should be collected on psychosocial factors that have been shown to influence behavior (e.g., self-efficacy and perceived barriers). To that end, it is recommended that a multidisciplinary study team be convened early in the planning process.

References

- Aherns, M. (2007). *U.S. Experience with smoke alarms and other fire detector/alarm equipment*. Quincy, MA: Fire Analysis & Research Division, National Fire Protection Association.
- Alsop, J., & Langley, J. (2001). Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accident Analysis & Prevention, 33*, 353–359.
- American Academy of Pediatrics Commission on Injury and Poison Prevention. (2001). Injuries associated with infant walkers. *Pediatrics, 108*, 790–792.
- Amoros, E., Martin, J. L., Lafont, S., & Laumon, B. (2008). Actual incidences of road casualties, and their injury severity, modeled from police and hospital data, France. *European Journal of Public Health, 18*, 360–365.
- Baker, S. P., O'Neill, B., Ginsburg, M. J., & Li, G. (1992). *The injury fact book* (2nd ed.). New York: Oxford University Press.
- Barton, B. K., & Schwebel, D. C. (2007). The roles of age, gender, inhibitory control, and parental supervision in children's pedestrian safety. *Journal of Pediatric Psychology, 32*, 517–526.
- Blais, E., & Gagné, M. (2010). The effect on collisions with injuries of a reduction in traffic citations issued by police officers. *Injury Prevention, 16*, 393–397.
- Block, A. W. (2000). *1998 Motor vehicle occupant safety survey: volume 2, seat belt report*. Washington, DC: National Highway Safety Administration.
- Boufous, S., Ivers, R., Senserrick, T., Stevenson, M., Norton, R., & Williamson, A. (2010). Accuracy of self-report of on-road crashes and traffic offences in a cohort of young drivers: the DRIVE study. *Injury Prevention, 16*, 275–277.
- Boyle, L. N., & Lee, J. D. (2010). Using driving simulators to assess driving safety. *Accident Analysis & Prevention, 42*, 785–787.
- Brener, N. D., Billy, J. O. G., & Grady, W. R. (2003). Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: evidence from the scientific literature. *Journal of Adolescent Health, 33*, 436–457.
- Brenner, R. A., Taneja, G. S., Haynie, D. L., Trumble, A. C., Qian, C., Klinger, R. M., et al. (2009). Association between swimming lessons and drowning in childhood: a case-control study. *Archives of Pediatrics & Adolescent Medicine, 163*, 203–210.

- Brock, K., & Lapidus, G. (2008). Police accident report forms: safety device coding and enacted laws. *Injury Prevention, 14*, 405–407.
- Centers for Disease Control and Prevention (CDC). (2004). Enhanced enforcement of laws to prevent alcohol sales to underage persons—New Hampshire, 1999–2004. *MMWR. Morbidity and Mortality Weekly Report, 53*, 452–454.
- Connor, S. M., & Wesolowski, K. (2004). Newspaper framing of fatal motor vehicle crashes in four Midwestern cities in the United States, 1999–2000. *Injury Prevention, 10*, 149–153.
- Connors, G. J., & Maisto, S. A. (2003). Drinking reports from collateral individuals. *Addiction, 98*(suppl 2), 21–29.
- Cummings, P., Mueller, B. A., & Quan, L. (2010). Association between wearing a personal flotation device and death by drowning among recreational boaters: a matched cohort analysis of United States Coast Guard data. *Injury Prevention*. doi:10.1136/ip.2010.028688.
- Degutis, L. C., Rabinovici, R., Sabbaj, A., Mascia, R., & D’Onofrio, G. (2004). The saliva strip is an accurate method to determine blood alcohol concentration in trauma patients. *Academic Emergency Medicine, 11*, 885–887.
- Del Boca, F. K., & Darkes, J. (2003). The validity of self-reports of alcohol consumption: state of the sciences and challenges for research. *Addiction, 98*(Suppl. 2), 1–12.
- DiGuseppi, C., & Roberts, I. G. (2000). Individual-level injury prevention strategies in the clinical setting. *The Future of Children, 10*, 53–82.
- Ginsburg, K. R., Durbin, D. R., Garcia-España, J. F., Kalicka, E. A., & Winston, F. K. (2009). Associations between parenting styles and teen driving, safety-related behaviors and attitudes. *Pediatrics, 124*, 1040–1051.
- Greenfield, T. K., & Kerr, W. C. (2008). Alcohol measurement methodology in epidemiology: Recent advances and opportunities. *Addiction, 103*, 1082–1099.
- Gulliver, P., Chalmers, D., & Cousins, K. (2009). Achieving compliance with pool fencing legislation in New Zealand: how much progress has been made in 10 years? *International Journal of Injury Control and Safety Promotion, 16*(3), 127–132.
- Haddon, W., Jr. (1980). Options for the prevention of motor vehicle crash injury. *Israel Journal of Medical Sciences, 16*, 45–68.
- Hatfield, P. M., Staresinic, A. G., Sorkness, C. A., Peterson, N. M., Schirmer, J., & Katcher, M. L. (2006). Validating self reported home safety practices in a culturally diverse non-inner city population. *Injury Prevention, 12*, 52–57.
- Hawton, K., Simkin, S., Deeks, J., Cooper, J., Johnston, A., Waters, K., et al. (2004). UK legislation on analgesic packs: before and after study of long term effect on poisonings. *BMJ*. doi:10.1136/bmj.38253.572581.7C.
- Hemenway, D. (2009). *While we were sleeping: success stories in injury and violence prevention*. Berkeley, CA: University of California Press.
- Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*, 2276–2284.
- Langhuag, L. F., Sherr, L., & Cowan, F. M. (2010). How to improve the validity of sexual behaviour reporting: systematic review of questionnaire delivery modes in developing countries. *Tropical Medicine & International Health, 15*, 362–381.
- Liu, B. C., Ivers, R., Norton, R., Boufous, S., Blows, S., & Lo, S. K. (2008). Helmets for preventing injury in motorcycle riders (review). *Cochrane Database of Systematic Reviews, (1)*, CD004333.
- Mayrose, J., Jehle, D., Hayes, M., Tinnesz, D., Piazza, G., & Wilding, G. E. (2005). Influence of unbelted rear-seat passengers on driver mortality: “the backseat bullet”. *Academic Emergency Medicine, 12*, 130–134.
- McCartt, A. T., Hellinga, L. A., & Braitman, K. A. (2006). Cell phones and driving: review of the research. *Traffic Injury Prevention, 7*, 89–106.
- Morrison, L., Chalmers, D. J., Langley, J. D., Alsop, J. C., & McBean, C. (1999). Achieving compliance with pool fencing legislation in New Zealand: a survey of regulatory authorities. *Injury Prevention, 5*(2), 114–118.
- Morrongiello, B. A., & House, K. (2004). Measuring parent attributes and supervision behaviors relevant to child injury risk: examining the usefulness of questionnaire measures. *Injury Prevention, 10*, 114–118.
- Morrongiello, B. A., & Lasenby, J. (2006). Supervision as a behavioral approach to reducing child-injury risk. In A. C. Gielen, D. A. Sleet, & R. DiClemente (Eds.), *Injury and violence prevention: behavioral science theories, methods, and applications*. San Francisco, CA: Jossey-Bass.
- Morrongiello, B. A., Cusimano, M., Barton, B. K., Orr, E., Chipman, M., Tyberg, J., et al. (2010). Development of the BACKIE questionnaire: a measure of children’s behaviors, attitudes, cognitions, knowledge, and injury experiences. *Accident Analysis & Prevention, 42*, 75–83.
- Morsbach, S. K., & Prinz, R. J. (2006). Understanding and improving the validity of self-report of parenting. *Clinical Child and Family Psychology Review, 9*, 1–20.
- National Fire Protection Association. (2011). Fire Sprinkler Initiative. <http://www.firesprinklerinitiative.org/legislation.aspx>. Accessed 2 Feb 2011.
- National Highway Traffic Safety Administration. (2010). Lives saved in 2009 by restraint use and minimum drinking age laws. Traffic Safety Facts: September 2010. Washington, DC: United States Department of Transportation. DOT HS 811 383.

- Nelson, D. E. (1996). Validity of self reported data on injury prevention behavior: lessons from observational and self reported surveys of safety belt use in the US. *Injury Prevention*, 2, 67–69.
- Nelson, D., Holtzman, D., Bolen, J., Stanwyck, C., & Mack, K. (2001). Reliability and validity of measurements from the Behavioral Risk Factor Surveillance System (BRFSS). *Sozial-und Praventivmedizin*, 46(Suppl 1), S3–S42.
- Peterson, K. (2004/2005). Biomarkers for alcohol use and abuse. *Alcohol Research & Health*, 28, 30–37.
- Plumridge, E., McCool, J., Chetwynd, J., & Langley, J. D. (1996). Purchasing a cycle helmet: are retailers providing adequate advice? *Injury Prevention*, 2, 41–43.
- Quinlan, K. P., Brewer, R. D., Sleet, D. A., & Dellinger, A. M. (2000). Characteristics of children passenger deaths and injuries involving drinking drivers. *JAMA*, 283, 2249–2252.
- Rapoport, M. J., Lanctôt, K. L., Streiner, D. L., Bédard, M., Vingilis, E., Murray, B., et al. (2009). Benzodiazepine use and driving: a meta-analysis. *Journal of Clinical Psychiatry*, 70, 663–673.
- Rogers, P. N., & Schoenig, S. E. (1994). A times series evaluation of California's 1982 driving-under-the-influence legislative reforms. *Accident Analysis & Prevention*, 26, 63–78.
- Room, R., Babor, T., & Rehm, J. (2005). Alcohol and public health. *Lancet*, 365, 519–530.
- Sawka, A. M., Ismaila, N., Cranney, A., Thabane, L., Kastner, M., Gafni, A., et al. (2010). A scoping review of strategies for the prevention of hip fracture in elderly nursing home residents. *PLoS One*, 5(3), e9515.
- Shaw, F. E., & Ogolla, C. P. (2006). Law, behavior, and injury prevention. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: behavioral science theories, methods, and applications*. San Francisco, CA: Jossey-Bass.
- Stigson, H., Krafft, M., & Tingvall, C. (2008). Use of fatal real-life crashes to analyze a safe road transport system model, including the road user, the vehicle and the road. *Traffic Injury Prevention*, 9, 463–471.
- SWOV. (2010). Naturalistic driving: observing everyday driving behaviour. SWOV Fact Sheet. Ledischednam, The Netherlands: SWOV.
- Thompson, D. C., & Rivara, F. (1998). Pool fencing for preventing drowning of children. *Cochrane Database of Systematic Reviews*, (1), CD001047. doi:10.1002/14651858.CD001047.
- Uherick, L., Gorelick, M. H., Biechler, R., & Brixey, S. N. (2010). Validation of two child passenger safety questionnaires. *Injury Prevention*, 16, 343–347.
- United States Government Accountability Office Report to the Chair, Committee on Transportation and Infrastructure, House of Representatives. (2008). *Traffic safety: improved reporting and performance measures would enhance evaluation of high-visibility campaigns* (Report No. GAO-08-477). Washington, DC: U.S. Government Accountability Office.
- van Weerdenburg, K., Mitchell, R., & Wallner, F. (2006). Backyard swimming pool safety inspections: a comparison of management approaches and compliance levels in three local government areas of NSW. *Health Promotion Journal of Australia*, 17, 37–42.
- Virginia Tech Transportation Institute (VTTI). (2011). 100-Car Naturalistic Study Fact Sheet. Blacksburg, VA: VTTI. http://www.vtti.vt.edu/PDF/100-Car_Fact-Sheet.pdf. Accessed 2 Feb 2011.
- Wagenaar, A. C., Tobler, A. L., & Komro, K. A. (2010). Effects of alcohol tax and price policies on morbidity and mortality: a systematic review. *American Journal of Public Health*, 100, 2270–2278.
- Watson, M., Kendrick, D., & Coupland, C. (2003). Validation of a home safety questionnaire used in a randomized controlled trial. *Injury Prevention*, 9, 180–183.
- Wegner, M. V., & Girasek, D. C. (2003). How readable are child safety seat installation instructions? *Pediatrics*, 111, 588–591.
- Williams, A. F., & Wells, J. K. (2004). The role of enforcement programs in increasing seat belt use. *Journal of Safety Research*, 35, 175–180.
- Williams, A. F., Lund, A. K., Preusser, D. F., & Blomberg, R. D. (1987). Results of a seat belt use law enforcement and publicity campaign in Elmira, New York. *Accident Analysis & Prevention*, 19, 243–249.
- Willis, G. B. (1994). *Cognitive interviewing and questionnaire design: a training manual*. Cognitive Methods Staff Working Paper Series, No. 7. Office and Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention.
- Wilson, C., Willis, C., Hendrikz, J. D., Le Brocque, R., & Bellamy, N. (2010). Speed cameras for the prevention of road traffic injuries and deaths (review). *Cochrane Database of Systematic Reviews*, (10), CD004607.
- World Health Organization (WHO). (2004). *Consequences of alcohol use: health effects and global burden of disease*. Global Status Report on Alcohol 2004. Geneva: Department of Mental Health and Substance Abuse, World Health Organization.
- Yan, X., Abdel-Aty, M., Radwan, E., Wang, X., & Chilakapati, P. (2008). Validating a driving simulator using surrogate safety measures. *Accident Analysis & Prevention*, 40, 274–288.
- Zimmerman, R. S., Atwood, K. A., & Cupp, P. K. (2006). Improving validity of self-reports for sensitive behaviors. In R. A. Crosby, R. J. DiClemente, & L. F. Salaza (Eds.), *Research methods in health promotion* (pp. 267–268). San Francisco, CA: Jossey-Bass.

Part III
Injury Outcome

Chapter 13

Injury Profiling

Limor Aharonson-Daniel

Background

The Development of Coding and Classification

The concept of profiling and categorization is well embedded into the medical profession. The classification of patients with similar clinical manifestations into meaningful categories termed diseases has characterized the medical profession since ancient times. Attempts to classify diseases systematically date back to Graunt's book "Natural and Political Observations Made upon the Bills of Mortality" (Graunt 1975) (1662), where he analyzed mortality trends in relation to demographic data, attempting to create a warning system against a plague in London. Later on in history, William Farr, founder of the General Register Office of England and Wales (1837), was a great contributor to the development of classifications and argued for international uniformity in their use (Stone 1997). At the first International Statistical Congress in 1853, William Farr and Marc d'Espine, of Geneva, were asked to prepare an internationally applicable, uniform classification of causes of death; by the second Congress, in 1855, they submitted two separate lists. Farr's classification was arranged under five groups including epidemic diseases, general diseases, local diseases arranged according to anatomical site, developmental diseases, and diseases that are the direct result of violence, while D'Espine classified diseases according to their nature. The Congress adopted a compromise list of 139 categories. While this classification underwent several revisions and was never universally accepted, the general arrangement proposed by Farr, including the principle of classifying diseases by anatomical site, survived as the basis of the International List of Causes of Death (Israel 1978).

The International Statistical Institute (the successor to the above-mentioned International Statistical Congress), at its 1891 meeting, charged a committee, chaired by Jacques Bertillon, Chief of Statistical Services of the City of Paris, with the preparation of a classification of causes of death. The report of this committee was presented by Bertillon at the next meeting in Chicago in 1893, where it was adopted. The classification represented a synthesis of English, German, and Swiss classifications and distinguished between general diseases and those localized to a particular organ or

L. Aharonson-Daniel, PhD (✉)

Department of Emergency Medicine, Faculty of Health Sciences, Ben-Gurion
University of the Negev, P.O. Box 653, Beer-Sheva 84105, Israel

PREPARED Center for Emergency Response Research, Ben-Gurion
University of the Negev, P.O. Box 653, Beer-Sheva 84105, Israel
e-mail: limorad@bgu.ac.il

anatomical site. The Bertillon Classification of Causes of Death received general approval and was adopted by several countries (Bertillon 1913). In 1898, the American Public Health Association recommended the adoption of the Classification by registrars of Canada, Mexico, and the USA and suggested that the classification be revised every 10 years. Ten years later, Bertillon presented a report on the progress of the classification, leading to the first International Conference for the Revision of the Bertillon or International List of Causes of Death attended by delegates from 26 countries. The classification of causes of death was adopted on 21 August 1900 and Bertillon continued to lead the International List of Causes of Death, through the revisions of 1910 and 1920 until his death in 1922. Following his death, the “Mixed Commission,” created with an equal number of representatives from the International Statistical Institute and the Health Organization of the League of Nations, drafted the proposals for the Fourth (1929) and the Fifth (1938) revisions of the International List of Causes of Death. In parallel with the progress of cause of death classification, William Farr had recognized that it was desirable to extend the same system of nomenclature to diseases which, though not fatal, cause disability in the population (Langmuir 1976; Lilienfeld 2007). A parallel classification of diseases for use in statistics of illness was, therefore, adopted at the second conference in 1909. The additional categories for nonfatal diseases were formed by subdividing a number of categories of the cause-of-death classification into two or three disease groups, each of these being designated by a letter.

In 1948, the First World Health Assembly endorsed the report of the Sixth Revision Conference and adopted World Health Organization Regulations No. 1, prepared on the basis of the recommendations of the Conference. The International Classification, including the Tabular List of Inclusions defining the content of the categories, was incorporated into the Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death (ICD) (World Health Organization 1949).

The Use of ICD Codes for Recording Injury

ICD is considered to be the global standard to report and categorize diseases, health-related conditions, and external causes of disease and injury. It is used for recording information related to mortality and morbidity. Reports generated using ICD are at the basis of planning health services and justifying public health programs and prevention strategies. ICD is used by many countries to record basic health statistics, thus enabling comparisons among countries.

From the sixth revision of the ICD named “Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death” and onwards, a designated chapter enhanced and improved the codes for recording injuries.

Often, the move from one version to the next in mortality data occurs prior to the move in morbidity, in a way preserving the original development of the classification. Currently, for injury morbidity data, information continues to be coded using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), long after the mortality data moved to the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) (World Health Organization 1992a). After several delays, ICD-10-CM, Diagnosis Codes is scheduled to replace the ICD-9-CM on October 1, 2013. Work on ICD-11 has already begun by Topic Advisory Groups (TAG) aiming for a brief version for use in primary care, a detailed version for use in specialty settings, and an advanced version for use in research (<http://www.who.int/classifications/icd/ICDRevision/en/>).

Chapter 17 of the ICD-9-CM includes classifications for recording injury and poisoning. The range of codes is 800–999. The chapter is divided by the type of injury such as fractures, internal injuries crush, burn, etc.; each injury type is then described by anatomical site. In ICD-10, the axes have been exchanged and injuries are grouped by anatomical site rather than by type of injury.

Several other features related to injury were updated in ICD-10-CM. These include laterality (left, right, and bilateral), blood alcohol level (known to increase vulnerability to injury), and the incorporation of the external causes of injury (E-codes) into the main classification rather than separated into supplementary classifications as they were in ICD-9-CM.

As the number of defined conditions and corresponding codes increased, a new instrument was devised to simplify the process of selecting ICD-9-CM diagnostic codes used to describe an injury and to standardize the definition and classification of injuries. This innovative concept and structure for monitoring, reporting, and analysis of records with multiple injuries coded appears below, under the “The Barell matrix” and “Multiple Injury Profiles” sections.

The Use of Abbreviated Injury Scale for Recording Injury

The Abbreviated Injury Scale (AIS) is an anatomically based classification. It is used to both categorize the type and nature of injury and classify the severity of single injuries. The AIS is updated approximately every 5 years since it was first published by a joint committee of the American Medical Association (AMA), Society of Automotive Engineers, and the AAAM in 1971 (Committee on Medical Aspects of Automotive Safety 1971). The most recent revision to-date (2008) introduced significant changes into the classification and added detail that is significant for injury profiling. This includes diagnoses for pelvic injuries, restructuring the diagnoses of upper and lower extremities and more.

Beyond the coding of the anatomic description of the injury, AIS injury codes are followed by a single digit that ranks severity on a scale of 1–6, with one being minor, five severe, and six currently unsurvivable. AIS serves as the basis for the Injury Severity Score (ISS) (Baker et al. 1974). ISS is formed by the summation of the squares of the severity digit in the AIS of the most severe injuries in any three of six predefined body regions.

Subsequent to the ISS, the New Injury Severity Score (NISS) was developed; this simplified coding of multiple injuries because it ignored body region, thus allowing two or three injuries to be included from the same body region. Many studies have shown that the NISS correlates better with mortality than the ISS, while maintaining its advantages such as being easy to use and understand (Osler et al. 1997). Studies have shown that despite the accuracy of ISS in predicting mortality, significant differences exist in mortality rates between patients with identical ISS from different AIS triplets (Aharonson-Daniel et al. 2006) as well as for patients with isolated injuries in different body regions and mechanism of injury (Aharonson-Daniel 2007).

This is only one of many reasons that it is beneficial to maintain information regarding the body region and nature of injury as proposed by the Barell matrix. The matrix that was initially built to simplify and unify the process of selecting diagnostic codes to describe an injury has become a tool for enhancing knowledge of injury characteristics. As described below, the Barell matrix has become a cornerstone of systematic injury profiling.

The Barell Matrix

Background

The Barell Body Region by Nature of Injury Diagnosis matrix is a tool for classifying injuries. It offers a standardized approach to data selection for analysis, using a two-dimensional array that includes all injury diagnosis codes. The matrix covers the range of ICD-9-CM codes 800–999 for injury and poisoning.

It serves to form uniform reports and as an elementary tool for standardized retrieval of injury cases for epidemiological, clinical, health promotion, economic, and management-oriented reports (Barell et al. 2002). Since its development and implementation it has been widely adopted and applied, leading to several derivative matrices as described below.

Original ICD-9-CM Matrix

Background

The matrix was conceptualized in 1996 by the late Vita Barell and a group of researchers in Israel and presented to members of the International Collaborative Effort (ICE) on Injury Statistics in 1997 (Barell 1996; Barell and Zadka 1996; Barell et al. 1999a, b). Concurrently, Mackenzie and Champion had worked on a parallel matrix in the USA (American college of surgeons, Committee On Trauma 1999). Collaborative work of both teams included experimentation on trauma registry and National Hospital Discharge Survey (NHDS) data until the unified version was finalized and approved at the ICE on injury statistics meeting of 2001 (Barell et al. 2002).

Conceptual Framework

As described above, the Injury and poisoning chapter details injury grouped by type or “nature” such as fractures, dislocations, sprains and strains, internal injuries, open wounds, amputations, injuries to blood vessels, contusions and superficial injuries, crush, burn, and nerve injuries. Each of these injury nature groups contains codes pertaining to potential harm to various parts of the human body; thus each injury “nature” sequence is organized similarly, from head to toe. An injury to a particular organ by various natures such as a combination of fractures and open wounds to a certain region necessitates looking for the body region of interest in each injury nature included. This action is not only tedious but also it may introduce variation when carried out by different people.

The original matrix displayed 12 “nature of injury” columns and 36 “body region rows” and placed each ICD-9-CM code in the range from 800 to 995 (traumatic injury) in a unique cell location. The matrix offered three predefined collapsed rows of body regions from 36 rows to 9 rows and then to 5 rows.

The matrix columns follow the sequence of ICD-9-CM codes where column A includes ICD-9 CM codes 800–829, column B contains codes 830–839, column C codes 840–848, and so on, with one exception made to separate amputations from open wounds (codes 885–887 and 895–897), as they were recognized as an important source of disability.

The rows differentiation is not as clear cut and was in fact the product of a long iterative process involving much thought and intent by a group of distinguished injury epidemiologists and based on statistical analysis of large databases. When applied in research, “Body region” rows are more likely to be collapsed or broken up for specific projects. Similarly, some specific injuries that were anticipated or proven to have different severity or outcome were separated in the original matrix. One such example are injuries to the spinal cord and spinal column that are separated due to increased severity in spinal cord injuries that are also associated with different types of rehabilitation needs and residual disability. Hip fractures are separated from other lower extremity fractures as they are often present as a single common diagnosis that affects different and specific populations (namely the elderly) (Barell et al. 2002). Hip fractures are also excluded from some registries and reports; thus, a simple elimination mechanism makes it easier to create a report that is internationally comparable.

Some concessions in assigning diagnoses to the matrix were unavoidable due to the structure of the ICD-9-CM coding itself, regarding conditions that could not be associated with a specific body region. Such were codes representing systemic injuries and codes that inherently refer to two or more body regions. As these codes did not fit into the matrix rows, they were placed in designated rows that were not classified by body region. Three types of entries are available in this group – codes that could not be fitted by the detailed body region row (the 36 row level) but could be fitted into a general body region category (the five row level) were put in a “regional other and unspecified” row such as “other and unspecified head and neck.” Codes describing multiple injuries, or generally defined injuries that were not in the same broad region, were placed in separate rows named “other and multiple specified” sites or “unspecified.” Finally, systemic injuries appear in the last row named “system-wide conditions.” These include: poisoning and toxic effects of substances, foreign bodies entering through orifice, early complications of trauma, late effects of injuries, and other and unspecified effects of external causes.

This last separation enables easy use of the matrix for analysis of trauma registry data that exclude cases with non-traumatic injuries based on the recommendations of the American College of Surgeons (American college of surgeons, Committee On Trauma 1999), in which diagnostic injury codes between 800 and 959.9 only are included.

ICD-9 CM codes for adverse effects (995.0–.4, .6–.7, .86–.89) and complications of surgical and medical care (996–999) are not included in the body of the matrix. This is consistent with the omission of comparable codes from the injury mortality diagnosis (IMD) tabulation framework developed by CDC (1997).

Injury Profiling Using the Matrix

In the USA, a SAS computer program was written that enabled reading a patient diagnosis and assigning it to a matrix cell. The SAS input statements can be found on the “injury data and resources” web page of the US Centers for Disease Control and Prevention.

Coders were instructed to choose the first-listed or “principal” diagnosis for this use. In Israel, availability of the matrix brought about a change in the approach toward recording multiple injuries. The Israeli SAS computer program took as input all diagnoses (up to 20 ICD-9-CM codes per patient) recorded in the medical file and updated the counts in the respective matrix cells. When doing so (using multiple diagnoses per patient), all injuries are counted, even if the patient has other, sometimes more severe injuries as well. This point is a central concept in the matrix application for data analysis and will be described in detail under the section “Multiple Injury Profiles” below.

A researcher attempting to profile an injury can define it by a cell (open wound of the chest), a combination of cells (open wound and contusion of the chest), a row (any injury to the chest), a column (casualties with a burn regardless of site), or any combination. The matrix can be used to identify common patterns of injuries in different circumstances, such as falls from height, etc. In brief, the matrix that originated from the need for a tool to select and retrieve data accurately and consistently provided the means for a major enhancement to injury statistics.

Matrix Use Review

The original manuscripts, describing the potential applications of the matrix, included data from the Israeli National Trauma Registry (ITR) and from the US NHDS (Barell et al. 2002). During subsequent years, the matrix has gained popularity. By 2010, over 100 studies using it had been published, most (69%) in peer reviewed journals, 28% in formal government or military periodicals. These publications originated in 13 countries in five continents written in four languages (English, Spanish, Italian, and Hebrew).

Most manuscripts used the matrix as a basis for epidemiological analyses, yet some used it as the basis for studies in health economics. The Matrix has been used either to select cases for inclusion in the analysis for a certain study or to classify and present data within the results. It has been used on various data sources such as hospital discharge data (Greenspan et al. 2006), trauma registry data (Aharonson-Daniel et al. 2007), injury surveillance data (Holder 2006; Horan and Mallonee 2003), and patient interview data (Warner et al. 2005).

In 2003, the Injury Surveillance Workgroup of the State and Territorial Injury Prevention Directors Association (STIPDA) published recommendations for systematic analysis of hospital discharge data (Injury Surveillance Workgroup 2003). The recommended analysis follows ten steps in a standard format to facilitate comparison. The Barell matrix is the basic construct of this report – providing the categorization of body region and nature of injury, and all but one step that deals with Ecodes instruct users to produce various analyses using the Barell matrix. One of the major reasons indicated for doing so is to enable comparative outputs. Assuming these guidelines will be followed, the use of the matrix is expected to grow in the future and the Matrix's preliminary aim of standardization and harmonization of injury data achieved. In Europe, as reflected in the final report of the European monitoring of trans-national injury and violence epidemiology project, EUROMOTIVE recommendations on “data quality and information exchange” included the adoption of a standardized diagnostic system similar to the Barell matrix, to establish a European framework to facilitate the presentation and comparison of national injury mortality rates throughout Europe (Christi and Stone 2004).

In an extensive evaluation of the comparability of different grouping schemes for mortality and morbidity, Lu et al. refer to the matrix, saying that the logic in grouping disease categories of similar etiology across different chapters has been widely adapted by the newly developed grouping schemes, such as the Global burden of Disease (GBD), Pan American Health Organization (PAHO) ICD-10 version, and the AHRQ-developed Clinical Classification Software (CCS) (Lu et al. 2005).

The “Barell+” Matrix

The “Barell+” Matrix was published in 2010 describing the creation of an expanded Barell matrix that aims to identify Traumatic Brain Injuries (TBI) of US Military Members (Wojcik et al. 2010). This matrix is actually a classification system for TBI that is derived by the Center for Army Medical Department Strategic Studies. The product named “Barell+” system is an expansion of the Barell body region by nature of injury diagnosis matrix, based on the Department of Defense severity classification system used for surveillance by the Defense and Veterans Brain Injury Center (DVBIC). The “Barell+” Matrix started with the Barell TBI category definitions adding 19 TBI-related diagnosis codes from the DVBIC classification into the resulting “Barell+” matrix.

Matrix for ICD-10

ICD-10 changed the direction of the classification system, so that the body region became the leading category (World Health Organization 1992b). This major change did not resolve the problem of seeking codes since to identify all fractures, all relevant codes must now be collated from the S and T sections of ICD-10, which are organized primarily by body region. A matrix was, therefore, devised to standardize the collection and presentation of the injury codes by body region of injury and nature of injury. As of March 2011, ICD-10-CM has not yet been put to use and ICD-10 is being used for mortality data only; this matrix was named the IMD matrix. The IMD matrix was designed to complement the Barell matrix, providing a tool for the organization and analysis of IMDs. The ICD-10 IMD matrix was developed to be as similar to the Barell matrix as possible, acknowledging the differences in data quality and coding systems (Fingerhut and Warner 2006). Differences between the matrices are primarily due to differences between the level of detail typically available for

morbidity and mortality. Concepts remained identical to those of the Barell matrix in regard to issues such as the exclusion of adverse effects, not elsewhere classified (T78), complications of surgical and medical care, not elsewhere classified (T80–T88) and their sequelae (T98.3).

Matrix for AIS

The matrix for AIS (Association for the Advancement of Automotive Medicine, Committee on Injury Scaling 1990) was created using AIS98 following the concept of the ICD-9-CM Barell matrix. Unlike the USA, where AIS is usually coded using ICDMAP (MacKenzie and Sacco 1997), in Israel, AIS is recorded and coded directly by trauma registrars. The AIS matrix is, therefore, valuable independently and not a derivative of the ICD-9-CM-based Barell matrix. AIS matrix rows depict the body regions and columns as follows: whole area, skin, blood vessels, nerves, internal organs, skeletal muscles tendons and ligaments, skeletal joints, skeletal bones, loss of consciousness, burn, and other. AIS injury codes are placed in the appropriate cells. The matrix is used in the same manner as the Barell matrix. An analysis conducted using this matrix is provided below.

Multiple Injury Profiles

Background

Multiple injuries in one patient often form a complex and challenging ailment with dire outcomes. The complexity associated with the treatment and outcome of multiple injuries brought about the development of various methods to assess risk to life or predict outcomes. The ISS (Baker et al. 1974; Baker and O’Neill 1976), which was the first widely acknowledged method and is still the most popular, was followed by the Anatomic Profile Score (Copes et al. 1990), the New ISS (Osler et al. 1997), Maximum AIS, ICISS, Revised Trauma Score, and others (Osler et al. 1996; MacKenzie et al. 1989; Gilpin and Nelson 1991). Nevertheless, these approaches focused on the contribution of multiple injuries to overall severity, to hospital workload, or to costs, without portraying the body regions or the nature of injury sustained. For many years, the primary diagnosis was used to describe an injury, losing information on the secondary, tertiary or fourth, fifth, and sixth injury in a single patient. In some cases, where selecting the primary diagnosis was not possible, “multiple injury” as a general category was classified (Champion et al. 2003), losing even the little information available in the coding of a primary diagnosis. Information regarding nature of additional injuries is important to compare case-mix and outcome between hospitals and countries and to support workforce planning that would take into account the specialties needed in a multidisciplinary trauma team. An injury profile that systematically looks at injury diagnosis combinations was first described in 2003, using the Barell body region by the nature of injury diagnosis matrix framework and named MIP (Aharonson-Daniel et al. 2003).

MIP Conceptual Framework

Detailing the injury natures and all body regions involved in an injury is valuable from clinical, epidemiological, and injury prevention perspectives.

Neither a primary diagnosis nor a severity score provide a comprehensive picture of the injury. MIP enable the identification of all cases with a specific injury and reflect both an accurate pattern of injury in the individual and a description of the hospital workload related to that injury.

		A	B	C	D	E
		FRACTURE	INTERNAL	OPEN WOUND	BURNS	OTHER
		800-829	850-854,860-869 952, 995.55	870-884, 890-894	940-949	830-854, 860-869,885-887, 895-897 900-904, 910-929, 950-957, 959
original Barell column		(A)	(D)	(E)	(J)	(B,C,F,G,H,I,K,L)
1	Head face and neck	800-804	850-854, 995.55	870-874	940-941	830, 848.0-.2
		807.5-.6			947.0	900, 910, 918, 920,921 925.1-.2, 950-951 953.0, 954.0, 957.0, 959.01,.09
2	Spine and back	805-806	952	/	/	839.0-.5
						847.0-.4
3	Torso	807.0-.4	860-868	875, 879.0-.7	942	839.6-.7, 846, 847.9, 848.3-.5
		808-809		876-878		901, 902.0-.5,.81-.82 911, 922, 926,959.1
				/		953.1-.3,.5, 954.1,.8-.9
4	Extremities	810-818	/	880-884	943-945	831-838, 840-845, 885-887, 895-897
		820-827		890-894		903, 904.0-.8, 912-917, 923, 924.0-.5 927-928, 953.4, 955, 959.2-.7
5	Other & unspecified	819, 828-829	869	879(.8-.9)	947.1-.2 , .8-.9	839.8-.9, 848.8-.9
					946, 948,949	902.87,.89 .,9, 904.9
						919, 924.8,.9, 929, 953.8, 956 953.9, 957.1,.8,.9, 959.8-.9

*The full Barell matrix appears in the June 2002 issue of Injury Prevention and can also be found at http://www.cdc.gov/nchs/injury/ice/barell_matrix.htm

Fig. 13.1 A modified 5×5 Barell Injury Diagnosis Matrix. Reproduced from “A new approach to the analysis of multiple injuries using data from a national trauma registry”, Aharonson-Daniel et al. (2003), with permission from BMJ Publishing Group Ltd, 2011

In essence, MIPs enable the use of combinations of multiple diagnoses to portray and analyze injuries in an individual and in a population.

The original framework for multiple injury diagnoses analysis was based on the Barell body region by the nature of injury matrix (Barell et al. 2002).

The level of detail and angle of inquiry is up to the individual researcher as the Matrix offers three general approaches or levels of analysis through a selection of diagnostic groups defined by matrix rows, columns, or cells.

Initially, patients’ ICD-9-CM diagnostic codes should be assigned to the appropriate matrix cells. Subsequently, analysis is conducted by matrix categories: rows, columns, or cells where the injuries are recorded. A program scans all patient records and builds vectors depicting the combination of cells forming each patients MIP. As doing so with the large 36×12 cell matrix often results in MIPs that are too sparse, grouping is applied so that a modified matrix is used for building the profiles. For example, the first published demonstration of building MIP using the Barell matrix (Aharonson-Daniel et al. 2003) took the most general level of five body regions (1) head and neck, (2) spine and back, (3) torso, (4) extremities, (5) other and unspecified and a modified display of Nature of injury columns: (A) fractures, (B) internal injuries, (C) open wounds, (D) burns, and (E) other as depicted in Fig. 13.1 above. For creating MIP, cells were represented by number–letter pairs defining patients’ placement in the matrix, similar to the way locations are noted on a chess board (A1-E5).

After placing the diagnoses into the 25-cell matrix, a patient can be described in a finite number of combinations while maintaining information on the body region and nature of his injury.

Table 13.1 A comparison of results obtained by “primary diagnosis” methodology with selected results obtained by multiple injury profiles (MIP) based on an AIS matrix

Injured body region		Primary diagnosis				Multiple injury profile (MIP)			
		Frequency distribution		ICU	Inpatient death	Frequency distribution		ICU	Inpatient death
MIP	Description	<i>n</i>	%	%	%	<i>n</i>	%	%	%
	Total	23,823	100.0	14.3	3.1	23,823	100.0		
H----	Head	8,263	34.7	22.5	5.8	3,036	12.7	13.1	3.1
-F----	Face	673	2.8	5.2	0.2	341	1.4	3.2	0.3
--C---	Chest	2,659	11.2	32.0	5.3	607	2.6	8.1	1.2
---A--	Abdomen	1,186	5.0	23.0	5.1	502	2.1	7.4	1.4
----E-	Extremities	7,512	31.5	4.6	0.6	4,992	21.0	2.7	0.3
----X	External	3,530	14.8	1.0	0.2	3,440	14.4	0.8	0.2
H---X	Head and external					1,949	8.2	9.2	1.1
---EX	Extremities and external					1,656	6.9	3.7	0.2
H--E-	Head and extremities					848	3.5	23.9	4.5
H-C---	Head and chest					274	1.2	49.8	21.3
H-C-E-	Head, chest, and extremities					254	1.1	64.2	22.6
	Other combinations					5,924	24.9	33.8	7.1

86 patients (0.4% had no diagnosis recorded)

“Other combinations” consists of a variety of combinations with 3, 4, 5, and 6 regions involved

Reproduced from Aharonson-Daniel et al. (2005), with permission from BMJ Publishing Group Ltd, 2011

A patient with an extremity fracture, an open wound to the head and an internal injury of the torso would be denoted by (A4-B3-C1). This combination is called an MIP. The meaning of “multiple” is then derived from the basic units in the analysis. Multiple injuries are injuries that fell into more than one group with the group defined as the basic unit studied: body region, injury nature, or matrix cells.

Often, for ease of presentation and to avoid lengthy lists of scarce profiles, common profiles are selected as a standard injury descriptor while infrequent profiles are combined in one category named “other multiple.” Each profile serves as a person’s characteristic (such as gender or ethnicity) and thus can then be examined for severity, treatments provided, service utilization, external cause of injury, and disposition.

Using the same conceptual approach, AIS injury diagnosis codes serve to report multiple injuries even more conveniently. As described above, AIS injury codes comprise six digits, the first of which depicts the body region of injury (Association for the Advancement of Automotive Medicine, Committee on Injury Scaling 1990). When using this digit to indicate the body region injured, a profile is created based on the six primary AIS body regions: *Head*, *Face*, *Chest*, *Abdomen*, *Extremities*, and *eXternal*. A patient with an isolated head injury would have a profile of H----, while a patient with a head and chest injury would have a profile of H-C---. Similar to the ICD-9-CM-based MIP, the frequency distribution of these profiles in the population can be studied and analyzed as any other categorical population characteristic.

Such an application is demonstrated in Table 13.1 where outcomes are compared with a simulation of a report using the common “primary injury” approach. The contribution of specific additional injuries to patient outcome is clearly demonstrated: Head injuries, usually associated with the highest (5.8%) inpatient death rate and high use of the intensive care unit (ICU) care (22.5%) are disclosed as not so severe when isolated (inpatient death rate of 3.1 and 13.1% ICU care).

The Application of MIP in Injury Epidemiology

Example 1: A study of 17,459 hospitalized patients recorded in the ITR following a road crash during a 4-year-period, 1997–2000 (Aharonson-Daniel et al. 2003). Injury data placed in a 5×5 matrix was analyzed using the approach described above.

Pedestrians and drivers had the highest proportion of multiple injuries. The analysis based on nature of injury demonstrated that 60% of the casualties sustained fractures, 57% of them as a multiple injury. Internal injuries were present in 43% of the population, 70% as part of a multiple injury. The most frequently injured body region in all road users except for motorcyclists was head and neck. Injuries to motorcyclists were mostly to the extremities. Pedestrians and drivers had the highest proportion of multiple injuries. When looking at the cell level, the two-dimensional distribution of injuries by body region and nature of injury shows that the most frequent injuries were internal injuries to the head and fractures of the extremities.

The use of MIP enabled counting even those diagnoses that would have been secondary using common analytic approaches, thus, often would not be recorded at all. The use of MIP exposes a different injury pattern than commonly seen. For example, the data demonstrated that in pedestrians, 25% had head injuries, 26% extremity injuries, and 15% had both. Drivers' injuries most commonly involved the head (21%), the torso (17%), or both (11%).

The contribution of a second and third injury to length of inpatient stay and death rates is clearly demonstrated in the data.

Example 2: A study of 23,909 casualties of transport accidents recorded in the ITR during 5 years (1998–2002) (Aharonson-Daniel et al. 2005). This study used AIS-based MIP and demonstrated the benefit of MIP that provided information on additional injuries when compared with traditional recording of primary injury. This addition reportedly ranged from 12% in head injuries to 270% for facial injuries. The results further demonstrated that MIP enabled a better description of the hospital workload. For example, it was known that among patients with an injury to the head, the proportion of patients staying in ICU was 22.5%. However, the examination of the MIP revealed that isolated head injuries resulted in only 13% ICU care, while among patients with MIP H---E- (head and extremities), 24% attended ICU, among patients with injuries to the head and chest (H-C---), 50% attended ICU, and among patients with injuries to the head, chest and extremities (H-C-E-), 64% attended ICU.

Assessing Injury Severity Using the Barell Matrix and MIP

The following chapter deals with injury severity measurement and survival outcome. Nevertheless, a few words are provided here regarding the relationship between severity indices and the instruments described above. The Barell matrix was not originally intended for the assessment of severity. However, several successful attempts to assign severity to matrix cells have been documented (Clark and Ahmad 2006).

Clark and Ahmad (2006) classified injury diagnoses of cases in the 2002 US Nationwide Inpatient Sample according to the Barell matrix. For each cell of the matrix, ICDMAP-90 (Association for the Advancement of Automotive Medicine, Committee on Injury Scaling 1990) was used to determine the predominant AIS score and body region and calculated the weighted proportion surviving among patients with any diagnosis in that cell. Maximum AIS, ISS, and minimum or product of cell for injured patients were calculated. Case survival was determined for different scores, and outcome models were compared to models using ISS calculated from ICDMAP-90 or using ICISS. The conclusion was that the Barell matrix allows severity scoring similar to that obtainable with ICDMAP-90 or ICISS.

Warner et al. (2006) attempted to rank severity among the injuries listed in the IMD matrix. Warner used the Barell matrix and the IMD matrix for categorizing diagnoses by body region and nature of injury. Bridging between morbidity data (ICD-9-CM) and mortality data (ICD-10) enabled determining survival risk ratios (SRR) by groups of ICD codes. After modifying the Barell matrix so it is comparable to the IMD matrix, an SRR was calculated for each cell by calculating the ratio between the number discharged alive divided by the number of all patients. A limitation of this method is that it does not take into account the existence of other injuries and calculates the risk of death for each cell as if it were an isolated injury. Accurate methods for tying between the Barell matrix derivatives, MIP, and severity, remain a future challenge.

Conclusion

For many years, methods that recorded injury severity did not monitor other injury characteristics. Epidemiological reports frequently describe only the primary diagnosis, thus losing information on additional injuries and underestimating the true burden of injury. Using the Barell matrix format to reduce the number of possible combinations of diagnoses, MIPs can be formed to enable presenting an improved picture of injury in a population.

References

- Aharonson-Daniel, L. (2007). *What affects outcomes of injuries with identical AIS severity? APHA 135th Annual Meeting*. Washington, DC. http://apha.confex.com/apha/135am/techprogram/paper_158822.htm
- Aharonson-Daniel, L., Boyko, V., Ziv, A., Avitzour, M., & Peleg, K. (2003). A new approach to the analysis of multiple injuries using data from a national trauma registry. *Injury Prevention, 9*, 156–162.
- Aharonson-Daniel, L., Giveon, A., & Peleg, K. (2005). Gaps in injury statistics: Multiple injury profiles reveal them and provide a comprehensive account. *Injury Prevention, 11*, 197–200.
- Aharonson-Daniel, L., Giveon, A., & Peleg, K. (2006). AIS triplets – different mortality predictions in identical ISS and NISS. *Journal of Trauma, 61*, 711–717.
- Aharonson-Daniel, L., Avitzour, M., Giveon, A., & Peleg, K. (2007). A decade to the Israeli Trauma Registry. *Israel Medical Association Journal, 9*, 347–351.
- American college of surgeons, Committee On Trauma. (1999). *Resources for optimal care of the injured patient*. Chicago, IL: American College of Surgeons.
- Association for the Advancement of Automotive Medicine, Committee on Injury Scaling. (1990). *The Abbreviated Injury Scale – 1990 revision (AIS-90)*. Des Plains, IL: Association for the Advancement of Automotive Medicine.
- Baker, S. P., & O'Neill, B. (1976). The injury severity score: An update. *Journal of Trauma, 16*(11), 882–885.
- Baker, S. P., O'Neill, B., Haddon, W., & Long, W. B. (1974). The severity score: A method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma, 14*, 187–196.
- Barell, V. (1996, November). *Assigning circumstances of injury in the emergency department*. Planning Session, International Collaborative Effort on Injury Statistics (ICE), Washington, DC.
- Barell, V., & Zadka, P. (1996, November). Matrix approach to classifying injury – application to Israel. *Proceedings of the international collaborative effort on injury statistics* (Vol. I), Washington, DC.
- Barell, V., Heruti, R. J., Abargel, A., & Ziv, A. (1999, June). The Israeli “Nature of Injury by Site” diagnostic matrix. *Proceeding of the International Collaborative Effort on Injury Statistics*, Symposium. Washington, DC: Centers for Disease Control and Prevention, National Center for Health Statistics.
- Barell, V., Heruti, R. J., Daniel-Aharonson, L., Ziv, A., & Abargel, A. (1999, June). Nature of injury by site diagnostic matrix: Differences between the Israeli and the U.S. versions. *Proceeding of the International Collaborative Effort on Injury Statistics*, Symposium. Washington, DC: Centers for Disease Control and Prevention, National Center for Health Statistics.
- Barell, V., Aharonson-Daniel, L., Fingerhut, L. A., Mackenzie, E. J., Ziv, A., Boyko, V., Abargel, A., Avitzour, M., & Heruti, R. J. (2002). An introduction to the Barell body region by nature of injury diagnosis matrix. *Injury Prevention, 8*, 91–96.

- Bertillon, J. (1913). Classification of the causes of death (abstract). In: *Transactions of the 15th International Congress on Hygiene Demography*. Washington, DC: Government Printing Offices.
- CDC. (1997). ICD-9 Framework for presenting injury mortality. *MMWR Recommendations and Reports*, 46(14), 1–30.
- Champion, H. R., Bellamy, R. F., Roberts, P., & Leppaniemi, A. (2003). A profile of combat injury. *Journal of Trauma*, 54, S13–S19.
- Christi, P., & Stone, D. H. (2004). *The final report of the European Monitoring of Trans-national Injury and Violence Epidemiology (EUROMOTIVE) project*. Glasgow, Scotland
- Clark, D. E., & Ahmad, S. (2006). Estimating injury severity using the Barell matrix. *Injury Prevention*, 12, 111–116.
- Committee on Medical Aspects of Automotive Safety. (1971). Rating the severity of tissue damage. I. The abbreviated scale. *JAMA*, 215, 277–280.
- Copes, W. S., Champion, H. R., Sacco, W. J., et al. (1990). Progress in characterizing anatomic injury. *Journal of Trauma*, 30, 1200–1207.
- Fingerhut, L. A., & Warner, M. (2006). The ICD-10 injury mortality diagnosis matrix. *Injury Prevention*, 12, 24–29.
- Gilpin, D. A., & Nelson, P. G. (1991). Revised trauma score: A triage tool in the accident and emergency department. *Injury*, 22(1), 35–37.
- Graunt, J. (1975). *Natural and political observations mentioned in a following index and made upon the bills of mortality*. New York: Arno Press. (Reprint of the 1662 ed. printed by T. Roycroft, London)
- Greenspan, A. I., Coronado, V. G., Mackenzie, E. J., Schulman, J., Pierce, B., & Provenzano, G. (2006). Injury hospitalizations: Using the nationwide inpatient sample. *Journal of Trauma*, 61(5), 1234–1243.
- Holder, Y. (2006). Injury surveillance systems in low and middle income countries (LMIC): Challenges prospects and lessons. *African Safety Promotion*, 4(1), 109–118.
- Horan, J. M., & Mallonee, S. (2003). Injury Surveillance. *Epidemiology Reviews*, 25, 24–42.
- Injury Surveillance Workgroup. (2003). *Consensus recommendations for using hospital discharge data for injury surveillance*. Marietta, GA: State and Territorial Injury Prevention Directors Association.
- Israel, R. A. (1978). The International Classification of Disease. Two hundred years of development. *Public Health Reports*, 93(2), 150–152.
- Langmuir, A. D. (1976). William Farr: Founder of modern concepts of surveillance. *International Journal of Epidemiology*, 5, 13–18.
- Lilienfeld, D. E. (2007). Celebration: William Farr (1807–1883) an appreciation on the 200th anniversary of his birth. *International Journal of Epidemiology*, 36(5), 985–987.
- Lu, T. H., Jen, I., Chou, Y. J., & Chang, H. J. (2005). Evaluating the comparability of different grouping schemes for mortality and morbidity. *Health Policy*, 71, 151–159.
- MacKenzie, E. J., & Sacco, W. J. (1997). *ICDMAP-90: A users guide*. Baltimore, MD: The Johns Hopkins University School of Public Health and Tri-Analytics, Inc.
- MacKenzie, E. J., Steinwachs, D. M., & Shankar, B. (1989). Classifying trauma severity based on hospital discharge diagnoses. *Medical Care*, 27, 412–422.
- Osler, T., Rutledge, R., Deis, J., & Bedrick, E. (1996). ICISS: An international classification of disease-based injury severity score. *Journal of Trauma*, 41, 380–388.
- Osler, T., Baker, S. P., & Long, W. (1997). A modification to the Injury Severity Score that both improves accuracy and simplifies scoring. *Journal of Trauma*, 43, 922–926.
- Stone, R. (1997). *Some British empiricists in the social sciences 1650–1900* (p. 260). Italy: Cambridge University Press.
- The International Classification of Diseases. <http://www.who.int/classifications/icd/ICDRevision/en/> Accessed February 15, 2011.
- Warner, M. W., Schenker, N., Heinen, M. A., & Fingerhut, L. A. (2005). The effects of recall on reporting injury and poisoning episodes in the National Health Interview. *Injury Prevention*, 11, 282–287.
- Warner, M., Fingerhut, L. A., & Chen, L. H. (2006). How can we pick which injury is most severe among all injury diagnoses listed on death certificates? *APHA 134th annual meeting*, Boston.
- Wojcik, B. E., Stein, C. R., Orosco, J., Bagg, K. A., & Humphrey, R. J. (2010). Creation of an expanded Barell matrix to identify traumatic brain injuries of U.S. military members. *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 7(3), 157–166.
- World Health Organization. (1949). *Manual of the international statistical classification of diseases, injuries, and causes of death* (6th rev.) Geneva: World Health Organization.
- World Health Organization. (1992). *International statistical classification of diseases and related health problems* (Vol. 1, 10th rev.) Geneva, Switzerland: World Health Organization.
- World Health Organization. (1992). *International statistical classification of diseases and related health problems* (10th rev.). Geneva: World Health Organization.

Chapter 14

Injury Severity Scaling

Maria Seguí-Gómez and Francisco J. Lopez-Valdes

Introduction

The notion of injury severity traces back to the book of Genesis (3, 15), where it is stated that “[God] will put enmity between you [the snake] and the woman, and between your offspring and hers; [the woman’s offspring] will crush your head, and you will strike his heel.” The term “injury severity,” however, was not used in an easily accessible research publication prior to a paper on aviation safety in 1962 (Pearson 1962). Ten more years elapsed before another indexed publication, this time on motor vehicle research, captured this term (Siegel 1972), even though one could claim an earlier publication using the term (Committee on Medical Aspects of Automotive Safety 1971) which is not indexed and thus not found in PubMed. Since then, more than 12,000 indexed papers (Pubmed accessed March 5, 2011) and countless other publications have included this term. Yet the actual meaning of the term “severe” as it applies to injury has never been formalized.

“Severity” is a word derived from Latin that describes the quality or condition of being “severe,” which in turn implies “harsh, strict or highly critical, as in treatment, serious or grave¹” (Guralnik 1986). Thus, in relation to injury, “severity” is a comparative term used in regard to a number of criteria including mortality risk, the need for more timely or more intensive care, or the risk of complications or lasting limitations in the future. That is, severity measures allow us to describe injuries above and beyond their existence or frequency (as with measures presented in the previous chapter) or to describe their long-term functional outcomes or even costs (described in separate chapters later in this book). Severity measures may have been included in any of the injury surveillance systems described elsewhere in this book, or they may be applied to special data collection efforts targeting

¹ In other Latin-derived languages, such as Spanish, the term “grave” is preferred to the term “severo” when using this concept as it applies to injury.

M. Seguí-Gómez, MD, ScD (✉)

European Center for Injury Prevention, Facultad de Medicina, Universidad de Navarra, c/Irunlarrea 1, Ed. Castaños S-200, Pamplona 31008, Spain

Johns Hopkins University, Baltimore, MD, USA

University of Virginia, Charlottesville, VA, USA

e-mail: msegui@unav.es

F.J. Lopez-Valdes, BEng

Center for Applied Biomechanics, University of Virginia, 4040 Lewis and Clark Drive, Charlottesville, VA 22911, USA

e-mail: fjl2j@virginia.edu

specific hypothesis testing. They may be used as outcome measures themselves when evaluating trends over time in effectiveness of different interventions, or they may be used as possible confounders (or case-mix adjusters) when evaluating the relationship between independent variables and any other outcome. In clinical settings, severity measures can be used in a variety of applications, including triaging (see Chap. 15), application of clinical guidelines, and prioritizing injuries amenable to environmental intervention, such as redesigned interiors of motor vehicles.

In 1984, MacKenzie summarized the then state-of-the-art severity measures as discussed in a conference on trauma indices cosponsored by the US National Center for Health Services Research and the American Trauma Society (Mackenzie 1984). In that meeting, criteria to evaluate the scales were developed, several scales were reviewed and the most promising ones identified, and, notably, three major reasons to develop and use these scales were characterized: (1) regionalization of trauma care (and, because of that, field triage), (2) policy development and funding (with related planning and retrospective evaluation), and (3) assistance in the development of intensive care units (with involvement in prognostic assessment and monitoring of patients in these units).

Since that 1984 publication, a few textbooks have included specific reference to the topic of injury severity measurement, either as a substantial component of a chapter (Berger and Mohan 1996, Chap. 2) or as full chapters (O’Keefe and Jurkovich 2001; Segui-Gomez 2007). Of the manuscripts that have been devoted to related issues, such as the choice of best scoring system to predict outcome after multiple trauma, we highlight Chawda et al. (2004).

In this chapter, we will present a concise description of the criteria or dimensions relevant to the development and application of injury severity measures, describe a selected number of them as well as the methodological implications related to their use, and suggest future areas of work.

Description of Dimensions Relevant to Injury Severity Scales

The growing number of injury severity scales over the last 50 years reflects the recognition of the importance of severity classification for patient care, resource planning, design of injury reduction interventions, and cost and effectiveness evaluation of injury prevention measures. Each scale has been developed according to specific aims that are not necessarily shared by the other systems, which can make the scales conceptually quite different. This chapter describes the most commonly used injury severity scales that apply to a broad range of patients, injuries, and injury mechanisms according to selected dimensions. The goal is that the reader may gain from the subsequent paragraphs a clear understanding of the development of each scale and guidance in selection of the appropriate scale to be used in a particular situation.

The selected dimensions are grouped into two different categories: those related to development of the scale and those related to usage and application of the scale. While a detailed description of the selected injury severity scales is given in the following section, we summarize here how each scale relates to identified dimensions. Table 14.1 shows the acronyms and full names of the injury scales selected to be discussed throughout this chapter.

1. Dimensions considered in the development of the scales:

- Population: whether the scale applies to people of all ages (to date, no scale has been developed targeting just one of the two genders).
- Injury nature: are all types of injuries included in the system? Or does it exclusively address head injury, long bone fracture, or others?
- Injury mechanism: was the scale developed to describe a single injury mechanism (e.g., motor vehicle crashes, falls, drowning) or is it applicable to several or all of them?
- Outcome: which is the outcome targeted by the scale (e.g., death, length of treatment, impairment)?

Table 14.1 Acronyms and full names of the injury severity systems described in the chapter

Acronym	Complete name
KABCOU	None specifically
AIS	Abbreviated Injury Scale
MAIS	Maximum AIS
ISS	Injury Severity Score
NISS	New Injury Severity Score
AP	Anatomic Profile
GCS	Glasgow Coma Scale
RTS	Revised Trauma Score
TRISS	Trauma and Injury Severity Score
DCP	Damage Control in Polytrauma
ICISS	International Classification of Injury Severity Score
HARM	Harborview Assessment for Risk of Mortality Score

- Observation unit: whether the scale evaluates the severity of a single injury or the likely overall effect of all the injuries in one subject.
 - Anatomical/physiological description: is the scale based on anatomical or physiological data or both?
 - Goal: does the scale predict or measure the outcome?
 - If the objective is to predict outcome, was this prediction based on real data or expert judgment? This dimension relates to how the injury score was assigned to the injuries (whether based on the opinion of experts assessing, for example, threat to life, or real data used in calculating probability of death).
 - If there is a prediction, has the scale been validated against real-world data?
 - Has the measure been validated in regard to reliability and other scale psychometric properties?
 - Source: whether the score is developed by collecting original data (primary) or is based on preexisting data, for example, codes collected for other related or unrelated purpose (secondary).
 - Dynamic/static: does the score evolve over time or remain the same regardless of the evolution of the patient?
2. Dimensions to be considered in the use of the scales:
- Continuous/categorical: the type of statistical values describing the severity score that the variable creates.
 - Best summary statistics: which statistics are recommended for use with that particular injury metric?
 - Translation into other scales: is that scale used to derive other injury severity scales?
 - Scenario (where the scale is applicable): whether it can be applied at the scene of injury to influence initial triage to a certain type of care, or during patient care (e.g., at hospital or intensive care unit), or only retrospectively.
 - Applications: whether the scale can be/has been used in field triage, in evaluation and planning, in clinical assessment and monitoring of patients, or in biomechanics research on injury thresholds.

Table 14.2 summarizes the information relating to the dimensions involved in the development of each injury severity system and Table 14.3 describes the main factors to be considered in the use and application of each scale.

Table 14.2 Comparison between selected injury severity scales across dimensions relevant to their development

Acronym	Population	Injury nature	Injury mechanism	Goal	Outcome	Observation unit	Anatomical/physiological description	Real data/expert judgment	Source	Dynamic/static	Validation
KABCOU	All	All	All	Measure	Death, impairment	Person	None	Real data	Primary	Static ^a	None
AIS	All ^b	All	Impact, penetrating, blast	Predict	Death and others	Injury	Anatomical ^c	Expert judgment	Primary, secondary (ICD)	Static	Some
ISS	All ^b	All	Same as AIS	Predict	Death	Person	Same as AIS	Same as AIS	Secondary (AIS)	Static	Some
NISS	All ^b	All	Same as AIS	Predict	Death	Person	Same as AIS	Same as AIS	Secondary (AIS)	Static	Some
AP	All	All	Same as AIS	Predict	Death	Person	Same as AIS	Same as AIS	Secondary	Static	Some
GCS	All	Head injury	All	Predict	Brain damage	Person	Physiological	Real data	Primary	Dynamic	Yes
RTS	All	All	All	Predict	Death and others	Person	Physiological	Real data	Primary	Dynamic	Some
TRISS	All	All	All	Predict	Death and others	Person	Combined	Combined	Secondary (AIS, RTS)	Dynamic	Some
DCP	All	Bone fractures	All	Measure	Damage	Person	Combined	Real data	Secondary (AIS, ISS)	Dynamic	Some
ICISS	All	All	All	Predict	Death	Person ^d	Combined ^e	Real data	Secondary (ICD)	Static	Some
HARM	All	All	All	Predict	Death	Person ^d	Combined ^e	Real data	Secondary (ICD, others)	Static	Little

^aLength of impairment and time interval until death are not defined uniquely

^bLimited application to the pediatric population

^cVery limited physiological description on a handful of injuries

^dSurvival risk ratios are calculated for individual injuries

^eCategorized as combined because ICD codes include a mix of anatomical and physiological data

Table 14.3 Comparison between the selected injury severity scales across dimensions relevant to their usage

Acronym	Continuous/ categorical	Best summary statistics	Translation into other scales	Scenario	Applications
KABCOU	Categorical	Frequencies	None	At scene	Evaluation and planning
AIS ^a	Ordinal	Frequencies, median, mode	Yes: ISS, NISS	Retrospectively (from medical chart)	Evaluation and planning, biomechanics
ISS ^a	Ordinal	Frequencies of ranges, median	TRISS	Retrospectively (from medical chart)	Evaluation and planning, triage ^b
NISS ^a	Ordinal	Frequencies of ranges, median	TRISS	Retrospectively (from medical chart)	Evaluation and planning
AP	Categorical	Frequencies	None	Retrospectively	Evaluation and planning
GCS ^a	Continuous (3–15)	Means (SD), frequencies	None	At scene	Triage, evaluation and planning, patient monitoring
RTS	Continuous (0–12)	Means (SD)	Yes: TRISS	At scene	Triage, patient monitoring
TRISS	Continuous	Means (SD)	None	At scene	Triage, patient monitoring
DCP	Categorical	Frequencies	None	At scene ^c	Patient monitoring
ICISS	Continuous	Means (SD)	None	Retrospectively	Evaluation and planning, biomechanics ^d
HARM	Continuous	Means (SD)	None	Retrospectively	Evaluation and planning

^aAIS, ISS, NISS, and GCS scores must be integers

^bSome trauma centers prioritize treatment to patients exhibiting ISS > 15 (although it is unclear how the score is computed at scene)

^cAs stated by author, however, inclusion of AIS/ISS information increases the feasibility of doing this

^dAs a potential application, the use of ICISS in biomechanics has not been reported to date

Description of Injury Severity Scales and Methodological Implications of Their Use

Traditionally, official figures, for example, on road or occupational injuries, have categorized victims into fatal, serious or slightly injured, or variations on this terminology, including the term “impaired.” For example, with regard to motor vehicle injuries, in most countries, the difference between serious and slight injury rests on whether the patient was admitted to the hospital, as reflected in the CARE or IRTAD international databases that collect official statistics on police-reported motor vehicle crashes (ETSC 2001; CARE 2011; IRTAD 2011).

KABCOU is a classification system closely related to the concept described in the previous paragraph (NASS CDS 2009) which classifies injuries into six levels: fatal (K), incapacitating (A), non-incapacitating (B), possible (C), no injury (O), or unknown injury (U). As far back as one can trace the use of injury severity, the desire to know which injuries can cause the death of the patient, which ones require hospital admission, or which ones result in the greatest disability to the subject has been a constant. Yet, there are numerous obstacles to an easy characterization of this concept. Even with an outcome as obvious as death, its use is generally restricted to death that occurs within a specified period after injury, and is therefore susceptible to discussion (e.g., 24 h or 30 days or any time for as long as the primary cause of death in the death certificate is noted to be an external cause (i.e., a V–Z code in the International Classification of Diseases 10th version)). For the remainder of this chapter, we will focus on measures primarily addressing threat to life of the subjects and leave the notions of disability and others for other chapters.

The *Abbreviated Injury Severity* scale (AIS) was developed in 1969 by a joint committee formed by the American Medical Association (AMA), the Society of Automotive Engineers (SAE), and the

Association for the Advancement of the Automotive Medicine (AAAM), which committed themselves to the development of a scale to classify injuries and their severity. The first scale was published in 1971 in the *Journal of the AMA (Committee on Medical Aspects of Automotive Safety 1971)*. The AAAM assumed the lead role for continued development of the scale in 1973. The goal of the new scale was to facilitate the classification of injuries with regard to both anatomic nature of the injury and an estimation of the severity of the injury. Severity was defined for the occasion as a combination of energy required to cause the injury, threat to life, risk of permanent impairment, length of treatment, and injury incidence. The 1985 revision of the scale (AAAM 1985) introduced a numeric system consisting of 7 characters that assigned a unique code number to each injury description contained in the dictionary. The first six digits of the code provide the description of the injury, while the seventh and last one (which is separated from the others by a dot) is the assigned AIS severity code that ranges from 1 (minor) to 6 (potentially unsurvivable).

The same format has prevailed intact until the last version of the scale, the 2005 revision (AAAM 2005), which has been recently improved into the AIS 2005 revision 2008 update (AAAM 2008). The 2005 revision offers the possibility of augmenting the code up to a total of 15 digits by using eight optional digits placed after the severity digit. The first four optional digits (called localizers) provide information on the location of the injury (e.g., proximal, distal, bilateral, anterior, superior), and the remaining four serve as descriptors of the intent or circumstances of injury (e.g., intentional vs. nonintentional, child seat forward facing, vehicle-occupant rear seat).

Table 14.4 shows the information that can be extracted from each of the digits of the code.

Without considering the multiplicity introduced by the localizers and descriptors of injury cause, the AIS 2005 revision 2008 update contains approximately 2,000 injury descriptors (four times more than the number published in the first dictionary). The AIS dictionary has been expanded over the years to deal with nonmechanical injuries to parallel the advancement of medical research and to develop comparability with other injury scoring systems. The last revision gives special importance to osseous injuries due to a close collaboration between the AIS Injury Scaling Committee and the Orthopedic Trauma Association (OTA).

For the sake of illustrating the successive changes undergone in the coding of one injury, a fracture of the tibia would have been assigned the code 92401.2 (tibia fracture, not further specified) in the AIS 1985 revision; in the AIS 1990 revision 1998 update, the same injury could have been classified with the code 853404.2 (tibia fracture, any type, without displacement), and the AIS 2005 revision 2008 update would have improved the description of the injury by using the code 854151.2 (proximal tibia fracture, extra-articular).

With regard to the six descriptive digits (predot code), the digit identifying the body region (the first digit) has been commonly used to describe the location of the injury, and therefore, it is common to find injuries classified according to the nine body regions defined in the AIS: head, face, neck, thorax, abdomen, spine (including spinal cord), upper extremities, lower extremities, and not further specified. This same body region classification is used to compute the maximum AIS or MAIS in patients sustaining multiple injuries, as it will be described later in this chapter.

Given the focus of the chapter on the measurement of injury severity, further discussion of the AIS will be limited to the seventh digit of the code. The severity of a given injury is defined by an ordinal scale from 0 to 6, in which 0 means no injury and 6 means maximal severity (virtually unsurvivable) injury. The codes in between these two extremes range from minor to critical severity. A value of 9 can be assigned to the seventh digit in the case of unknown injury severity.

The AIS severity score has been developed by consensus of a group of experts over the life of the AIS. Although threat to life is one of the most important dimensions used by the expert panel to assign the severity score, there are other dimensions that were also considered within the AIS (energy dissipation, impairment, temporary disability, treatment cost). Therefore, the fact that a patient has an injury with an AIS score of 6 does not necessarily mean that he died. Similarly, the fact of the death of the victim does not necessarily mean that the AIS score was 6. Each injury descriptor (predot digits) is associated with a unique severity score. Given the high number of codes and the several

Table 14.4 Interpreting AIS 2005 update 2008 numerical codes by location of digit

1st digit	2nd digit	3rd and 4th digits	5th and 6th digits	7th digit	(8th and 9th digits)	(10th and 11th digits)	(13th through 15th digits)
Body region	Type of anatomical structure	Anatomical structure	Level	Severity	Localizer 1	Localizer 2	Injury cause
1 = Head	1 = Depends on body region (see manual)	Values change by body region (see manual)	In the head: 02 = length of unconsciousness; 04, 06, 08 = level of consciousness; 10 = concussion	1 = Minor	Right, left, middle, bilateral, multiple, upper, lower, and other combinations (see manual)	Finger/toe, ribs, teeth (see manual)	Intent/type of vehicle/others (see manual)
2 = Face	2 = Vessels			2 = Moderate			
3 = Neck	3 = Nerves			3 = Serious			
4 = Thorax	4 = Organs (including muscles and ligaments)			4 = Severe			
5 = Abdomen and pelvic contents	5 = Skeletal (including joints)			5 = Critical			
6 = Spine	6 = Head — loss of consciousness		In the spine 02 = Cervical 04 = Thoracic 06 = Lumbar	6 = Maximal (currently untreatable)			
7 = Upper extremity							
8 = Lower extremity, pelvis, and buttocks							
9 = External and thermal injuries				9 = Not further specified			

(continued)

Table 14.4 (continued)

1st digit	2nd digit	3rd and 4th digits	5th and 6th digits	7th digit	(8th and 9th digits)	(10th and 11th digits)	(13th through 15th digits)
0 = Other trauma	0 = Whole area		In whole area				
			02 = Abrasion				
			04 = Contusion				
			06 = Laceration				
			07 = Avulsion				
			10 = Amputation				
			20 = Burn				
			30 = Crush				
			40 = Degloving				
			50 = Not further specified				
			60 = Penetrating				
			90 = Nonmechanic trauma				
			<i>n</i>				

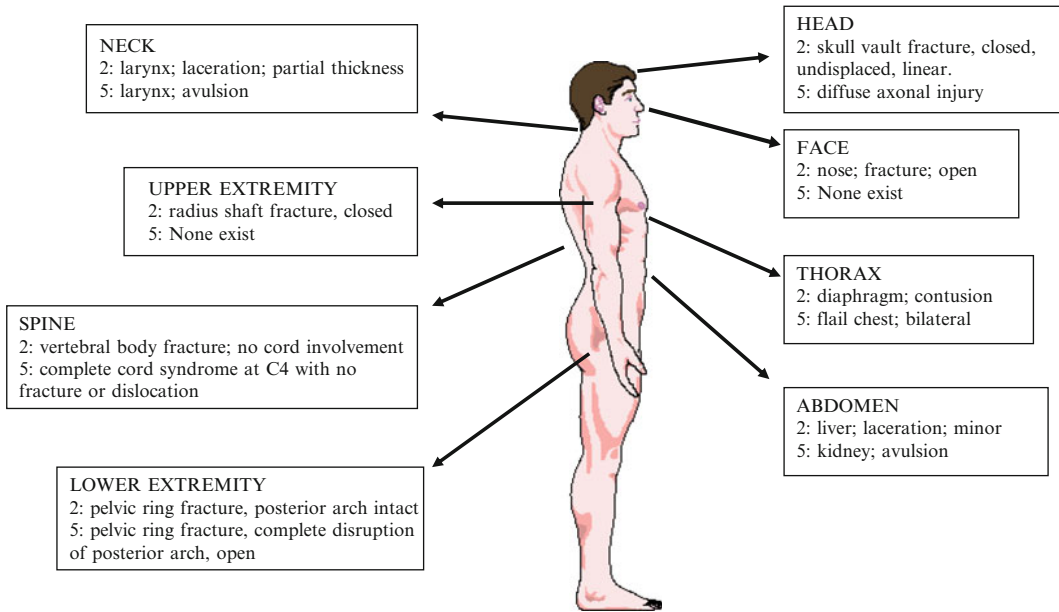


Fig. 14.1 Selected AIS 2005 update 2008, injuries assigned codes of 2 or 5 (in severity) by body region

coding rules that have been developed over the years, training is virtually essential for anyone choosing to use AIS to code injuries.

In the process of assigning the severity scores to specific injuries, the members of the AIS Injury Scaling Committee were asked to make their assessments independent of age and sex of the victim (actually, a few codes were assigned different severity codes if the victim was a child) and to assume the patient would undergo optimal medical treatment. Also, the victims were to be assumed to be perfectly healthy before sustaining the injury. This is of particular relevance since there are other systems that are based on empirically derived severity instead of consensus-based severity as in the case of the AIS. Last, panel members were to focus on the anatomical injury to produce a severity score; only for a handful of injuries is physiological information (e.g., in relation to the duration of the loss of consciousness) used as a modifier to the severity score. Thus, AIS is considered an “anatomical” severity score that remains constant from the time of event onward.

There are numerous studies confirming that the consensus of the experts of the AIS Injury Scaling Committee performs well as a measure of mortality. In an analysis of the correlation between AIS 1990 revision 1998 update and survival using data from the National Trauma Data Bank (NTDB) from patients sustaining a single injury, it was shown that AIS 6 correlates with 80% likelihood of death, AIS 5 with 40%, AIS 4 with 15%, AIS 3 with 3%, AIS 2 with 1%, and AIS 1 with less than 1% probability of death (AAAM 2005).

Figure 14.1 shows an example of several injuries of different severity distributed according to the AIS body regions.

Some suggest that one deficiency of the AIS is that the likelihood of death is different depending on the body region (O’Keefe and Jurkovich 2001), for example, the threat to life of an AIS 5 head injury would be higher than that of an AIS 5 lower-extremity injury. The origin of this criticism can be found in an official document issued by the AIS Committee (Petrucci et al. 1981), which describes how the ordinal nature of the scale results in some injuries that actually differ in severity being assigned to the same severity category. Another limitation of the AIS is the poor correlation found between the severity scores and the prediction of impairment or length of treatment. Last, the

fact of having a single code per each injury poses some difficulties at the time of measuring the injury severity in polytraumatized patients such as the ones typically found in car crashes.

Despite its limitations, the AIS is the injury classification most used in injury prevention research worldwide (and even more so in motor vehicle-related injuries and biomechanics). Numerous information systems have adopted the AIS to classify the injuries resulting from a crash. In the USA, the National Automotive Sampling System Crashworthiness Data System (NASS CDS) slightly modified the conventional AIS (1990 revision), adding an additional code to specify the lateral aspect of the injury. The British Cooperative Crash Investigation Study (CCIS) would be the best example of the use of the AIS in a motor vehicle-crash database in Europe. Both the NASS CDS and the CCIS have incorporated the AIS 2005 revision 2008 update in their investigations.

Over the years, the AIS dictionary and manual have been translated into French, German, Italian, Chinese, Spanish, and Japanese. As mentioned earlier, the AIS has also become the standard injury severity measurement used in experimental biomechanics. A vast majority of the biomechanical injury criteria correlate a physical property (e.g., force, acceleration, deflection) with the probability of sustaining an injury of a specific AIS level. The use of the scale is so widespread in the automotive sector that most standards and regulations related to the passive safety of the occupants are expressed in terms of the AIS (US NCAP, FMVSS, ECE regulations, Euro NCAP). For a review of uses of the scale, please see Watchko et al. (under review).

The extensive use of the AIS propitiated the interest of the clinicians in obtaining an algorithm that allowed converting from ICD codes to AIS codes in order to extend its application to larger databases containing ICD information. The first and best known is software called ICDMAP, created to convert ICD-9-MC into AIS 1985 revision (MacKenzie et al. 1989) although it was updated to transform to AIS 1990 revision (Center for Injury Research and Policy of the Johns Hopkins University School of Public Health 1998). Over the years, several other programs to convert between the two systems have been published (Kingma et al. 1994a, b), recently including one allowing the conversion between ICD-10 and AIS 1990 revision (European Center for Injury Prevention 2006). Similarly, the AAAM in its 2005 AIS version provides a conversion algorithm into AIS 1990 update 1998 codes.

The *Maximum AIS* (MAIS) is probably the first attempt of using AIS information to reflect overall injury severity in a patient sustaining multiple injuries. The MAIS (which is simply the highest severity value of any of the considered AIS codes) can be readily assigned for each of the AIS-defined body regions or as an overall score of the patient regardless of the location. However, Baker et al. (1974) found that the overall most severe AIS injury was nonlinearly correlated with death rates and that the severity of injuries in other body regions strongly influenced mortality. These findings resulted in the creation of the Injury Severity Score.

The *Injury Severity Score* (ISS) is the sum of the squares of the three most severe AIS scores in three different body regions (Baker et al. 1974). Six body regions are considered in the case of the ISS: head and neck, face, thorax, abdominal and pelvic organs, extremities and pelvis, and external structures. Like the AIS, the ISS is an ordinal scale (Stevenson et al. 2001a, b) ranging from 0 (no injury) to 75 (if there is at least one AIS 6 code among the three or three AIS 5 codes). The ISS takes the value of 99 in case the severity of any of the injuries is unknown (AIS 9). Some values in the interval (0, 75) are not possible to obtain, such as 7, 15, 23, 28, 31, 37, 39, 40, 44, 46, 47, 49, 52, 53, 55, 56, 58, 60–65, and 67–74.

The *New ISS* (NISS) (Osler et al. 1997) is a variation of the ISS where, instead of using the three most severe AIS codes from three different body regions, one considers the three most severe injuries regardless of their location. NISS is also an ordinal scale ranging between 0 and 75, and, as with the ISS, some values in that interval are not possible. ISS and NISS are well correlated with mortality, which was the main goal behind the development of both systems; several studies have shown that the NISS correlates more closely with mortality than the ISS (Brenneman et al. 1998; Lavoie et al. 2004; Sacco et al. 1999). In addition, the NISS is much simpler to calculate because body regions can be ignored.

The *Anatomic Profile* (AP) is the most recent scale derived from the AIS (Copes et al. 1990). AP also combines AIS codes to produce an alphabetic scale from A to D. Injuries within the A category correspond to AIS 3+ injuries to the head, brain, or spinal cord. AIS 3+ injuries to the thorax or the front of the neck would be classified as B. All other injuries coded as serious (AIS 3) or severe (AIS 4) through maximal (AIS 6) in the remaining body regions would fall into the C category. Facial injuries were placed in D, given the low mortality rate associated with these injuries in the absence of serious brain or head injury. AP was developed with the goal of predicting survival probability and as an alternative to ISS.

The *Glasgow Coma Scale* (GCS) is one of the best known severity scores and the oldest standing injury severity measure based on physiological data only (Teasdale and Jennett 1974). It was also developed during the 1970s and was originally designed to measure the severity of brain injuries with traumatic, vascular, or infectious origin. This scale is also ordinal, and the values range between 3 (deep coma) and 15 (normal). The global score is obtained by adding three individual scores corresponding to motor response, verbal response, and eye response. Despite being widely used, some current prehospital admission practices such as intubation and administration of pain killers can preclude its use.

The *Revised Trauma Score* (RTS) combines information of several physiological parameters such as cardiac frequency, systolic arterial pressure, and breathing frequency of the subject (Champion et al. 1989).

The *Trauma and Injury Severity Score* (TRISS) combines the RTS with information from the injury mechanism (e.g., penetrating yes/no), age, and the ISS to estimate the probability of death.

Borrowing from the concept of “damage control” implemented in the USA to treat penetrating abdominal trauma in the 1980s, Giannoudis (2003) proposed the *Damage Control in Polytrauma* (DCP) which is a severity rating system targeting multiply-injured patients sustaining fractures of long bones and/or pelvis (albeit not exclusively). This system aimed to incorporate the notion that inflammatory reaction to the injuries could dramatically alter the outcome of the patients, and it was designed to assist practitioners in deciding on whether and when to intervene. In this system, patients are categorized into being stable, borderline, and unstable in extremis, depending on several anatomical or physiological parameters. For example, for a patient to be labeled borderline, he or she must exhibit at least one of the following: (a) multiple injuries with ISS > 20 including a thoracic injury AIS > 2, (b) multiple injuries with abdominal/pelvic trauma and initial systolic blood pressures < 90 mmHg, (c) ISS > 40, (d) radiographic evidence of bilateral pulmonary contusion, (e) initial mean pulmonary arterial pressure > 24 mmHg, or (f) pulmonary artery pressure increase during intramedullary nailing > 6 mmHg.

Because of the amount of clinical information that is needed to use any of the above four measures (GCS, RTS, TRISS, or DCP), they are not as frequently used as those based on an anatomical description. Further, all these measures that use physiological parameters have to incorporate the fact that physiological conditions change in time. Therefore, comparison between different scores would make sense only if the time since injury is also prescribed (and comparable between the systems).

The *International Classification of Injury Severity Score* (ICISS) derived *Survival Risk Ratios* (SRRs) for every ICD-9 injury category using the North Carolina Hospital Discharge Registry (Osler et al. 1996). These SRRs are calculated as the ratio of the number of times of a given ICD-9 code occurs in a surviving patient to the total number of occurrences of that code. The ICISS is defined as the product of all the survival risk ratios for each of an individual patient’s injuries (for as many as ten different injuries) (Osler et al. 1996). Thus, the survival risk for a given subject decreases if either there is an injury with a very low-associated survival risk or there are multiple injuries even if their survival risks are moderate. Table 14.5 shows a selection of ICD-9-CM codes with the highest mortality risk as derived by the ICISS.

Although over the years other parameters have been considered for addition to ICISS (such as age, injury mechanism, or even the RTS), the ICISS remains as an injury measurement based on the ICD description of the injuries.

Table 14.5 Random selection of 10 out of 100 ICD-9-MC codes with the lowest SRRs presented, sorted from lowest to highest SRR value (source: Osler et al. 1996)

SRR	ICD-9-CM	Description
0	852.35	Subdural hemorrhage, continuing LOC
0.41	902.33	Portal vein injury
0.51	902.0	Abdominal aorta injury
0.53	901.2	Superior vena cava injury
0.64	850.4	Concussion, continuing LOC
0.68	902.53	Iliac artery injury
0.72	958.4	Traumatic shock
0.74	902.31	Superior mesenteric vein injury
0.79	806.04	Cervical fracture, C1–C4
0.79	902.42	Renal vein injury

The *Harborview Assessment for Risk of Mortality Score* (HARM) (Al West et al. 2000) groups the ICD-9-CM codes into 109 categories and also incorporates information on the injury mechanism (e.g., traffic crash, fall), intentional vs. unintentional causes, preexisting medical conditions, and age and gender of the subject. HARM can also handle multiple injuries in one subject. All the information needed to use HARM is generally available in hospital admission databases. A comparison between the survival risk associated with ICISS and the different ICD-9 codes in HARM reveals a great similarity between both systems. The ten most severe injuries in HARM would be those that most increase the risk of death. Thus, the most lethal injury according to HARM is loss of consciousness for more than 24 h (95% increase in mortality risk), followed by full-thickness cardiac lacerations (67% increase) and unspecified cardiac injuries (32%), and next, complete spinal cord injury at the level of C4 or above (31% increase of the risk of death), injuries to the superior vena cava or innominate vein (28%), pulmonary laceration (27%), cardiac contusion (22%), traumatic amputation above the knee (21%), major laceration of the liver (15%), and injuries to the thoracic aorta or great vessels (14%). The reader is reminded that these estimations of risk are adjusted by sex and gender, injury mechanism, and all other aforementioned variables involved in HARM. It is relevant to note here that this measure is not to be confused with HARM as defined by the US National Highway Traffic Safety Administration, which is a metric to value cost of injuries (Miller et al. 1995).

In fact, the two last measures are efforts to provide a severity score, as with the AIS. However, the methods used within each of the systems to derive the mortality risk estimations from empirical data can be also questionable. For instance, both ICISS and HARM make use of hospital data, and therefore, mortality is frequently calculated at discharge ignoring all deaths prior to hospitalization, which in some instances can amount to more than 50% of the deaths. On a more general note, the transferability to other circumstances and locations of these systems (that have been developed based on information of specific regions and hospitals within the USA) must be assessed. ICISS is being more commonly used than HARM in the literature.

Challenges for Future Development

As stated in Chawda et al. (2004), “the plethora of available scoring systems for trauma [severity] suggests that there is a need for a universally applicable system, but this goal may be difficult to achieve.” Part of this difficulty may relate to the fact that the concept of severity is somewhat ill-defined. Since the 1960s, short-term survivability seems to be at the heart of most developed metrics, yet other concepts, such as difficulty of treatment and likelihood of long-term impairment, have cluttered its operational definition. Out of a plethora, this chapter presents a selection of measures that apply

to most populations, injuries, and injury mechanisms and that are widely found in the literature. Even these lack definitional precision of the term “severity.” Of the scales presented here, only the AIS, the ISS, and the RTS (in its first version, called Trauma Score) were included in the 1984 review by MacKenzie (1984).

As with any other health measure, severity scores should be subjected to rigorous evaluation for validity and reproducibility. Validity can only be measured if the outcome under evaluation is clearly defined. For example, how and whether to combine AIS scores into any mathematical model to derive patient-based severity scores can only be determined if, for example, predicting death, is set as the objective. In this regard, definitional issues need to be addressed across all measures, and whether their validity differs depending on the subpopulation must also be considered. For example, whether the pediatric-related modification of a few AIS scores in the 2005 version is sufficient to adjust the validity of the measure in this subpopulation needs to be investigated.

Regarding reliability, since the mid-1980s, there is a call for the rigorous application of scoring criteria (MacKenzie 1984). In the case of the AIS, its parent organization (AAAM) has developed an extensive in-person and online training program around the world (www.aaam.org). However, the number of users trained to code AIS or most other scales remains low, as revealed when publications indicate misuse or misunderstanding of the codes (Watchko et al. under review).

When scores range between several values and decisions to transfer and/or treat patients are made based on those scores, rigorous analysis of specificity and sensitivity, including development of receiver operating characteristic (ROC) curves, is due. Due to the insufficient research on these topics for most of the scales, more work is needed, particularly in the triage and decision-making application of these measures.

These measures also vary in the mathematical nature of the numbers produced; some are categorical variables, others, ordinal, yet others, continuous. Often, they are all used as continuous variables, resulting in inappropriate arithmetical operations and statistical analysis. Users need to be mindful of the actual analytical possibilities of the measures.

Since the objectives are severalfold, it is likely that no scale serves best for all purposes, particularly in triage and clinical applications. Yet, when it comes to evaluation and planning or biomechanical applications, the AIS, SRRs, and injury-specific HARM scores, as well as their composites to address overall severity, are being widely used and in somewhat of a competition. Some researchers argue against the consensus-derived AIS as assessed by experts who belong to the AIS Committee. Some even produced real-world probability-of-death ratios for the predot AIS codes of motor vehicle-injury victims collected under the US National Highway Traffic Safety Administration National Automotive Sampling System Crashworthiness Dataset (Martin and Eppinger 2003). Nevertheless, the fact is that real-world probability-based measures such as ICISS or HARM are not exempted from criticism. For example, which data to use and where to apply become crucial. For example, are SRRs derived from hospital discharges in the 1990s in North Carolina applicable to 2010 hospitalized injury patients in Spain? Time- and space-external validity becomes an important parameter to assess.

In the years ahead, it is possible that redefinition and refinement of the concept of injury severity will allow for further development of existing or newly developed scales. At the population level, and in regard to program evaluation purposes, severity measures derived from already collected data will continue to prevail both as outcome variables and as independent variables (and possible confounders) in multivariate analyses. It will be interesting to see whether the field will be dominated by SRRs (and derivatives) or the AIS (or derivatives) computed using algorithms based on ICD.

Acknowledgments Dr. Segui-Gomez’s efforts were supported by the European Center for Injury Prevention at Universidad de Navarra and the AAAM AIS Reference Center funding. Dr. Segui-Gomez chaired the AAAM AIS Committee at the time of writing; AIS related contents have been reviewed and approved by the AAAM Board. Mr. Lopez-Valdes’ efforts were supported by the Center for Applied Biomechanics at the University of Virginia. We thank Montserrat Ruiz Perez for her assistance in developing the text.

References

- AAAM. American Association for Automotive Medicine (now Association for the Advancement of Automotive Medicine) (1985) The Abbreviated Injury Scale. Des Plaines, IL, USA
- AAAM. Association for the Advancement of Automotive Medicine (2005) The abbreviated injury scale. In T. Gennarelli, E. Wodzin (Eds.), Barrington, IL, USA: AAAM
- AAAM. Association for the Advancement of Automotive Medicine. (2008) The Abbreviated Injury Scale 2005, updated 2008. In T. Gennarelli, E. Wodzin, (Eds.), Barrington, IL, USA
- Al West, T., Rivara, F. P., Cummings, P., Jurkovich, G. J., & Maier, R. V. (2000). Harborview Assessment of Risk of Mortality: An improved measure of injury severity on the basis of ICD-9-CM. *Journal of Trauma*, 49, 530–41.
- Baker, S. P., O'Neill, B., Haddon, W., & Long, W. B. (1974). The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma*, 14, 187–96.
- Berger, L. R., & Mohan, C. D. (1996). *Injury control: A global overview*. India: Oxford University Press.
- Brenneman, F. D., Boulanger, B. R., McLellan, B. A., & Redelmeier, D. A. (1998). Measuring injury severity: time for a change? *Journal of Trauma*, 44(4), 580–2.
- CARE. Community database on Accidents on the Roads in Europe. (2011). http://erso.swov.nl/safetynet/content/wp_1_care_accident_data.htm (Accessed Mar 2011).
- Center for Injury Research and Policy of the Johns Hopkins University School of public Health and Tri-Analytics, Inc. ICDMAP-90: A program to Map ICD-9CM diagnoses into AIS and ISS severity scores. Baltimore, Maryland, 1998
- Champion, H. R., Sacco, W. J., & Copes, W. S. (1989). A revision of the trauma score. *Journal of Trauma*, 26, 623–9.
- Chawda, M.N., Hildebrand, F., Pape, H.C. & Giannoudis, P.V. (2004). Predicting outcome after multiple trauma: Which scoring system? *Injury*, 35(4), 277–58.
- Committee on Medical Aspects of Automotive Safety. (1971). Rating the severity of tissue damage. *Journal of the American Medical Association*, 215(2), 277–80.
- Copes, W. S., Champion, H. R., Sacco, W. J., et al. (1990). Progress in characterizing anatomic injury. *Journal of Trauma*, 30, 1200–7.
- ETSC (2001) EU transport accident, incident and casualty databases: current status and future needs. Brussels: European Transport Safety Council. Available at: www.etsc.eu (Accessed Mar 2011).
- European Center for Injury Prevention. (2006). Algorithm to transform ICD-10 codes into AIS 90. Pamplona, Spain: University of Navarra
- Giannoudis, P. V. (2003). Surgical priorities in damage control in Polytrauma. *Journal of Bone and Joint Surgery*, 85B, 478–83.
- Guralnik, D. B. (Ed.). (1986). *Webster's new world dictionary of the American language, 2nd College Edition*. Upper Saddle River, NJ: Prentice Hall.
- IRTAD (2011). International Traffic Safety Data and Analysis Group. www.internationaltransportforum.org/irtad (Accessed Mar 2011).
- Kingma, J., TenVergert, E., & Klasen, H. J. (1994). SHOWICD: a computer program to display ICD-9CM coded injury diagnoses and their corresponding injury severity scores for a particular patient. *Perceptual Motor and Skills*, 78, 939–46.
- Kingma, J., TenVergert, E., Werkman, H. A., Ten Duis, H. J., & Klasen, H. J. (1994). A Turbo Pascal program to convert ICD-9-CM coded injury diagnoses into injury severity scores: ICDTOAIS. *Perceptual Motor and Skills*, 78, 915–36.
- Lavoie, A., Moore, L., LeSage, N., Liberman, M., & Sampalis, J. S. (2004). The new injury severity score: a more accurate predictor of in-hospital mortality than the injury severity score. *Journal of Trauma*, 56(6), 1312–20.
- MacKenzie, E. J. (1984). Injury severity scales: overview and directions for future research. *American Journal of Emergency Medicine*, 2(6), 537–49.
- MacKenzie, E. J., Steinwachs, D. M., & Shankar, B. (1989). Classifying trauma severity based on hospital discharge diagnoses. Validation of an ICD-9-CM to AIS-85 conversion table. *Medical Care*, 27, 412–22.
- Martin, P. G., & Eppinger, R. H. (2003). Ranking of NASS injury codes by survivability. *Association for the Advancement of Automotive Medicine Annual Proceedings*, 47, 285–300.
- Miller, T. R., Pindus, N. M., Douglass, J. G., et al. (1995). *Databook on nonfatal injury: incidence, costs and consequences*. Washington, DC: The Urban Institute Press.
- NASS CDS. National Highway Traffic Safety Administration. National Automotive Sampling System Crashworthiness Data System. (<http://www.nrd.nhtsa.dot.gov/department/nrd-30/nca/CDS.html>).
- NASS CDS. National Automotive Sampling System Crashworthiness Data System (2009) Coding and editing manual. National highway traffic safety administration. USA: Department of Transport

- O'Keefe, G., Jurkovich, G. J., (2001). Measurement of injury severity and co-morbidity. In Rivara et al. (Eds.), *Injury control: a guide to research and program evaluation*. Cambridge: Cambridge University Press
- Osler, T., Baker, S. P., & Long, W. (1997). A modification of the injury severity score that both improves accuracy and simplifies scoring. *Journal of Trauma*, *43*, 922–5.
- Osler, T., Rutledge, R., Deis, J., & Bedrick, E. (1996). ICISS. An international classification of disease-9 based injury severity score. *Journal of Trauma*, *41*, 380–8.
- Pearson, R. G. (1962). Determinants of injury severity in light plane crashes. *Aviation, Space, and Environmental Medicine*, *33*, 1407–14.
- Petruccelli, E., States, J. D., & Hames, L. N. (1981). The abbreviated injury scale: evolution, usage, and future adaptability. *Accident Analysis and Prevention*, *13*, 29–35.
- Sacco, W. J., MacKenzie, E. J., Champion, H. R., Davis, E. G., & Buckman, R. F. (1999). Comparison of alternative methods for assessing injury severity based on anatomic descriptors. *Journal of Trauma*, *47*(3), 441–6. discussion 446–7.
- Segui-Gomez, M., (2007). [Injury Frequency and Severity Measures] In C Arregui et al. (Eds.), [Principles on Motor Vehicle Injury Biomechanics] DGT.
- Siegel, A. W. (1972). Automobile collisions, kinematics and related injury patterns. *California Medicine*, *116*, 16–22.
- Stevenson, M., Segui-Gomez, M., DiScala, C., Lescohier, J., & McDonald-Smith, G. (2001). An overview of the injury severity score and the new injury severity score. *Injury Prevention*, *7*, 10–3.
- Stevenson, M., Seguí-Gómez, M., DiScala, C., Lescohier, J., & McDonald-Smith, G. (2001). An overview of the injury severity score and the new injury severity score. *Injury Prevention*, *7*, 10–3.
- Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness – a practical scale. *Lancet*, *7*, 81–3.
- Watchko, A. Y., Abajas-Bustillo, R., Segui-Gomez, M., Sochor, M. R., (2011). Current uses of the abbreviated injury scale: a literature review. (Under Review)

Chapter 15

Triage

Craig Newgard

Brief History and Introduction to Field Trauma Triage

The basis for trauma triage is rooted in military medicine and the need to use limited resources in a manner that allows for the greatest benefit (Iserson and Moskop 2007; Moskop and Iserson 2007). Civilian triage has many similarities to military settings, but also unique differences requiring development of triage guidelines specific to a civilian population. In the early 1970s, before the development of trauma centers and trauma systems, injured patients were simply taken to the closest hospital for care. In 1976, the American College of Surgeons Committee on Trauma (ACSCOT) initiated two processes that would prove pivotal in the development of trauma systems and field trauma triage: the earliest version of a trauma triage protocol (including the concept of bypassing a closer hospital for a trauma center) and accreditation of trauma centers (American College of Surgeons 1976; Mackersie 2006). With the concentration of specialized resources, personnel, and expertise at trauma centers, there was a growing need for early identification of seriously injured patients that could be directed to such specialized centers (i.e., triage). Because the majority of seriously injured patients access trauma care through the 9-1-1 emergency medical services (EMS) system, development of formal field trauma triage guidelines was an obvious element in the development of regionalized trauma care.

The Field Triage Decision Scheme represents a combination of science and expert opinion, built largely by consensus of trauma experts and interpretation of research on individual criteria or portions of the triage algorithm. After development of the initial “Triage Decision Scheme” in 1976, the algorithm was revised and reformatted to the “Field Triage Decision Scheme” in 1987 to represent a template very similar to what is used today in most US trauma systems (American College of Surgeons 1986, 1987). The 1987 triage algorithm was the first template to integrate an ordered progression of three “steps” (physiologic, anatomic, and mechanism), organized by likelihood of serious injury. The triage algorithm was revised in 1990 with integration of a fourth step for age and comorbidity factors (Am Coll Surg 1990), again in 1993 (Am Coll Surg 1993), in 1999 (Am Coll Surg 1999), and most recently in 2006 (Fig. 15.1) (Am Coll Surg 2006). The 2006 revision was developed with support from the Centers for Disease Control and Prevention and includes a detailed assessment of both the evidence for and knowledge gaps related to the triage algorithm (CDC 2009). A more recent revision, completed in 2011, is pending release at the time of publication of this text.

C. Newgard, MD, MPH (✉)

Department of Emergency Medicine, Center for Policy and Research in Emergency Medicine,
Oregon Health and Science University, Portland, OR, USA
e-mail: newgardc@ohsu.edu

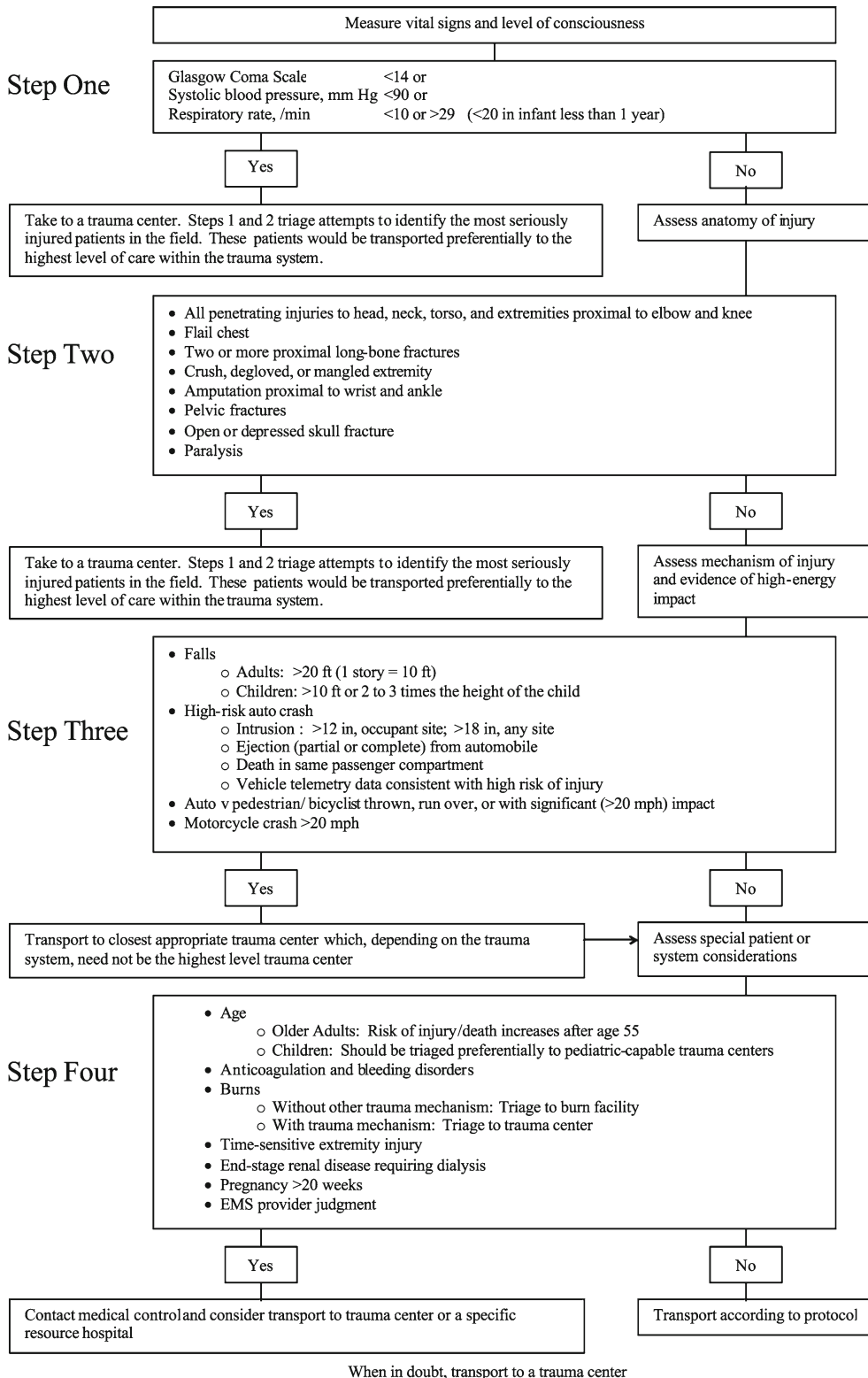


Fig. 15.1 The 2006 National Trauma Triage Protocol. Reprinted with permission from the American College of Surgeons (Am Coll Surg 2006)

The Field Triage Decision Scheme is assumed to be highly sensitive in identifying seriously injured persons (Lerner 2006). However, key limitations in our understanding of field triage (including the true accuracy of the scheme) persist. The decision scheme is organized as an algorithmic decision process, proceeding through four “steps” to identify seriously injured patients. While it is assumed that EMS providers follow the algorithm, inconsistencies in the application of triage criteria have been noted (Pointer et al. 2001; Ma et al. 1999; Baez et al. 2003), and the true process for identifying seriously injured patients in the frequently chaotic out-of-hospital environment remains incompletely understood. The algorithm suggests a formal, methodical process for identifying seriously injured persons, though the realities of field triage are much more complicated. The need to apply the trauma triage guidelines to a heterogeneous patient population in a variety of clinical, environmental, and situational settings, where the occurrence of occult injury is common (particularly early after injury), all contribute to an inherently imperfect and challenging task of identifying those most in need of trauma center care.

This chapter provides a critical evaluation of the existing literature on trauma triage, including the reasons for triage, the ideal patient population targeted for identification in triage, primary and secondary triage, components of the current triage algorithm, under- and over-triage, available accuracy estimates of the decision scheme, important limitations and knowledge gaps, populations with unique triage issues, cost implications, out-of-hospital cognitive reasoning, and future directions.

The Impetus for Trauma Triage: Improved Outcomes and Finite Resources

The Benefits of Trauma Center Care

Trauma systems and the development of trauma centers hinge on the belief that regionalized trauma care (i.e., the concentration of specialized personnel, resources, and expertise in specific hospitals) improves the outcomes of seriously injured persons and provides the most efficient use of limited resources. This belief has been well-substantiated among adults treated in urban/suburban settings (MacKenzie et al. 2006; Mullins et al. 1994, 1996, 1998; Mullins and Mann 1999; Sampalis et al. 1999; Demetriades et al. 2005; Pracht et al. 2007; Jurkovich and Mock 1999; Shafi et al. 2006; Nathens et al. 2000) and to a lesser extent among children (Cooper et al. 1993; Hulka et al. 1997; Johnson and Krishnamurthy 1996; Hall et al. 1996; Pracht et al. 2008). Among seriously injured adults, the survival benefit of early trauma center care has been shown to persist up to 1-year post-injury (MacKenzie et al. 2006). Therefore, field triage seeks to maximize the concentration of such patients in trauma centers, while not overwhelming precious resources. Definitions of “serious injury” (i.e., those shown to benefit from regionalized trauma care) have included Abbreviated Injury Scale (AIS) score ≥ 3 (MacKenzie et al. 2006), Injury Severity Score (ISS) ≥ 16 (Mullins et al. 1994, 1996; Mullins and Mann 1999; Jurkovich and Mock 1999; Hulka et al. 1997), ISS >12 or ≥ 2 injuries with AIS ≥ 2 (Sampalis et al. 1999), specific “index” injuries (Mullins et al. 1998; Demetriades et al. 2005), and certain International Classification of Disease-9 (ICD-9) diagnoses (Pracht et al. 2007, 2008). The specifics of these definitions are important in matching the target of triage to the type of patient shown to derive a measurable outcome benefit from specialized trauma care.

Finite Trauma Resources

Although trauma center care has been demonstrated to improve survival among the seriously injured, the resources that allow for such outcomes are finite. Trauma centers and trauma systems face continued threats to maintaining key resources, including hospital and emergency department closures

Table 15.1 Definitions used to denote trauma center “need” in previous triage studies^a

Adults	
ISS ≥ 16	(Knopp et al. 1988; Esposito et al. 1995; Norcross et al. 1995; Long et al. 1986; Bond et al. 1997; West et al. 1986; Cooper et al. 1995; Smith and Bartholomew 1990; Cottington et al. 1988)
ISS ≥ 20	(Cottington et al. 1988)
ISS ≥ 10 plus LOS	(West et al. 1986)
ISS plus major non-orthopedic surgery, ICU, death, and other resources	(Norcross et al. 1995; Simon et al. 1994; Newgard et al. 2005, 2007a, b)
Emergency operative intervention within 1 h of emergency department arrival	(Steele et al. 2007)
Major non-orthopedic surgery or death	(Henry et al. 1996)
Major non-orthopedic surgery, ICU, death, and other resources	(Gray et al. 1997; Phillips and Buchman 1993; Baxt et al. 1990; Fries et al. 1994; Zechnich et al. 1995; Newgard et al. 2005)
Death or LOS	(Newgard et al. 2010a, b)
Children	
ISS ≥ 16	(Tepas et al. 1988; Eichelberger et al. 1989; Chan et al. 1989; Kaufmann et al. 1990; Phillips et al. 1996; Qazi et al. 1998; Newgard et al. 2002)
ISS ≥ 16, plus major non-orthopedic surgery, ICU, death, and other resources	(Newgard et al. 2005, 2007a, b)
Major non-orthopedic surgery, ICU, death, and other resources	(Engum et al. 2000; Qazi et al. 1998)
Death or LOS	(Newgard et al. 2010a, b)

^aSome studies assessed multiple outcomes and are therefore listed more than once. *ISS* injury severity score, *ICU* intensive care unit, *LOS* length of stay

(2006a, b), difficulty maintaining on-call panels (McConnell et al. 2007, 2008), increasing economic threats and competition for state and federal budgets (Mann et al. 2005), and a declining workforce of trauma surgeons (Green 2009). Sending all injured patients directly to trauma centers would overwhelm the capacity to provide such specialized care and would result in very inefficient use of resources. Some research also suggests that emergency department resources required for trauma patients pull critical staff and resources away from other high-acuity patients (e.g., acute cardiac patients) that can result in worse outcomes for such non-trauma patients (Fishman et al. 2006). Trauma centers also tend to serve as specialty centers for other conditions (e.g., ST-elevation myocardial infarction, stroke, cardiac arrest, oncology, transplant, etc.), frequently have high clinical volumes, and spend more time on ambulance diversion than non-trauma hospitals (Sun et al. 2006). Triage is an important aspect in preserving trauma resources for those most in need and those shown to benefit from comprehensive trauma care.

Defining the Target of Triage

A logical next question is which patients should be targeted in the development and assessment of field trauma triage guidelines? An evidence-based approach to trauma triage would seek to identify patients that have been shown to benefit from care at major trauma centers (Section “The Benefits of Trauma Center Care”). However, previous triage research has used a vast array of definitions to denote the target population of trauma triage, including different measures of injury severity, length of stay (LOS), resource use, and death (Table 15.1). While a resource-based definition of trauma

center “need” is a practical method for defining the target population for triage, such definitions are subject to many potential biases and variability in practice patterns. For example, a given procedure (e.g., splenectomy) performed in one hospital may not be performed on a similar patient in another hospital (Todd et al. 2004) or even by another surgeon at the same hospital, so using this operative intervention in a composite triage outcome is confounded by such variations in hospital and provider practice patterns. The integration of time-based resource definitions [e.g., major operative intervention within 1 h (Steele et al. 2007)] is also potentially confounded by variability in surgical decision-making, clinical practice patterns, operative resource constraints (e.g., operating room availability), and unique issues with certain patients (e.g., obtaining parental consent for a minor or contacting a person with medical decision-making capacity for an elder with dementia). To complicate matters further, previous studies have demonstrated only fair correlation between resource use and anatomic injury measures (Baxt and Upenieks 1990; Newgard et al. 2008), suggesting that using only anatomic injury or only resource measures to define the object of triage may miss important patients.

Research and development of trauma triage decision rules should seek to match the target of triage criteria with the type of patient shown to benefit from trauma center care, yet there is disagreement on the exact definition. Comparison of the various definitions for patients shown to benefit from trauma center care suggests a common denominator of having at least one “serious” injury (AIS \geq 3). However, such a definition could be considered too liberal, as patients with more severe injuries (e.g., AIS \geq 4, ISS \geq 16) are more closely tied to mortality risk and appear to drive the primary outcome benefit and cost-effectiveness of trauma center care (MacKenzie et al. 2006, 2010). The issue of including a resource-based definition (possibly in addition to a measure of anatomic injury severity) remains unresolved, because a definition based purely on injury severity may miss important patients requiring trauma center care or may not match the geographic distribution of trauma resources (Newgard et al. 2008). Further, the ideal target for trauma triage practices may differ by region, depending on resource availability and trauma system design. Defining the “major trauma patient” (the patient requiring immediate transport to a trauma center) remains an active area of triage research.

Primary and Secondary Triage

Primary Trauma Triage

There are two general types of trauma triage: primary and secondary. While these terms are used in different ways, a practical definition of each is offered here. *Primary triage* is generally performed in the out-of-hospital environment (i.e., by EMS providers), prior to any emergency department or hospital-based evaluation, and involves the process of actively matching the receiving hospital to the patient’s medical and surgical needs based on presumed injury severity and/or resource need. Sometimes this distinction is termed a field “trauma activation” or “trauma entry” to delineate active enrollment into a trauma system with subsequent protocolized care for the patient. Primary triage is the basis for the Field Triage Decision Scheme. Simply assessing the type of hospital to which a patient was transported does not necessarily constitute an accurate reflection of primary triage, as injured patients may be transported to major trauma centers for a variety of reasons unrelated to triage (e.g., proximity, patient request). This distinction is important when calculating accuracy estimates for primary triage and field triage guidelines, as using receiving hospital to define triage processes may over-estimate the true accuracy of primary triage (e.g., a patient with unrecognized serious injury who happened to be transported to a trauma center due to proximity or patient request).

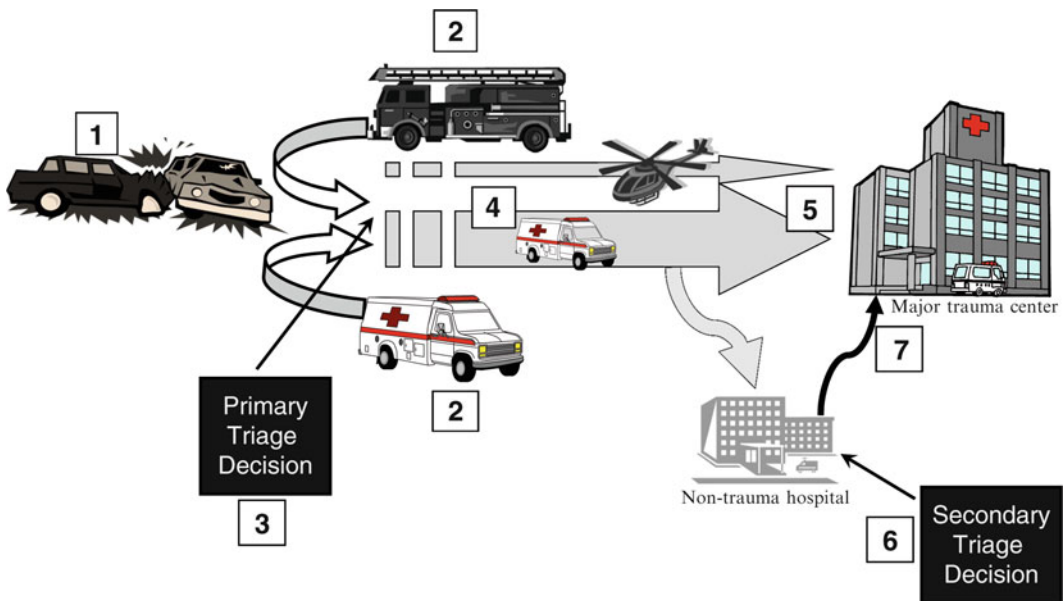


Fig. 15.2 Primary and secondary triage processes in a sample trauma system

Secondary Trauma Triage

Secondary triage generally represents emergency department- or hospital-based triage. Secondary triage may not only occur following primary triage (i.e., after EMS transport) but can also occur without primary triage (e.g., for a patient transported to a hospital by private auto without EMS contact). The intent of secondary triage differs based on the hospital setting where it is performed. Secondary triage at trauma centers is often done to guide the need for immediate trauma resources and staff present upon patient arrival. Many trauma centers will use an initial graded response to determine which members of the “trauma team” are involved in the initial emergency department assessment and care. Alternatively, secondary triage in non-trauma hospitals has the goal of identifying seriously injured patients for inter-hospital transfer to a major trauma center. Depending on the trauma system and region (e.g., rural and frontier settings), there may be protocols for EMS to initially transport patients to the closest hospital for stabilization (with subsequent inter-hospital transfer to a tertiary trauma center as needed), even if the patient meets primary field triage guidelines. Particularly at non-trauma hospitals (or lower level trauma centers), secondary triage is crucial in the early identification of seriously injured patients either missed by primary triage processes, presenting without primary triage (e.g., transport by private auto) or otherwise requiring higher level of care services. While some trauma systems have criteria to guide secondary triage processes at non-trauma facilities, there is relatively little research investigating secondary triage practices. Existing research suggests that there is substantial variability in secondary triage practices among non-tertiary hospitals, (Newgard et al. 2006) yet that there is also a measureable outcome benefit from secondary triage (Newgard et al. 2007). This is an area ripe for additional research, as effective secondary triage practices can help improve the concentration of seriously injured patients in major trauma centers and increase the efficiency of a trauma system.

The schematic in Fig. 15.2 depicts primary and secondary triage processes in a sample trauma system. There are many variations on this theme, though the majority of the elements depicted here are represented in some way in most US trauma systems. The process begins with an injury event

(1), followed by 9-1-1 notification and an EMS response (2). The type of EMS response (e.g., advanced life support vs. basic life support) differs by region, including the number of vehicles that initially respond. Figure 15.2 illustrates a dual-response EMS system, with a first responder (e.g., fire department) and transport vehicle (ambulance) both responding to the 9-1-1 call. After initial assessment of the scene and the patient, a primary triage decision is made (3). For patients who are transported, there is a decision about selecting a receiving facility (i.e., based on the primary triage assessment and other factors) and mode of transport (i.e., ground ambulance vs. air medical transport) (4). For patients meeting trauma triage criteria, there is often advance notification to the receiving trauma center to allow preparation for patient arrival (5). Patients that do not meet field triage criteria may be transported to a non-trauma hospital or to a trauma center, depending on proximity, patient preference and other factors. At both trauma centers and non-trauma hospitals, there is generally a secondary triage decision (6). For trauma centers, secondary triage may help determine which members of the trauma team are present for the initial emergency department patient assessment and management. For non-trauma hospitals, secondary triage typically involves a decision of whether to transfer the patient to a trauma center for further management based on known or suspected serious injuries (7). The secondary triage decision at non-trauma hospitals may be made early in the course of hospital care (e.g., in the emergency department) or days later following admission to the hospital.

The Field Triage Decision Scheme: Deciphering Components of the Algorithm

The most recent version of the Field Triage Decision Scheme entails four “steps” listed in order of decreasing risk for serious injury: physiologic (step 1), anatomic (step 2), mechanism (step 3), and special considerations (step 4). The decision scheme is generally viewed as a template for systems to follow, but one that can be modified to fit the unique complexities of individual systems. While many of the triage criteria have research to demonstrate their predictive value in identifying seriously injured patients, other factors (e.g., comorbidities) have been added based on expert opinion and/or indirect evidence of risk. Especially for steps 3 and 4 (mechanism and risk factors), each revision of the scheme has included additions and deletions of different criteria. This aspect, along with variable uptake among EMS and trauma systems, has created a situation where “old” criteria frequently remain in use, even after deletion from the revised algorithm, creating variability in the criteria used in practice between different trauma systems.

Step 1: Physiologic Criteria

The physiologic step has remained fairly consistent across multiple revisions of the Field Triage Decision Scheme, except for slight changes in the cut point for Glasgow Coma Scale (GCS) score. Step 1 consists of measures of physiologic compromise, including mentation (GCS), hypotension (systolic blood pressure), and respiratory distress (respiratory rate). Some systems recognize airway compromise as a separate criterion, while others lump airway issues into the respiratory rate criterion. There have been multiple studies demonstrating the high-risk nature and predictive value of physiologic compromise among injured adults and children (Cottingham et al. 1988; Esposito et al. 1995; Henry et al. 1996; Hannan et al. 2005; Baxt et al. 1990; Franklin et al. 2000; Lipsky et al. 2006; Newgard et al. 2007 a, b, 2010a, b; Kaufmann et al. 1990; Engum et al. 2000). Whether there

is a benefit of using pediatric-specific physiologic values to better identify seriously injured children remains unclear (Eichelberger et al. 1989; Nayduch et al. 1991; Phillips et al. 1996; Newgard et al. 2007a, b, 2009).

While the predictive value of physiologic compromise is generally high, such patients constitute a minority of patients with serious injury. That is, there are a substantial number of seriously injured patients with normal/compensated physiology during the initial field evaluation. Physiologic measures have therefore generally been shown to be insensitive, yet highly specific, for identifying seriously injured patients (Cottingham et al. 1988; Esposito et al. 1995; Kane et al. 1985; Norcross et al. 1995; Henry et al. 1996; Knopp et al. 1988; Long et al. 1986; Bond et al. 1997; Baxt et al. 1990; Zechnich et al. 1995; Lerner 2006). Triage algorithms that rely exclusively on physiologic measures to identify those in need of trauma center care are likely to miss a sizable portion of seriously injured patients.

Step 2: Anatomic Criteria

In the anatomic step, specific anatomic injuries diagnosed during field assessment are used to identify patients requiring immediate trauma center care. These criteria include such factors as: penetrating injuries of the head, neck, or torso; flail chest; multiple proximal long-bone fractures; proximal amputation; pelvic fracture; skull fracture; and spinal injury/paralysis. Though also highly predictive of serious injury and resource need (Esposito et al. 1995; Henry et al. 1996; Knopp et al. 1988; Lerner 2006), many of these diagnoses are difficult to make in the field, and a minority of patients actually meet such specific criteria. Similar to the physiologic criteria, anatomic triage criteria are highly specific for serious injury and need for trauma center care, but are generally insensitive. That is, the absence of anatomic criteria does not substantially reduce the likelihood of serious injury.

Steps 3 and 4: Mechanism and Special Considerations Criteria

The mechanism and risk factor steps have generally demonstrated less predictive value and therefore have been considered more controversial as independent triage criteria. Some have suggested that patients meeting only mechanism-of-injury criteria contribute to over-triage rates (Simon et al. 1994; Shatney and Sensaki 1994). However, many patients with serious injury do not manifest physiologic abnormality or anatomic injury during the initial field assessment. As detailed above, some of this phenomenon may be explained by early physiologic compensation following injury and the difficulty in making anatomic injury diagnoses in the field. Therefore, mechanism and risk factor criteria are felt to play important roles in identifying seriously injured patients missed by physiologic and anatomic criteria. There are multiple studies supporting the inclusion of mechanism triage criteria (Cottingham et al. 1988; Esposito et al. 1995; Henry et al. 1996; Knopp et al. 1988; Long et al. 1986; Cooper et al. 1995; Newgard et al. 2005; Burd et al. 2007), though debate continues regarding which of these criteria should be recognized as independent criteria. For special considerations, there is little data to directly support their inclusion, but they have logical utility in identifying high-risk patients who often require specialized care, and therefore have been retained in the triage scheme (CDC 2009).

EMS Provider Judgment was added as an independent criterion to Step 4 in the 2006 version of the Field Triage Decision Scheme (CDC 2009). However, this criterion has been used in many EMS and trauma systems for years and has been indirectly supported by previous versions of the triage algorithm stating “When in doubt, take to a trauma center” (MacKersie 2006). There have been mixed results regarding the utility of EMS provider judgment in identifying patients with serious injury (Qazi et al. 1998; Fries et al. 1994; Simmons et al. 1995; Mulholland et al. 2005). However,

provider judgment likely plays a significant role in interpreting the presence and application of other triage criteria and navigating the many clinical and environmental scenarios not depicted in individual criteria that pose the potential for serious injury and resource need.

The Concepts of Under-Triage, Over-Triage and Overall Accuracy of Trauma Triage

Under-Triage

In the context of primary (field) triage, *under-triage* represents the proportion of seriously injured patients transported from the scene to non-trauma hospitals (Am Coll Surg 2006). The under-triage rate can be calculated directly from a sensitivity value for identifying seriously injured patients ($1 - \text{sensitivity}$). The target for under-triage rates in a trauma system is less than 5% (Am Coll Surg 2006). While seemingly straightforward, the definition for this term becomes less clear when considering inter-hospital transfers and patients cared for in rural locations. In rural settings (or in regions with long transport times to a major trauma center), some trauma systems recommend transport to the closest appropriate hospital for initial evaluation and stabilization even when triage criteria are met, with subsequent inter-hospital transfer to a major trauma center as needed. That is, some systems may consider the definition of “under-triage” based on the ability to identify and concentrate seriously injured patients in major trauma centers within a fixed time period (e.g., the first 24 h) rather than direct transport from the scene. MacKenzie et al. used such a practical definition to define and quantify the benefit of early trauma center care (MacKenzie et al. 2006). Previous primary triage research has demonstrated the under-triage rate of the trauma triage guidelines to generally be low (Lerner 2006); however, these estimates are subject to many methodological limitations. Recent research suggests the under-triage rate may be much higher than previously known for both adults and children (Vassar et al. 2003; Wang et al. 2008; Hsia et al. 2010) and varies significantly by age (Vassar et al. 2003; Hsia et al. 2010). Another aspect of calculating under-triage rates is accounting for unrecognized seriously injured patients who are still transported to major trauma centers (e.g., based on proximity or patient request). While some may not consider such patients under-triaged because they ultimately arrive at the correct type of hospital, they should be considered near-misses (or true misses) because they were not prospectively identified by triage criteria.

Over-Triage

Over-triage generally represents the proportion of patients with minor injuries that are transported to major trauma centers (Am Coll Surg 2006). Patients with minor injuries have not been shown to have a measurable benefit from care at trauma centers and therefore constitute inappropriate use of specialized resources with increased expense (Hoff et al. 1992). The over-triage rate can be calculated directly from the specificity of identifying minimally injured patients ($1 - \text{specificity}$). Per ACSCOT, the target over-triage rate in a trauma system should be less than 50% (Am Coll Surg 2006). Previous triage research suggests the over-triage rate for field trauma triage to be in this range or higher (Lerner 2006), yet these estimates are subject to the same limitations noted in calculating rates of under-triage. Because the number of patients with minor injuries is substantially greater than those with serious injuries (Newgard et al. 2012), moderate to high over-triage rates translate into vastly larger volumes of persons cared for at major trauma centers, increased costs, and magnification of system inefficiencies.

The Balance Between Under- and Over-Triage

There is an inevitable trade-off between under- and over-triage. In general, as one goes down, the other goes up. To date, the culture of out-of-hospital triage has favored the minimization of under-triage at the expense of over-triage, thus maximizing the capture of seriously injured patients. However, the consequences of such overuse of resources and expense remains poorly defined. Major trauma centers have been shown to have high rates of ambulance diversion (Sun et al. 2006), frequently function at or above capacity, and have a questionable ability to handle a significant surge in clinical care (e.g., during a major disaster) (Rivara et al. 2006). As trauma centers also frequently serve as specialized care centers for other medical conditions, the ability to care for patients with such non-trauma conditions may also be affected by liberal over-triage rates. Finally, while there are guidelines for “acceptable” under- and over-triage rates, these targets may not be appropriate in all settings, depending on available resources, funding, patient volume, geographic location, and other factors.

The Accuracy of Field Trauma Triage

Although there is a relatively large body of literature assessing individual triage criteria and segments of the triage algorithm, few studies have evaluated the decision scheme in its entirety. Henry et al. evaluated the full Field Triage Decision Scheme among patients involved in motor vehicle crashes (Henry et al. 1996). Two additional studies evaluated the full triage algorithm (Esposito et al. 1995; Norcross et al. 1995). These and other studies suggest the sensitivity and specificity of the decision scheme to range from 57–97% and 8–55%, respectively (Lerner 2006). However, as of the writing of this chapter, there have been no rigorous, prospective studies validating the full decision scheme among a broad out-of-hospital injury population, though such efforts are currently underway. This limitation has been noted in the most recent revision of the triage guidelines (CDC 2009) and by other groups dedicated to critical evaluation of the existing triage literature (EAST 2010). Some research suggests that the under-triage rate may be much higher (Vassar et al. 2003; Wang et al. 2008). Prospective validation of the Field Triage Decision Scheme is needed to guide future revisions of the guidelines and enhance the efficiency of regionalized trauma care.

Important Limitations in Previous Trauma Triage Research

While the body of literature for trauma triage is relatively large, there are key limitations that have persisted in almost all previous trauma triage studies.

Study Design

The majority of previous trauma triage studies have used retrospective study designs and data from trauma registries. While retrospective studies are integral to describing relevant issues, testing associations, and formulating hypotheses to be further evaluated in prospective studies, research on field triage has generally not moved into a rigorous prospective phase of evaluation. The retrospective nature of previous triage research has created the potential for selection bias, variable definitions of key predictor terms and triage criteria, variable inclusion criteria, and other potential threats to the validity of study results. While prospective trauma triage research has been conducted (Esposito

et al. 1995; Norcross et al. 1995; Henry et al. 1996; Knopp et al. 1988; Baxt et al. 1990; Phillips and Buchman 1993; Cooper et al. 1995), most of these studies have other key limitations.

Defining the Relevant Out-of-Hospital Injury Population (i.e., the Denominator)

Another substantial limitation of previous trauma triage research has been defining and studying the appropriate out-of-hospital population. Most previous studies have used non-population-based sampling and integration of hospital-based inclusion criteria (e.g., only admissions, ISS above a certain threshold value, only trauma center patients, restriction to patients entered into a trauma registry). Such variable inclusion criteria limit the generalizability of findings and create the potential for selection bias. Limiting field data collection to single EMS agencies or single modes of transport (e.g., air medical) can also detract from population-based sampling and integrate bias to the calculation of accuracy measures. Because most previous triage studies have focused on patients transported to trauma centers, the population of patients initially transported to non-trauma hospitals has remained essentially invisible, except for those subsequently transferred to major trauma centers. Such scenarios suggest a strong potential for inflated sensitivity estimates of the trauma triage guidelines. In summary, most previous triage studies have used a narrower denominator of patients than those for whom the decision scheme is routinely applied (i.e., all injured patients evaluated by out-of-hospital personnel).

Data Quality and Definitions of Field Triage Criteria

The out-of-hospital setting is complex, often with multiple EMS agencies and providers caring for the same patient. This scenario is common in tiered EMS systems and dual-response EMS systems. Failure to adequately capture data from all EMS agencies participating in the care of a given patient may unintentionally omit important clinical and triage information. Further, defining and recording field triage criteria should ideally be done prospectively by field providers to avoid skewed definitions or use of information that was not available (or not appreciated) at the scene. Missing data are also common in EMS charts, creating the need for attention to appropriately handling missing values. Failure to account for missing data can introduce bias into the results and reduce study power (Little and Rubin 2002; Van Der Heijden et al. 2006; Crawford et al. 1995; Newgard and Haukoos 2007). These complexities have not been accounted for in most previous triage research. Failure to appreciate and account for such subtleties can result in potentially inaccurate, misclassified, or biased data on field triage.

Variability in the Target of Triage

As detailed in Section “Defining the Target of Triage,” there have been a multitude of definitions used for patients targeted by triage criteria. This variability has reduced comparability between studies and allowed questions to persist. Many studies have used definitions inconsistent with literature defining patients shown to benefit from trauma center care. Because previous research has demonstrated an outcome benefit of trauma center care for patients with injuries of AIS ≥ 3 severity (MacKenzie et al. 2006), setting the definition for “serious injury” above this level (e.g., ISS ≥ 20) results in misclassification of some patients that may otherwise serve to benefit from trauma center care. As previously detailed, there are also challenges to using resource-based definitions.

These findings strongly suggest the need to define the target population of triage using measures with face validity, demonstrated association of benefit from trauma center care, and free from practice variability.

Lack of Full Clinical Decision Rule Methodology

Although the Field Triage Decision Scheme has been developed and widely implemented over the past 2 decades as a clinical decision rule, several aspects of decision rule development have yet to be conducted. These include assessing the inter-rater reliability of field triage criteria, appropriate selection of subjects (the out-of-hospital injury denominator), matching the sample size to planned analyses (i.e., power calculations), prospective validation, understanding how the decision rule is used in practice and economic impact of the rule (e.g., whether the rule is cost-effective) (Stiell and Wells 1999; Laupacis et al. 1997). These are areas of need for future trauma triage research.

Timing of Triage

The concept of the “golden hour” has been deeply entrenched in the development of trauma systems, trauma triage, and EMS systems, yet evidence demonstrating a clear link between time and outcome among injured patients is sparse (Lerner and Moscati 2001). There is likely a subset of injured patients where minutes (or hours) do affect survival; however, this association has not been substantiated in most research to date. Two studies from Quebec in the 1990s demonstrated an association between shorter out-of-hospital times and increased survival (Sampalis et al. 1993, 1999), yet more recent studies have failed to replicate such a link, even among injured patients meeting Step 1 physiologic criteria (Newgard et al. 2010a,b). Several studies have also compared trauma patients transported directly to major trauma centers versus those first evaluated in non-trauma hospitals with subsequent transfer to trauma centers (Nirula et al. 2010; Nathens et al. 2003; Sampalis et al. 1997; Young et al. 1998). Most (though not all) of these studies suggest that patients transported direct from the scene to major trauma centers have better outcomes, though it is unclear whether selection bias and unmeasured confounding may explain these findings. In the context of trauma triage, the issue of timing is important because some research suggests that seriously injured patients missed by primary triage processes or transported to a lower level trauma center for initial evaluation and stabilization may still have a window of time for secondary triage with improved outcomes (Newgard et al. 2007a,b). While the details of such a time window remain poorly understood, inclusive trauma systems with efficient primary and secondary triage processes are likely to maximize outcomes and efficient resource use by effectively matching patient need to varying levels of care (Utter et al. 2006).

Populations with Unique Triage Issues

It is unlikely that a single triage algorithm will be accurate and effective in all settings. That is, expecting a “one-size-fits-all” approach to trauma triage is unrealistic. In this section, three populations with unique triage issues (children, elders, patients injured in rural locations) that are likely to affect the performance and accuracy of triage guidelines are briefly discussed.

Children

Children are a unique and under-researched population with regard to trauma triage. There are many issues unique to injured children, including physiologic response to injury, injury patterns and mechanisms, differences in clinical and operative management, pediatric versus adult trauma centers, and the need for practitioners experienced in the care of acutely ill children. A 2006 Institute of Medicine report on the state of emergency care highlighted these issues, along with the many deficiencies in pediatric emergency care in the US healthcare system (2006a, b). Some trauma systems have integrated child-specific triage guidelines (e.g., age-specific systolic blood pressure (SBP) and respiratory rates), yet the evidence for and utility of such modifications remain unclear. Some studies have demonstrated age-specific associations between physiologic measures and outcomes (e.g., SBP, respiratory rate, heart rate) (Newgard et al. 2007a, b; Potoka et al. 2001; Coates et al. 2005), while others have shown no difference (Eichelberger et al. 1989; Kaufmann et al. 1990; Nayduch et al. 1991; Newgard et al. 2009). A recent population-based assessment of field physiologic measures among 10 sites across North America was unable to demonstrate utility in age-specific physiologic measures in identifying high-risk injured children (Newgard et al. 2009). The same study also found a significant proportion of missing out-of-hospital values (e.g., SBP) among injured children that differed by age and outcome. Gausche et al. previously demonstrated that out-of-hospital providers are uncomfortable measuring vital signs (especially blood pressure) in young children and frequently simply forego such efforts (Gausche et al. 1990), further calling into question the use of age-specific pediatric physiologic values for triage. While the current framework for pediatric trauma triage is generally no different than for adults (Am Coll Surg 2006), the question remains whether a completely different algorithm would better identify seriously injured children and better meet the practical realities of caring for injured children in the out-of-hospital environment.

Elders

As with children, injured elders also have unique triage considerations that are not reflected in the current triage guidelines. Although existing research is limited, some research has suggested that the current triage guidelines are relatively insensitive for identifying seriously injured elders (Scheetz 2003) and that many injured elders are cared for in non-trauma hospitals (Hsia et al. 2010). Other perplexing issues include the questions of whether injured elders benefit from care in major trauma centers (MacKenzie et al. 2006) and the cost-effectiveness of caring for seriously injured older patients in major trauma centers (MacKenzie et al. 2010). Injured elders frequently have unique issues not present in younger patients (e.g., different physiologic response to injury, increased comorbidity burden, more complex considerations with operative intervention and medical management, end-of-life considerations, different preferences regarding the location of care, etc.). Whether elder-specific triage criteria should be developed remains unclear and is another important area of future trauma research in the setting of an aging population.

Rural Patients

A large number of Americans live more than 60 min from the closest major trauma center and 28% of the US population is able to access specialized trauma care within 60 minutes only by helicopter (Branas et al. 2005). Previous research has demonstrated that persons injured in rural locations tend

to have worse outcomes (Gomez et al. 2009), possibly secondary to long EMS response times, decreased access to high-quality trauma care, and other factors. Other research has shown that while survival improved in urban areas during implementation of a statewide trauma system, there was no measurable change in rural regions (Mann et al. 2001). Another study demonstrated that mortality for patients injured in rural locations worsened after removal of air medical transport (Mann et al. 2002), suggesting that air transport services are particularly important in rural settings. Additional research has shown variability in inter-hospital transfer practices among rural hospitals (Newgard et al. 2006). These findings all suggest that primary and secondary triage issues are different in rural regions and likely play a role in determining outcomes among persons injured in rural settings. Unfortunately, research to better understand and guide triage protocols in such settings is sparse. Triage guidelines developed exclusively in urban/suburban locations with relatively close proximity to major trauma centers may not apply to rural settings. Rural trauma triage is an area of great need for future triage and trauma research.

Cost Implications of Field Triage

While the cost-effectiveness of trauma center care has been demonstrated among seriously injured adults (MacKenzie et al. 2010), there is little research on the cost implications of trauma triage. The cost of care is notably higher in trauma centers, even after accounting for injury severity and other important confounders (Goldfarb et al. 1996; MacKenzie et al. 2010). Although these costs are justifiable among seriously injured patients, it is quite possible that trauma systems with high over-triage rates are not cost-effective. Because field triage has substantial downstream effects on care (e.g., location of care, type of care received, inter-hospital transfers, etc.), there are likely to be substantive cost implications stemming from prehospital triage decisions. Future research is needed to better define these costs and financial implications in concert with patient outcomes to maximize the efficiency of trauma systems.

Field Provider Cognitive Reasoning in Trauma Triage

The current model for field trauma triage is algorithmic (Am Coll Surg 2006). Since its inception, there has been an assumption that field providers will simply follow the algorithm to make triage decisions. While this may be true for new field providers, a recent study suggests that field providers may use cognitive reasoning processes closer to that of experienced clinicians to make triage decisions, rather than following a highly structured, algorithmic approach (Newgard et al. 2011). Such rapid cognitive processing, termed “Type 1” by Croskerry (Croskerry 2009), is fast, heuristic, and intuitive – all factors encouraged and rewarded in EMS systems favoring short scene times and rapid transport for trauma patients. This rapid decision-making is partially captured under the criterion “EMS Provider Judgment” in the 2006 Field Triage Decision Scheme and is likely to be closely tied to provider experience. Better understanding the cognitive reasoning processes used by out-of-hospital providers during field triage may help explain the variable application of triage criteria. The influence, role, and predictive value of “EMS Provider Judgment” as an individual criterion requires additional research and may offer insight into the practice of trauma triage in the dynamic and often chaotic out-of-hospital setting.

Future Directions with Trauma Triage

Primary and secondary trauma triage practices play critical roles in trauma systems. Current processes used for trauma triage in the USA have been developed over the past 3 decades, but have important limitations and many areas for further research and development. As regionalized care becomes increasingly integrated to healthcare delivery systems for a variety of high-acuity conditions (i.e., ST-elevation myocardial infarction, stroke, cardiac arrest), trauma triage processes and trauma systems will continue to serve as models for such care. To achieve the Institute of Medicine's vision of a fully integrated emergency care system, primary and secondary triage processes (for trauma and other conditions) will need continued development and evaluation. Future directions in trauma triage involve defining the "major trauma" patient (i.e., those most in need of immediate transport to major trauma centers), defining the role of time (when, where, and how trauma care should be delivered), improved matching of patient need to varying levels of care, geographic and age-specific differences in triage, addressing limitations in previous trauma triage research, understanding and applying cognitive reasoning models to triage guidelines, and maximizing triage in an increasingly cost-constrained healthcare environment.

References

- American College of Surgeons (1976). Optimal hospital resources for care of the seriously injured. *Bulletin of the American College of Surgeons*, 61:15–22.
- American College of Surgeons (1986). *Hospital and prehospital resources for the optimal care of the injured patient*. Chicago, IL: American College of Surgeons.
- American College of Surgeons (1987). *Hospital and prehospital resources for the optimal care of the injured patient, Appendices A through J*. Chicago, IL: American College of Surgeons.
- American College of Surgeons (1990). *Resources for the optimal care of the injured patient*. Chicago, IL: American College of Surgeons.
- American College of Surgeons (2006). *Resources for the optimal care of the injured patient*. Chicago, IL: American College of Surgeons.
- Baez, A. A., Lane, P. L., & Sorondo, B. (2003). System compliance with out-of-hospital trauma triage criteria. *Journal of Trauma*, 54, 344–351.
- Baxt, W. G., Jones, G., & Fortlage, D. (1990). The Trauma Triage Rule: A new, resource-based approach to the out-of-hospital identification of major trauma victims. *Annals of Emergency Medicine*, 19, 1401–1406.
- Baxt, W. G., & Upenieks, V. (1990). The lack of full correlation between the injury severity score and the resource needs of injured patients. *Annals of Emergency Medicine*, 19, 1396–1400.
- Bond, R. J., Kortbeek, J. B., & Preshaw, R. M. (1997). Field trauma triage: Combining mechanism of injury with the out-of-hospital index for an improved trauma triage tool. *Journal of Trauma*, 43, 283–287.
- Branas, C. C., MacKenzie, E. J., Williams, J. C., et al. (2005). Access to trauma centers in the United States. *Journal of the American Medical Association*, 293, 2626–2633.
- Burd, R. S., Jan, T. S., & Nair, S. S. (2007). Evaluation of the relationship between mechanism of injury and outcome in pediatric trauma. *Journal of Trauma*, 62, 1004–1014.
- Centers for Disease Control and Prevention. (2009). Guidelines for field triage of injured patients: Recommendations of the national expert panel on field triage. *Morbidity and Mortality Weekly Report*, 57, 1–35.
- Chan, B. S. H., Walker, P. J., & Cass, D. T. (1989). Urban trauma: An analysis of 1,116 paediatric cases. *Journal of Trauma*, 29, 1540–1547.
- Coates, B. M., Vavilala, M. S., Mack, C. D., et al. (2005). Influence of definition and location of hypotension on outcome following severe pediatric traumatic brain injury. *Critical Care Medicine*, 33, 2645–2650.
- Cooper, A., Barlow, B., DiScala, C., et al. (1993). Efficacy of pediatric trauma care: Results of a population-based study. *Journal of Pediatric Surgery*, 28, 299–303.
- Cooper, M. E., Yarbrough, D. R., Zone-Smith, L., et al. (1995). Application of field triage guidelines by prehospital personnel: Is mechanism of injury a valid guideline for patient triage? *American Surgeon*, 61, 363–367.

- Cottingham, E. M., Young, J. C., Shuffelbarger, C. M., et al. (1988). The utility of physiologic status, injury site, and injury mechanism in identifying patients with major trauma. *Journal of Trauma*, 28, 305–311.
- Crawford, S. L., Tennstedt, S. L., & McKinlay, J. B. (1995). A comparison of analytic methods for non-random missingness of outcome data. *Journal of Clinical Epidemiology*, 48, 209–219.
- Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, 84, 1022–1028.
- Demetriades, D., Martin, M., Salim, A., et al. (2005). The effect of trauma center designation and trauma volume on outcome in specific severe injuries. *Annals of Surgery*, 242, 512–519.
- Eichelberger, M. R., Gotschall, C. S., Sacco, W. J., et al. (1989). A comparison of the trauma score, the revised trauma score, and the pediatric trauma score. *Annals of Emergency Medicine*, 18, 1053–1058.
- Engum, S. A., Mitchell, M. K., Scherer, L. R., et al. (2000). Prehospital triage in the injured pediatric patient. *Journal of Pediatric Surgery*, 35, 82–87.
- Espósito, T. J., Offner, P. J., Jurkovich, G. J., et al. (1995). Do out of hospital trauma center triage criteria identify major trauma victims? *Archives of Surgery*, 130, 171–176.
- Fishman, P. E., Shofer, F. S., Robey, J. L., et al. (2006). The impact of trauma activations on the care of emergency department patients with potential acute coronary syndromes. *Annals of Emergency Medicine*, 48, 347–353.
- Franklin, G. A., Boaz, P. W., Spain, D. A., et al. (2000). Prehospital hypotension as a valid indicator of trauma team activation. *Journal of Trauma*, 48, 1034–9.
- Fries, G. R., McCalla, G., Levitt, M. A., et al. (1994). A prospective comparison of paramedic judgment and the trauma triage rule in the prehospital setting. *Annals of Emergency Medicine*, 24, 885–889.
- Future of emergency care series: Emergency care for children, growing pains (2006a). Committee on the Future of Emergency Care in the United States Health System, Board on Health Care Services. Washington, DC: Institute of Medicine of the National Academies. The National Academy Press
- Future of emergency care series: Hospital-based emergency care, at the breaking point (2006b). Committee on the Future of Emergency Care in the United States Health System, Board on Health Care Services. Washington, DC: Institute of Medicine of the National Academies. The National Academy Press
- Gausche, M., Henderson, D. P., & Seidel, J. S. (1990). Vital signs as part of the prehospital assessment of the pediatric patient: A survey of paramedics. *Annals of Emergency Medicine*, 19, 173–8.
- Goldfarb, M. G., Bazzoli, G. J., & Coffey, R. M. (1996). Trauma systems and the costs of trauma care. *Health Services Research*, 31, 71–95.
- Gomez, D., Berube, M., Ziong, W., et al. (2010). Identifying targets for potential interventions to reduce rural trauma deaths: A population-based analysis. *Journal of Trauma*, 69, 633–9.
- Gray, A., Goyder, E. C., Goodacre, S. W., et al. (1997). Trauma triage: A comparison of CRAMS and TRTS in a UK population. *Injury*, 28, 97–101.
- Green, S. M. (2009). Trauma surgery: Discipline in crisis. *Annals of Emergency Medicine*, 54, 198–207.
- Hall, J. R., Reyes, H. M., & Meller, J. L. (1996). The outcome for children with blunt trauma is best at a pediatric trauma center. *Journal of Pediatric Surgery*, 31, 72–77.
- Hannan, E. L., Farrell, L. S., Cooper, A., et al. (2005). Physiologic trauma triage criteria in adult trauma patients: Are they effective in saving lives by transporting patients to trauma centers? *Journal of the American College of Surgeons*, 200, 584–592.
- Henry, M. C., Hollander, J. E., Alicandro, J. M., et al. (1996). Incremental benefit of individual American College of Surgeons trauma triage criteria. *Academic Emergency Medicine*, 3, 992–1000.
- Hoff, W. S., Tinkoff, G. H., Lucke, J. F., et al. (1992). Impact of minimal injuries on a level I trauma center. *Journal of Trauma*, 33, 408–412.
- Hsia, R. Y., Wang, E., Torres, H., et al. (2010). Disparities in trauma center access despite increasing utilization: Data from California, 1999 to 2006. *Journal of Trauma*, 68, 217–24.
- Hulka, F., Mullins, R. J., Mann, N. C., et al. (1997). Influence of a statewide trauma system on pediatric hospitalization and outcome. *Journal of Trauma*, 42, 514–519.
- Iserson, K. V., & Moskop, J. C. (2007). Triage in medicine, part I: Concept, history, and types. *Annals of Emergency Medicine*, 49, 275–281.
- Johnson, D. L., & Krishnamurthy, S. (1996). Send severely head-injured children to a pediatric trauma center. *Pediatric Neurosurgery*, 25, 309–314.
- Jurkovich, G. J., & Mock, C. (1999). Systematic review of trauma system effectiveness based on registry comparisons. *Journal of Trauma*, 47, S46–55.
- Kane, G., Engelhardt, R., Celentano, J., et al. (1985). Empirical development and evaluation of out of hospital trauma triage instruments. *Journal of Trauma*, 25, 482–9.
- Kaufmann, C. R., Maier, R. V., Rivara, F. P., et al. (1990). Evaluation of the pediatric trauma score. *Journal of the American Medical Association*, 263, 69–72.
- Knopp, R., Yanagi, A., Kallsen, G., et al. (1988). Mechanism of injury and anatomic injury as criteria for out of hospital trauma triage. *Annals of Emergency Medicine*, 17, 895–902.
- Laupacis, A., Sekar, N., & Stiell, I. G. (1997). Clinical prediction rules: A review and suggested modifications of methodological standards. *Journal of the American Medical Association*, 277, 488–494.

- Lerner, E. B. (2006). Studies evaluating current field triage: 1966–2005. *Prehospital Emergency Care, 10*, 303–306.
- Lerner, E. B., & Moscati, R. M. (2001). The golden hour: Scientific fact or medical “urban legend”? *Academic Emergency Medicine, 8*, 758–760.
- Lipsky, A. M., Gausche-Hill, M., Henneman, P. L., et al. (2006). Prehospital hypotension is a predictor of the need for an emergent, therapeutic operation in trauma patients with normal systolic blood pressure in the emergency department. *Journal of Trauma, 61*, 1228–33.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.
- Long, W. B., Bachulis, B. L., & Hynes, G. D. (1986). Accuracy and relationship of mechanisms of injury, trauma score, and injury severity score in identifying major trauma. *American Journal of Surgery, 151*, 581–584.
- Ma, M. H., MacKenzie, E. J., Alcorta, R., et al. (1999). Compliance with prehospital triage protocols for major trauma patients. *Journal of Trauma, 46*, 168–75.
- MacKenzie, E. J., Rivara, F. P., Jurkovich, G. J., et al. (2006). A national evaluation of the effect of trauma-center care on mortality. *New England Journal of Medicine, 354*, 366–378.
- MacKenzie, E. J., Weir, S., Rivara, F. P., et al. (2010). The value of trauma center care. *Journal of Trauma, 69*, 1–10.
- Mackersie, R. C. (2006). History of trauma field triage development and the American College of Surgeons criteria. *Prehospital Emergency Care, 10*, 287–294.
- Mann, N. C., MacKenzie, E., Teitelbaum, S. D., et al. (2005). Trauma system structure and viability in the current healthcare environment: A state-by-state assessment. *Journal of Trauma, 58*, 136–147.
- Mann, N. C., Mullins, R. J., & Hedges, J. R. (2001). Mortality among seriously injured patients treated in remote rural trauma centers before and after implementation of a statewide trauma system. *Medical Care, 39*, 643–653.
- Mann, N. C., Pinkney, K. A., Price, D. D., et al. (2002). Injury mortality following the loss of air medical support for rural inter-hospital transport. *Academic Emergency Medicine, 9*, 694–698.
- McConnell, K. J., Johnson, L. A., Arab, N., et al. (2007). The on-call crisis: A statewide assessment of the costs of providing on-call specialist coverage. *Annals of Emergency Medicine, 49*, 727–733.
- McConnell, K. J., Newgard, C. D., & Lee, R. (2008). Changes in the cost and management of emergency department on-call coverage: Evidence from a longitudinal statewide survey. *Annals of Emergency Medicine, 52*, 635–642.
- Moskop, J. C., & Iserson, K. V. (2007). Triage in medicine, part II: Underlying values and principles. *Annals of Emergency Medicine, 49*, 282–287.
- Mulholland, S. A., Gabbe, B. J., & Cameron, P. (2005). Is paramedic judgment useful in prehospital trauma triage? *Injury, International Journal of the Care Injured, 36*, 1298–1305.
- Mullins, R. J., & Mann, N. C. (1999). Population-based research assessing the effectiveness of trauma systems. *Journal of Trauma, 47*, S59–66.
- Mullins, R. J., Mann, N. C., Hedges, J. R., et al. (1998). Preferential benefit of implementation of a statewide trauma system in one of two adjacent states. *Journal of Trauma, 44*, 609–617.
- Mullins, R. J., Veum-Stone, J., Hedges, J. R., et al. (1996). Influence of a statewide trauma system on location of hospitalization and outcome of injured patients. *Journal of Trauma, 40*, 536–545.
- Mullins, R. J., Veum-Stone, J., Helfand, M., et al. (1994). Outcome of hospitalized injured patients after institution of a trauma system in an urban area. *Journal of the American Medical Association, 271*, 1919–1924.
- Nathens, A. B., Jurkovich, G. J., & Rivara, F. P. (2000). Effectiveness of state trauma systems in reducing injury-related mortality: A national evaluation. *Journal of Trauma, 48*, 25–30.
- Nathens, A. B., Maier, R. V., Brundage, S. I., et al. (2003). The effect of interfacility transfer on outcome in an urban trauma system. *Journal of Trauma, 55*, 444–449.
- Nayduch, D. A., Moylan, J., Rutledge, R., et al. (1991). Comparison of the ability of adult and pediatric trauma scores to predict pediatric outcome following major trauma. *Journal of Trauma, 31*, 452–8.
- Newgard, C. D., Cudnik, M., Warden, C. R., et al. (2007). The predictive value and appropriate ranges of prehospital physiological parameters for high-risk injured children. *Pediatric Emergency Care, 23*, 450–456.
- Newgard, C. D., & Haukoos, J. (2007). Missing data in clinical research – part 2: Multiple imputation. *Academic Emergency Medicine, 14*, 669–678.
- Newgard, C. D., Hedges, J. R., Diggs, B., et al. (2008). Establishing the need for trauma center care: Anatomic injury or resource use? *Prehospital Emergency Care, 12*, 451–458.
- Newgard, C. D., Hui, J., Griffin, A., et al. (2005). Prospective validation of a clinical decision rule to identify severely injured children at the scene of motor vehicle crashes. *Academic Emergency Medicine, 12*, 679–687.
- Newgard, C. D., Lewis, R. J., & Jolly, B. T. (2002). Use of out-of-hospital variables to predict severity of injury in pediatric patients involved in motor vehicle crashes. *Annals of Emergency Medicine, 39*, 481–491.
- Newgard, C. D., McConnell, K. J., & Hedges, J. R. (2006). Variability of trauma transfer practices among non-tertiary care hospital emergency departments. *Academic Emergency Medicine, 13*, 746–754.
- Newgard, C. D., McConnell, K. J., Hedges, J. R., et al. (2007). The benefit of higher level of care transfer of injured patients from non-tertiary care hospital emergency departments. *Journal of Trauma, 63*, 965–971.
- Newgard, C. D., Nelson, M. J., Kamp, M., et al. (2011). Out-of-hospital decision-making and factors influencing the regional distribution of injured patients in a trauma system. *Journal of Trauma, 70*, 1345–53.

- Newgard, C. D., Rudser, K., Atkins, D. L., et al. (2009). The availability and use of out-of-hospital physiologic information to identify high-risk injured children in a multisite, population-based cohort. *Prehospital Emergency Care, 13*, 420–31.
- Newgard, C. D., Rudser, K., Hedges, J. R., et al. (2010). A critical assessment of the out-of-hospital trauma triage guidelines for physiologic abnormality. *Journal of Trauma, 68*, 452–62.
- Newgard, C. D., Schmicker, R., Hedges, J. R., et al. (2010). Emergency medical services time intervals and survival in trauma: Assessment of the “Golden Hour” in a North American prospective cohort. *Annals of Emergency Medicine, 55*, 235–246.
- Newgard C. D., Zive D, Holmes J. F., et al. (2012). A multi-site assessment of the ACSCOT field triage decision scheme for identifying seriously injured children and adults. *Journal of American College of Surgeons* (In Press).
- Nirula, R., Maier, R., Moore, E., et al. (2010). Scoop and run to the trauma center or stay and play at the local hospital: Hospital transfer's effect on mortality. *Journal of Trauma, 69*, 595–601.
- Norcross, E. D., Ford, D. W., Cooper, M. E., et al. (1995). Application of American college of surgeons' field triage guidelines by pre-hospital personnel. *Journal of the American College of Surgeons, 181*, 539–544.
- Phillips, J. A., & Buchman, T. G. (1993). Optimizing out of hospital triage criteria for trauma team alerts. *Journal of Trauma, 34*, 127–32.
- Phillips, S., Rond, P. C., Kelly, S. M., et al. (1996). The need for pediatric-specific triage criteria: Results from the Florida trauma triage study. *Pediatric Emergency Care, 12*, 394–398.
- Pointer, J. E., Levitt, M. A., Young, J. C., et al. (2001). Can paramedics using guidelines accurately triage patients? *Annals of Emergency Medicine, 38*, 268–77.
- Potoka, D. A., Schall, L. C., & Ford, H. R. (2001). Development of a novel age-specific pediatric trauma score. *Journal of Pediatric Surgery, 36*, 106–112.
- Pracht, E. E., Tepas, J. J., & Celso, B. G. (2007). Survival advantage associated with treatment of injury at designated trauma centers. *Medical Care Research and Review, 64*, 83–97.
- Pracht, E. E., Tepas, J. J., Langland-Orban, B., et al. (2008). Do pediatric patients with trauma in Florida have reduced mortality rates when treated in designated trauma centers? *Journal of Pediatric Surgery, 43*, 212–221.
- Qazi, K., Kempf, J. A., Christopher, N. C., et al. (1998). Paramedic judgment of the need for trauma team activation for pediatric patients. *Academic Emergency Medicine, 5*, 1002–1007.
- Resources for the Optimal Care of the Injured Patient (1993). Chicago, IL: American College of Surgeons
- Resources for the Optimal Care of the Injured Patient (1999). Chicago, IL: American College of Surgeons
- Resources for the Optimal Care of the Injured Patient (2006). Chicago, IL: American College of Surgeons
- Rivara, F. P., Nathens, A. B., Jurkovich, G. J., et al. (2006). Do trauma centers have the capacity to respond to disasters? *Journal of Trauma, 61*, 949–953.
- Sampalis, J. S., Denis, R., Frechette, P., et al. (1997). Direct transport to tertiary trauma centers versus transfer from lower level facilities – impact on mortality and morbidity among patients with major trauma. *Journal of Trauma, 43*, 288–296.
- Sampalis, J. S., Denis, R., Lavoie, A., et al. (1999). Trauma care regionalization: A process-outcome evaluation. *Journal of Trauma, 46*, 565–581.
- Sampalis, J. S., Lavoie, A., Williams, J. I., et al. (1993). Impact of on-site care, prehospital time, and level of in-hospital care on survival in severely injured patients. *Journal of Trauma, 34*, 252–261.
- Scheetz, L. J. (2003). Effectiveness of prehospital trauma triage guidelines for the identification of major trauma in elderly motor vehicle crash victims. *Journal of Emergency Nursing, 29*, 109–115.
- Shafi, S., Nathens, A. B., Elliott, A. C., et al. (2006). Effect of trauma systems on motor vehicle occupant mortality: A comparison between states with and without a formal system. *Journal of Trauma, 61*, 1374–8.
- Shatney, C. H., & Sensaki, K. (1994). Trauma team activation for “mechanism of injury” blunt trauma victims: Time for a change? *Journal of Trauma, 37*, 275–281.
- Simmons, E., Hedges, J. R., Irwin, L., et al. (1995). Paramedic injury severity perception can aid trauma triage. *Annals of Emergency Medicine, 26*, 461–468.
- Simon, B. J., Legere, P., Emhoff, T., et al. (1994). Vehicular trauma triage by mechanism: Avoidance of the unproductive evaluation. *Journal of Trauma, 37*, 645–649.
- Smith, J. S., & Bartholomew, M. J. (1990). Trauma index revisited: A better triage tool. *Critical Care Medicine, 18*, 174–180.
- Steele, R., Gill, M., Green, S. M., et al. (2007). Do the American College of Surgeons' “Major Resuscitation” trauma triage criteria predict emergency operative management? *Annals of Emergency Medicine, 50*, 1–6.
- Stiell, I. G., & Wells, G. A. (1999). Methodologic standard for the development of clinical decision rules in emergency medicine. *Annals of Emergency Medicine, 33*, 437–447.
- Sun, B. C., Mohanty, S. A., Weiss, R., et al. (2006). Effects of hospital closures and hospital characteristics on emergency department ambulance diversion, Los Angeles County, 1998 to 2004. *Annals of Emergency Medicine, 47*, 309–316.
- Tepas, J. J., Ramenofsky, M. L., Mollit, D. L., et al. (1988). The pediatric trauma score as a predictor of injury severity: An objective assessment. *Journal of Trauma, 28*, 425–429.

- The EAST Practice Management Guidelines Work Group (2010). Practice management guidelines for the appropriate triage of the victim of trauma. Eastern Association for the Surgery of Trauma
- Todd, S. R., Arthur, M., Newgard, C., et al. (2004). Hospital factors associated with splenectomy for splenic injury: A national perspective. *Journal of Trauma*, 57, 1065–1071.
- Utter, G. H., Maier, R. V., Rivara, F. P., et al. (2006). Inclusive trauma systems: Do they improve triage or outcomes of the severely injured? *Journal of Trauma*, 60, 529–35.
- Van Der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T., et al. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, 59, 1102–9.
- Vassar, M. J., Holcroft, J. J., Knudson, M. M., et al. (2003). Fractures in access to and assessment of trauma systems. *Journal of the American College of Surgeons*, 197, 717–725.
- Wang, N. E., Saynina, O., Kuntz-Duriseti, K., et al. (2008). Variability in pediatric utilization of trauma facilities in California: 1999 to 2005. *Annals of Emergency Medicine*, 52, 607–615.
- West, J. G., Murdock, M. A., Baldwin, L. C., et al. (1986). A method for evaluating field triage criteria. *Journal of Trauma*, 26, 655–9.
- Young, J. S., Bassam, D., Cephas, G. A., et al. (1998). Inter-hospital versus direct scene transfer of major trauma patients in a rural trauma system. *American Surgeon*, 64, 88–91.
- Zechnich, A. D., Hedges, J. R., Spackman, K., et al. (1995). Applying the trauma triage rule to blunt trauma patients. *Academic Emergency Medicine*, 2, 1043–1052.

Chapter 16

Clinical Prediction Rules

James F. Holmes

Introduction

A “clinical prediction rule” is a set of variables used to assist clinicians in their evaluation of a patient at risk for a particular disease or outcome from a disease. Such tools are increasingly developed by the medical community to optimize the decision-making process (Laupacis et al. 1997; Stiell and Wells 1999). Due to the nature of injured patients, prediction rules have an important role in maximizing the evaluation and management of trauma victims as they help trauma physicians cope with the diagnostic and therapeutic uncertainties inherent to this setting. Patients with injuries to the ankle, knee, cervical spine, and head are more appropriately managed with the use of prediction rules (Perry and Stiell 2006).

Unfortunately, “prediction rule” terminology varies. The term “rule” is frequently interchanged with “tool” or “instrument” and the term “prediction” is frequently interchanged with “decision.” Although the general concept is the same, different implications exist with different terminologies, as the terms “decision” and “rule” imply a course of action must be taken, whereas “tool” and “instrument” provide guidance to the clinician and do not mandate action. Furthermore, “prediction” implies the patient is categorized into one of several classes, whereas “decision” implies the patient is categorized into one of two classes (yes/no obtain imaging studies, yes/no patient has disease, etc.). In this chapter, the term “prediction rule” is used as it was suggested in the original description articles (Laupacis et al. 1997; Wasson et al. 1985). Regardless of the terminology, the general concepts for the idea are the same. Ultimately, the prediction rule is evidence based and used to assist the clinician in patient management. Although some consider the prediction rule to mandate a particular action (such as obtaining a diagnostic test), others consider the prediction rule to simply guide or assist clinicians in their patient care. The clinician must be aware of the methods by which the prediction rule was developed (particularly the patient population studied and successful validation of the rule) and the intended action the prediction rule imparts.

Clinical prediction rules are well suited for the evaluation and management of patients with traumatic injuries. Errors in the evaluation and management of trauma patients are often preventable when prediction rules or guidelines are followed (Hoyt et al. 1994). Implementing formal, defined trauma protocols into the emergency departments (EDs) has demonstrated improved resource utilization and improved patient care (Nuss et al. 2001; Palmer et al. 2001; Sarioe 2000; Tinkoff et al.

J.F. Holmes, MD, MPH (✉)

Department of Emergency Medicine, University of California at Davis School of Medicine,
2315 Stockton Blvd. PSSB 2100, Sacramento, CA 95817, USA
e-mail: jfholmes@ucdavis.edu

1996). Trauma patients require rapid assessment, appropriate diagnostic imaging, and treatment based upon the diagnostic evaluation. Much variation exists with the diagnostic evaluation of the injured patient and such variation limits optimal care. Well-developed prediction rules can guide clinicians to collect the important clinical data pieces and to provide evidence-based care. Applying these prediction rules removes variability and minimizes missed injuries and excessive utilization of diagnostic testing and limited resources.

Numerous examples of prediction rules for the evaluation of injured patients now exist. Radiographic imaging for the diagnosis of traumatic injuries is perhaps the ideal setting for clinical prediction rules because many diagnostic evaluation schemes for trauma are protocol driven (Blackmore 2005; Hunink 2005). Most trauma prediction rules focus on appropriate radiographic evaluation (Blackmore 2005), especially CT scan utilization (Haydel et al. 2000; Holmes et al. 2002a, 2009a; Kuppermann et al. 2009; Mower et al. 2005; Stiell et al. 2001a) and to a lesser extent plain radiography (Holmes et al. 2002b; Stiell et al. 1993, 1995a, 2001b; Rodriguez et al. 2011). Prediction rules, however, have also been developed for numerous other trauma scenarios including: use of laboratory testing (Langdorf et al. 2002), performing a rectal examination (Guldner et al. 2004), determining appropriate trauma transfer (Newgard et al. 2005a), performing a laparotomy after a positive abdominal ultrasound examination (Rose et al. 2005), and both primary and secondary trauma triage (Newgard et al. 2005b; Steele et al. 2006).

Grading the Clinical Prediction Rules

Investigators have arbitrarily suggested levels of evidence for prediction rules (McGinn et al. 2000). Although this stratification provides a template, it is limited in its ability to clarify certain degrees of differences and validation. Table 16.1 builds on this prior description and more definitively classifies levels of prediction rule quality and implementation. Prior to implementation of any prediction rule, it is critical that appropriate validation is accomplished. This chapter highlights the different criteria to develop and grade prediction rules.

Table 16.1 Grades of clinical prediction rules

	A	B	C	D
Stage of prediction rule development	<ul style="list-style-type: none"> Prospective validation in separate, large cohort Impact analysis demonstrates improved patient care 	<ul style="list-style-type: none"> Prospective validation in separate cohort Prospective split sample validation in very large sample No impact analysis 	<ul style="list-style-type: none"> Prospective derivation with retrospective validation Prospective split sample validation in small/moderate size sample Retrospective derivation and validation with very large samples 	<ul style="list-style-type: none"> Retrospective derivation and validation in small/moderate sample Prospective derivation and validation solely with statistical techniques
Appropriate use	Actively disseminate and implement rule	Implement in appropriate settings	Use rule with caution	None

Development of the Clinical Prediction Rule

Methodologic criteria for the development of clinical prediction rules were initially described in the mid-1980s (Wasson et al. 1985; Feinstein 1987). Subsequently, development of prediction rules became increasingly popular and the appropriate methodologic standards are now well established (Laupacis et al. 1997; Stiell and Wells 1999; McGinn et al. 2000). Figure 16.1 provides an overview of the process of prediction rule development.

Need for a Clinical Prediction Rule

Prior to the actual development of a prediction rule, a clinical need for the prediction rule must exist. This includes addressing the following (1) variation in clinician practice, (2) risk/cost of the resource, and (3) physician desire/perceived need for a rule. Some investigators suggest developing prediction rules only for common clinical problems (Stiell and Wells 1999), but significant variation in practice likely occurs more frequently with less common injuries (e.g., aortic injury) and prediction rules are almost assuredly helpful for patients with rare injuries (Ungar et al. 2006). Unfortunately, instances where the disease or disease outcome is rare, collecting a sufficient sample to prospectively derive

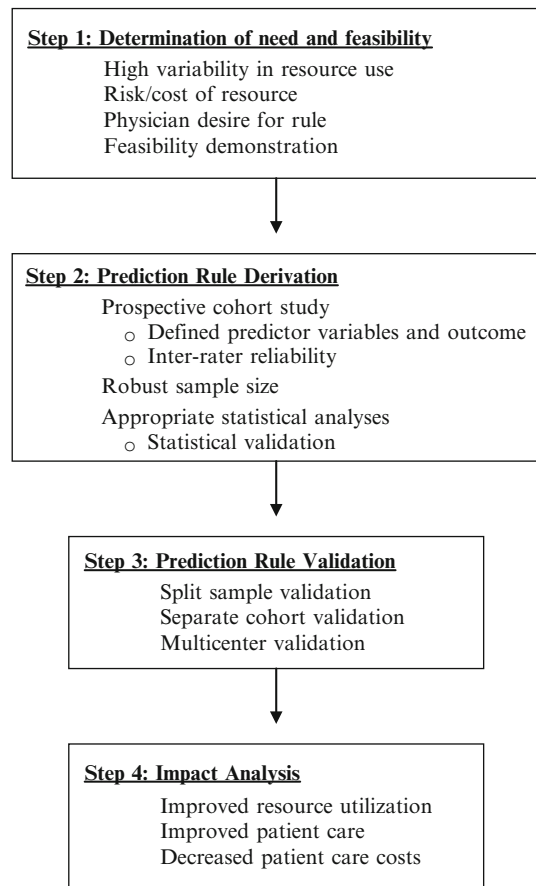


Fig. 16.1 Development of a clinical prediction rule

and validate a prediction rule is logistically difficult. In these scenarios, evaluating large retrospective databases may serve as the first step in the development of a prediction rule (Holmes et al. 1999; Fine et al. 1997).

Variation in care is a source of clinical inefficiency, especially in trauma care (Glance et al. 2010; Minei et al. 2010; Culica and Aday 2008; Bowman et al. 2005). Variation in resource utilization appropriate for a prediction rule includes diagnostic test utilization, providing specific therapy, or determining appropriate patient disposition. Significant variation existed among physician ordering of cervical spine radiographs after trauma (Stiell et al. 1997). Subsequently two prediction rules for trauma cervical spine radiography were developed (Stiell et al. 2001b; Hoffman et al. 2000). Furthermore, demonstrating the magnitude of clinical inefficiency strengthens the cause for prediction rule development. Examples of such include the inefficiency of abdominal CT use in trauma (Garber et al. 2000), cranial CT use in children with minor head trauma (Klassen et al. 2000), trauma knee radiography (Stiell et al. 1995b), and intensive care unit utilization in patients with traumatic brain injury (Nishijima et al. 2010). Demonstrating variation and inefficient resource utilization provides the background for prediction rule development.

Generally, some risk or drawback to the resource being used should exist. Radiologic testing is now a focus of prediction rules due to concerns of overuse and the risk of radiation-induced malignancy, especially with CT scanning (Brenner and Hall 2007). In the current economic environment of expanding healthcare costs, cost savings is driving development of prediction rules as inefficiency of resource use significantly impacts hospital costs (Nishijima et al. 2010).

Finally, physician willingness for the rule and desire to use the rule should exist. A methodologically sound prediction rule that improves patient care is ideal, but if physicians never utilize the rule, it is simply wasted. Surveys suggest emergency physicians routinely order radiographic imaging to “rule out” fracture despite believing the patient is at very low risk (Stiell et al. 1995b) and these physicians are truly interested in implementing well-developed prediction rules (Graham et al. 1998). Determining actual physician desire for a prediction rule, however, is likely more difficult than simply surveying physicians, because discrepancies exist between actual physician practice and survey reports of behavior (Bandiera et al. 2003).

After demonstrating prediction rule need but prior to expending the considerable energy to derive and validate a rule, prediction rule feasibility is determined. Such a feasibility assessment is frequently combined with the need assessment. This feasibility/need assessment is often a retrospective analysis of the problem of interest and includes gathering data on the variability of care, potential predictor variables to be studied, and prevalence of the outcome of interest in the anticipated study population (Holmes et al. 1999; Klassen et al. 2000; Nishijima and Sena 2010). Results from this study provide necessary data to determine overall feasibility of prospectively deriving a prediction rule by (1) providing insight into the probability of the outcome being predicted by the variables of interest, (2) estimating the approximate sample size for the derivation study, and (3) determining the time needed for the sample to be collected. If the feasibility study demonstrates appropriate use of the resource, the lack of variables predictive of the outcome of interest, or a non-feasible sample size, then the investigator may wish to abort the process.

Prediction Rule Sensibility

To be clinically useful, the prediction rule must be sensible (i.e., clinically rational) (Feinstein 1987) and investigators must consider this in their planning. The rule should have face validity in that the predictor variables are anticipated by the clinicians and have biologic plausibility. A prediction rule for CT scanning patients with head trauma that includes a variable of “leg pain” lacks clinical sensibility

and is unlikely to be implemented. Furthermore, clinicians will have reservation in utilizing a prediction rule lacking a variable believed very important. A recently derived and validated prediction rule for avoiding cranial CT scanning in children with blunt head trauma does not include vomiting in those younger than 2 years (Kuppermann et al. 2009). The variable was not independently important in the derivation or validation of the rule, but physicians' beliefs regarding the importance of this variable must be overcome for successful implementation of the rule.

Prediction Rule Derivation

Once need and feasibility of a prediction rule are described, the initial derivation of the prediction rule is performed following rigorous methodologic standards. The derivation study cohort involves gathering either prospective or retrospective data. Unless an inherent necessity for retrospective data exists (see below), prediction rules are most appropriately developed from prospective data (Stiell and Wells 1999). The multiple advantages of prospective data as compared to retrospective data include the following:

1. *Documentation of variables prior to clinician knowledge of the outcome of interest.* Researchers can mandate specific variable documentation prior to knowledge of the outcome of interest. Such action is impossible in retrospective cohorts as clinicians frequently complete medical record documentation after knowledge of the outcome of interest and bias is introduced into their documentation of potential predictor variables. For example, a clinician is more likely to document the presence of abdominal tenderness if an abdominal injury is known to be present on abdominal CT and more likely to document no abdominal tenderness if an abdominal injury is known to be absent on CT. Thus, the variables of interest are most reliably documented prior to knowledge of the outcome of interest.
2. *Explicit variable definition.* Prospective data collection allows for explicitly defining variables of interest. In prospective data collection, a "seat belt sign" was defined as a continuous area of erythema/contusion across the abdomen secondary to a lap restraint (Sokolove et al. 2005). Such a definition excludes lap belt related abrasions located only on the anterior iliac crests that are not continuous. Some physicians will document abrasions solely on the iliac crests or over the chest wall as "seat belt sign" in the patient's medical record. In a retrospective study, the medical record abstractors would document a seat belt sign present in these cases. In such a scenario, the frequency of the intended variable is overestimated and the actual association with the outcome of interest is diluted.
3. *Collection of all the variables of interest.* A variable of interest that is not routinely included in the clinician's history and physical examination can be explicitly recorded in a prospective study. Bowel sound auscultation is often not performed during the abdominal evaluation of the injured patient and would not be routinely documented in a medical record review.
4. *Missing data is minimized.* Retrospective data have more missing data than data collected prospectively (Stiell and Wells 1999). Abdominal inspection and palpation are routine parts of the trauma examination, but clinicians may fail to document a complete abdominal examination in the medical record leading to missing data.

Despite the benefits of prospective data collection, instances exist where retrospective data collection is necessary. If the disease is rare, the disease outcome is rare, or a specific complication/treatment is rare, then prospective data collection (especially at a single center) is extremely difficult and potentially impossible. In such cases, investigators may proceed with retrospective data collection or analyzing large databases to gather a sample sufficient to derive a prediction rule.

The increasing frequency of multicenter research networks makes the need for retrospective data collection less necessary.

Population (subject selection). The sampled population is critical to the performance of the prediction rule. The study population must be generalizable and representative such that a successfully derived and validated prediction rule can be implemented into clinical care. Reporting the study population includes the well-defined inclusion/exclusion criteria and appropriate demographic (age, gender, and race) and historical (mechanism of injury) information. Explicitly defined inclusion/exclusion criteria along with a well-described population provide the reader with the ability to appropriately implement the prediction rule to the correct population. Finally, the study site(s) (trauma center, urban, patient volume, teaching hospital, etc.) must be described in detail as important differences in patient populations may exist among hospitals.

Although enrolling a too-restrictive sample limits the generalizability of the results, the inclusion of “inappropriate” subjects must be limited. For example, in the creation of a prediction rule for determining cranial CT use in patients with blunt head trauma, patients on warfarin or those presenting more than 24 h after the traumatic event are unlikely to be representative of the intended population. Thus, such patients are appropriately excluded (Haydel et al. 2000; Kuppermann et al. 2009; Stiell et al. 2001a). However, including only patients with certain mechanisms of injury (e.g., creating a prediction rule for CT use in patients with blunt head trauma from a motor vehicle collision) is overly restrictive and not clinically useful. Thus, the inclusion and exclusion criteria must be well described so that the clinician understands the appropriate population to apply the prediction rule. It is inappropriate to apply to patients with GCS scores <14 a prediction rule developed for abdominal CT scanning in trauma patients with GCS scores of 14 and 15.

Prediction rules are often derived on a population selected because they are undergoing a particular diagnostic test. Such an enrollment strategy has the benefit of being logistically simple, as it requires fewer resources than a strategy of enrolling all patients meeting predefined historical and physical examination criteria. Physicians simply complete a data collection form once they decide to obtain the diagnostic test being studied. Thus, the inclusion criterion is obtaining the specific diagnostic test. These strategies were used for large studies on patients with cervical spine, head, and abdominal trauma (Holmes et al. 2009a; Mower et al. 2005; Hoffman et al. 2000). Some experts have criticized this methodology by noting that physicians may use different thresholds for ordering the diagnostic test, resulting in a biased sample (Stiell and Wells 1999). These prediction rules, however, provide useful information when applied in appropriate context. This strategy (including only those patients undergoing a particular diagnostic test) results in a prediction rule that identifies a population that does not benefit from the diagnostic test. Patients without any of the variables in the prediction rule should not undergo the diagnostic test under study. The rule, however, does not mandate obtaining the diagnostic test, if the patient is positive for one of the variables in the rule as the clinician must make this determination. This particular study design is aimed only at reducing use of the diagnostic test (usually radiographic imaging).

In a prediction rule for CT scanning following abdominal trauma, the absence of all important variables indicates that the patient is at such low risk for abdominal injury requiring therapy that abdominal CT scanning is not indicated (Holmes et al. 2009b). The presence of one of the variables increases the patient’s risk for abdominal injury but does not necessarily mandate CT scanning. Furthermore, the strategy of including all patients undergoing the diagnostic test ensures that patients have a diagnostic test to identify the injury being studied and in cases of CT utilization, this diagnostic test is also the gold standard. Patients enrolled based on predefined historical and physical examination criteria but not undergoing a diagnostic test to determine presence or absence of the outcome become problematic, because it is not known whether the patient actually has the injury. In such cases, methods of follow-up including telephone or clinical follow-up, hospital continuous quality improvement record and/or death record reviews are completed (Kuppermann et al. 2009; Holmes et al. 2002b; Stiell et al. 1996).

Predictor variables. Potential variables considered for a prediction rule require explicit definition. This allows the clinicians completing the data form to understand the intent of the investigators and enhances replication of study results. Therefore, limited training of the clinicians completing the data collection forms is necessary. Care must be taken not to over-train those completing the data form as this may limit the generalizability of the prediction rule. Poorly defined predictor variables threaten the internal validity of the prediction rule. Inclusion of variables termed as “soft” or “subjective” is acceptable as long as they are adequately described and carry the same meaning between physicians (Feinstein 1977).

All variables considered for the prediction rule are included on the paper/electronic data collection form. Collecting “too much” data, however, may dilute data reliability. Investigators must weigh the need to collect the necessary data versus the time needs of the clinician. Prediction rule development relies on busy clinicians to complete data forms. Thus, excessively long forms may not be completed; also, questions may be left blank due to perceived time limitation of the clinician or may be answered incorrectly due to physician inattentiveness. Pilot testing the data collection forms and limiting any unnecessary data points minimizes these risks.

Determining the Variables for Consideration in the Rule. Ultimately, a prediction rule is intended for clinical use. Thus, the rule must be simple to implement and consist of biologically plausible variables. Prediction rules that are difficult to remember due to excessive or unusual variables are unlikely to be implemented by clinicians. Continuous variables are often dichotomized for simplicity. Although this action typically costs statistical power, dichotomizing a variable simplifies the variable. Appropriate categorization of a continuous variable is determined by the following methods (1) identifying acceptable cutpoints within the literature, (2) calculating a receiver operating characteristic curve and selecting the best cutpoint, or (3) entering the variable as a continuous variable in the recursive partitioning model and allowing the software to identify the appropriate cutpoint. The latter method, however, may result in a categorization that is nonsensical to the clinician. A prediction rule with a variable of systolic blood pressure <97 mmHg is unlikely to be remembered by clinicians as opposed to a variable of systolic blood pressure <90 or <100 mmHg. With the increased availability of electronic support to assist clinicians, this limitation is likely to be minimized in the future.

Missing data. Regardless of the data collection method, missing data occurs. This problem is especially likely with retrospective data collection, but it also occurs, to a lesser degree, with prospective data collection (Stiell and Wells 1999). Prior to study implementation, investigators should design methods to assess and limit missing data. Preventing missing data involves pilot testing the data collection form to identify locations at risk for missing data. A data point may be left blank because of poor wording or the location of the question on the data collection form. Bolding and shading areas of the data collection form can help to prevent this. Variables left missing in prospective data collection in more than 5% of cases are often considered inappropriate for inclusion into a prediction rule (Kuppermann et al. 2009; Holmes et al. 2002b), based on the belief that inherent problems with the variable exist if it is missing to such a degree.

Statistical methods for handling missing data include excluding cases with any missing data (complete case analysis), overall mean imputation (Donders et al. 2006), and the missing-indicator method (Altman and Royston 2000), although these methods may lead to bias (Greenland and Finkle 1995; Little 1988). Imputation is a better option than excluding the case or using simple methods to replace data, because bias is avoided, assuming the data are missing at random (Donders et al. 2006; Janssen et al. 2010; van der Heijden et al. 2006). Acceptable methods for imputation of missing data are available and included in statistical software (Vergouwe et al. 2010a). Multiple imputation is preferred over single imputation as single imputation produces unsatisfactory small standard estimates. Simplistically, imputation predicts an “imputed” value for the missing variable by using available data.

An important data point not often collected is the clinician's impression of the outcome of interest. Such a variable demonstrates physician variability and allows comparison between prediction rule performance and physician performance. Investigators are cautioned, however, that physician reported belief of the patient's risk of the outcome and their actual belief may not be congruent. A physician may report a belief that the patient is at no risk for a particular outcome but if the physician orders the diagnostic test then one must question if the physician truly believes the patient does not have the outcome (Bandiera et al. 2003). In addition, some investigators collect data (especially in validation studies of previously derived prediction rules) measuring physician acceptance of the prediction rules.

Outcome Definition. Carefully defining the outcome of interest is critical to the performance of the rule. The outcome must be reliably assessed and clinically sensible.

Historically, clinical prediction rules are derived to identify all patients who are diagnosed with the particular injury or disease of interest (*disease-oriented outcome*). These outcomes are most appropriate when the disease requires specific treatment in essentially all cases (i.e., pulmonary embolus or myocardial infarction). More recent investigations have generated clinical prediction rules focusing on *patient-oriented outcomes* by defining the outcome of interest to be "clinically significant injuries" or "injuries requiring specific treatment" (Holmes et al. 2009a; Kuppermann et al. 2009; Stiell et al. 2001a; Hoffman et al. 2000; Palchak et al. 2003).

Patient-oriented outcomes do not consider minor injuries or injuries not requiring therapy as outcomes of interest. Several reasons justify the use of clinically important or patient-oriented outcomes, especially in trauma. First, failure to identify an injury never requiring intervention does not alter the patient's clinical course and, thus, has minimal or no impact on the patient's care. Second and more importantly, utilizing a patient-oriented outcome minimizes potential bias introduced with outcome misclassification bias. If the criterion standard test has a low specificity, a substantial number of patients without the disease will be diagnosed as having the disease (i.e., a false-positive test). In this case, the misclassification bias with regard to the presence of the disease would be high. If the prediction rule is developed with inclusion of a substantial number of patients with false-positive outcome results, bias is introduced into the prediction rule.

An example of this important bias is as follows. Investigators attempting to derive a clinical prediction rule to determine the indications for abdominal CT in patients at risk for intra-abdominal injury define the outcome of interest to be any intra-abdominal injury diagnosed (i.e., a disease-oriented outcome). In this study, abdominal CT is the gold standard test. The investigators anticipate enrolling a large sample size (5,000 subjects undergoing abdominal CT scan) to ensure a robust model and narrow confidence intervals. The study is conducted at a busy trauma center where the true prevalence of abdominal injury among eligible subjects is 10%. Thus, the investigators anticipate enrolling approximately 500 subjects with abdominal injury and 4,500 subjects without abdominal injury. Assuming abdominal CT has a specificity of 99% (unlikely that it is truly this high), 4,455 subjects without intra-abdominal injury have a normal abdominal CT scan (4,500 subjects without disease multiplied by the CT specificity of 0.99). Thus, 45 subjects have a false-positive abdominal CT interpretation and are now classified (inappropriately) as having the outcome of interest. Therefore, 45 subjects (8%) of the 545 considered as having intra-abdominal injury are misclassified. These 45 patients are unlikely to undergo any specific therapy (angiographic embolization or surgical repair) for their injury as no injury actually exists. However, these patients are considered positive for the outcome when the model is constructed, and in all likelihood some of these patients will certainly not be identified by the derived prediction rule since they do not have an injury. Thus, expecting a sensitivity of a clinical prediction rule to be 100% is unrealistic when testing the prediction rule against a disease-oriented outcome as this type of misclassification bias is a major problem (Holmes et al. 2002a).

Care must be taken in defining patient-oriented outcomes to avoid potential bias resulting from behavioral outcomes. In general, outcomes heavily influenced by behavioral aspects are avoided as

both environmental and physician variability substantially impact the results. As previously discussed, emergency department disposition is potentially impacted by numerous variables not related to the disease process, including insurance status, type of hospital, and physician beliefs. Care must be taken to appropriately define the outcome such that these biases are minimized. In addition, the outcome definition cannot include any variables considered for inclusion in the prediction rule. A predictor variable also included in the outcome definition will have its predictive capability falsely elevated. An example of such would be an outcome definition of a study of patients with rib fractures. If the outcome, rib fracture, is defined as any patient with rib tenderness to palpation that subsequently increases with inspiration, then chest wall tenderness cannot be included as a predictor variable as its characteristics will be biased.

Finally, it is critical that the outcome of interest be applicable in all settings. Investigators wishing to develop a prediction rule that appropriately determines disposition from the emergency department (e.g., discharge, ICU admission, and ward admission) should define an appropriate outcome that warrants such disposition. Emergency department disposition is impacted by a multitude of factors and varies significantly among hospitals and physicians. Therefore, determining actual patient need for ICU or hospital admission is more appropriate than simply modeling on an outcome of patient disposition. Defined criteria for ICU admission from the emergency department exist (Nishijima et al. 2010). A study to derive a prediction rule for ICU admission would more appropriately define the outcome of interest as requiring an ICU intervention as opposed to simply being admitted to the ICU.

Outcome Assessment To avoid ascertainment bias, the methods to assess the outcome should be similar in all subjects. If the outcome is intra-abdominal injury, outcome assessment by abdominal CT on some patients and abdominal ultrasound on others introduces bias. In addition, the outcome should be assessed without awareness of the predictor variables. Knowledge of the predictor variables is most likely to impact investigator classification of the outcome if the outcome is subjective (patient quality of life) but less likely for the hard outcomes (patient disposition or death).

Statistical techniques. Appropriate statistical techniques are available for the derivation of the prediction rule and must be adequately described. Univariate analyses are frequently performed, although demonstrating association between a single variable and the outcome of interest provides useful but limited information as clinical medicine is a multifactorial process. Variables not demonstrating significant association (p value < 0.05) with univariate testing may still be predictive of the outcome, and this association may only be demonstrated via multivariate modeling. Furthermore, variables identified as associated with the outcome of interest in a univariate analysis may lose independent association in a multivariate model (Guyatt et al. 1995). Therefore, multivariate analysis is required to derive prediction rules.

Two processes for selecting variables for inclusion in a multivariate model exist. Some investigators conduct a univariate analysis and then select those variables with a predefined p value below a certain cut-point (p value < 0.20 or < 0.10) (Stiell and Wells 1999). Other investigators select the variables of interest based upon prior evidence and biological plausibility, and then enter all possible predictor variables into the model (Kuppermann et al. 2009). The latter method minimizes bias within a dataset. Drawbacks of this method include the potential failure to identify an important variable and the need for a larger sample size, as a more extensive list of variables is usually considered.

Once the candidate variables are determined, multivariate modeling is performed to create the prediction rule. The two multivariate statistical techniques most frequently utilized for prediction rule development are *regression analysis* and *binary recursive partitioning*. In addition, discriminant analysis and neural networks are occasionally utilized but have not gained popularity (Baxt 1995; Rudy et al. 1992). Available software packages can easily perform these analyses but the novice is cautioned to include appropriate statistical expertise in this endeavor. Although no consensus exists

regarding the best statistical technique for creating a prediction rule, general concepts exist. Regression modeling derives a rule with a higher overall accuracy (better overall classification of patients), while binary recursive partitioning produces a rule with better sensitivity (Laupacis et al. 1997). If identification of the disease/injury is critical and a high sensitivity is required, recursive partitioning is likely the most appropriate technique. For example, a prediction rule identifying patients with brain injury requiring neurosurgery requires a rule with very high sensitivity. Clinicians are unlikely to accept a prediction rule with a sensitivity for a life-threatening event that is less than 100% (Graham et al. 1998). If properly classifying patients is more important and near-perfect sensitivity is unnecessary, logistic regression is probably the most appropriate technique. Such a situation exists if the treatment of a particular disease is risky or failure to identify all patients with a particular outcome carries minimal risk. For example, determining appropriate hospital admission location (intensive care unit versus ward) likely requires more appropriate classification as opposed to high sensitivity due to the high resource costs. A rule with very high sensitivity but low specificity wastes valuable and limited resources. A regression model with a high overall accuracy more appropriately distributes patients and saves resources.

Regression analysis is a classic statistical method identifying the association between the outcome variable (dependent variable) and the predictor variable (independent variable). The regression model predicts changes in the outcome as the predictor variable(s) changes. Although single variable regression models are easily calculated, multivariate modeling is mandatory for the creation of prediction rules.

Most prediction rules are designed to identify a binary outcome (disease: yes/no; treatment: yes/no; etc.). In these instances, logistic regression is the appropriate regression technique. Regression modeling is limited by missing data as the statistical program drops observations with missing variables, and thus imputation is required to maintain adequate sample size (see missing data section). Most importantly, a sufficient number of subjects with the outcome of interest must be in the dataset before regression analysis is considered. Ten subjects with the outcome of interest are required for every predictor variable entered into the model (Concato et al. 1993; Harrell et al. 1985). Study sample size is often determined by this requirement.

In most instances, the results of the regression analysis become the prediction rule (the variables with independent association in the regression model compose the prediction rule) regardless of the weight of association. A patient who is positive for any of the prediction rule variables is thus considered positive for the rule. This method increases the sensitivity of the prediction rule but sacrifices rule specificity. Although this method is common, more precision is possible.

Alternatively, regression analysis provides the researcher with the ability to develop a *score-based prediction rule* for the outcome of interest. Once the regression model is created, the variables' regression coefficients (not the odds ratios) are used to assign a "score" for each variable (Moons et al. 2002). The patient is given the appropriate "score" for each variable, and the overall sum is the patient's "score" for the outcome of interest. This final "score" categorizes the patient into a risk class for the outcome of interest, and evaluation/therapy is based on the risk category. A hypothetical example is presented in Table 16.2. The PORT score is commonly used and classifies adult patients with community-acquired pneumonia for 30-day mortality (Fine et al. 1997). Calculation of a patient's pneumonia score places the patient in one of five risk categories such that the clinician can determine appropriate location for management (home, ward, or intensive care).

Statistical validation of regression models. Bootstrapping is a statistical technique to validate regression models. It is not a replacement for validation in a separate cohort but is especially appropriate if a validation cohort is not available. Bootstrap validation randomly samples cases from the original dataset and creates new datasets of similar size as the original dataset. The original regression model is tested by performing multiple regression analyses on the "new" datasets (Altman and Andersen 1989; Chen and George 1985; Efron and Tibshirani 1991). The percentage of iterations in

Table 16.2 Hypothetical regression analysis calculating a score for abdominal injury requiring therapeutic laparotomy

Variable retaining significance	Regression coefficient	Score assigned to variable
Age < 18 years	-1.5	-1.5
Hypotension	2.93	3
Abdomen tender to palpation	1.22	1
Seat belt sign	3.55	3.5
Peritoneal irritation on palpation	7.14	7
Unexplained pneumoperitoneum on CT	8.90	9
Contrast extravasation on CT	6.38	6
Intraperitoneal fluid on >4 CT images	3.7	4
Hematocrit < 25%	2.8	3
	Laparotomy risk score	Rate of therapeutic laparotomy (%)
Class I	<2	1
Class II	2-4	5
Class III	5-8	20
Class IV	≥9	85

The patient's score for therapeutic laparotomy is calculated by summing the points for all the patient's variables. The patient with abdominal injury is then placed into a class based on the total score. The patient's risk for requiring therapeutic laparotomy is then determined and educated treatment decisions are determined. Patients in Class IV would likely be taken to the operating suite immediately, whereas those in Class III could be managed in the intensive care unit and those in Class I or Class II observed closely on the ward as their risk for a therapeutic laparotomy is low

which a variable is identified as an independent predictor in the new regression models is reported, and the variable is considered validated if identified in more than 50% of the bootstrap regression analyses (Chen and George 1985). This method utilizes the variability in the derivation sample to predict prediction rule performance in a new sample.

Binary recursive partitioning is a nonparametric multivariable analytic technique used to classify observations based on the risk for the outcome of interest, using a tree-like structure with decision “nodes” (Breiman et al. 1984). In binary recursive partitioning analysis, groups of patients, represented by a node in the decision tree, are split into two nodes, depending on risk stratification for the outcome. The most important variable for dividing the population into low- and high-risk groups is selected as the first node. This partitioning process is repeated until certain preselected tree-building criteria are met (subgroups are either too small to be further divided, homogeneous with regard to the outcome, or further subdivision no longer improves model accuracy). The tree enables users to visualize the hierarchical interaction of the variables. Thus, the clinician can determine the risk of the outcome for each step of the decision tree. An advantage of this approach is the exploration of interactions between predictor variables inherently in the analytic algorithm. This is in contrast to regression analysis, in which interaction terms must be created to explore these interactions. Therefore, recursive partitioning allows identification of predictor variables having differential relevance in different subgroups.

In the construction of decision trees, misclassification “costs” specific to misclassification errors are determined. Making specific misclassification errors more costly (e.g., the outcome is present but not predicted by the rule), results in a rule minimizing these mistakes at the expense of overclassifying patients without injury (i.e., specificity is sacrificed for improved sensitivity). These costs affect the growth and pruning of trees. The assigned value of the misclassification cost can vary widely depending on the clinical risk of missing the outcome. An outcome of “required neurosurgery for brain injury” would be assigned a higher misclassification cost than an outcome of “any brain injury.”

Table 16.3 Level of agreement for kappa ranges

Kappa value	Agreement
<0.0	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Failing to identify a case of brain injury requiring neurosurgery has much greater consequences than failing to diagnose a brain injury that requires no therapy.

Statistical validation of binary recursive partitioning models. The decision tree is routinely internally validated using tenfold cross-validation. This validation technique is performed by partitioning the data into ten strata, with each strata containing equal likelihood of the outcome. Ten different subanalyses are then performed, in which decision trees are derived from analysis of 90% of the data and tested on the remaining 10% of the data which was initially withheld. Different unique subsets of derivation and test data are used in each iteration. The average performance of these subanalyses is an estimate of how the tree derived from 100% of the data will perform on subsequent data samples and is used to determine the “best” tree.

Inter-rater reliability. For a prediction rule to be clinically useful, variable assessment must be reliably reproduced not only by the same clinician but also between different observers. Obviously, if different clinicians cannot reliably agree on the presence or absence of a particular variable, the variable has very limited utility in a prediction rule. The term for this assessment is *inter-rater* or *inter-observer reliability*. Variables with poor inter-rater reliability are not considered for inclusion in the prediction rule. Measuring inter-rater reliability is a critical but sometimes neglected step in the successful development of prediction rules. Inter-rater reliability data are routinely collected as part of the derivation rule study, and investigators either report the results in the prediction rules manuscript (Holmes et al. 2009a; Stiell et al. 2001a) or in a separate manuscript prior to publication of the prediction rule (Gorelick et al. 2008; Hollander et al. 2003). The latter option allows an in-depth presentation of the data and deeper insight into the issues related to inter-rater reliability.

Collecting inter-rater reliability data requires special considerations as substantial logistical issues must be overcome. Two physicians must document their responses to the same variables, independently of each other and within a narrow time window to ensure potential changes in patient history or physical examination results are minimized. Physical examination findings during the initial trauma examination may change significantly 6 h later, especially if the patient has received sedative or pain medications. Thus, the two physicians must, *independently*, document findings within a narrow timeframe to minimize errors occurring with time. If the secondary examination occurs more than 1 h after the initial examination, it would be excluded from analyses (Gorelick et al. 2008).

Inter-rater reliability results are not reported as the joint probability of agreement (number of times the raters agree/total number of ratings) as this method is too simplistic and fails to account for agreement due to chance. Instead, inter-rater reliability is appropriately reported using Cohen’s kappa statistic for categorical data (Cohen 1960) and intraclass correlation coefficient for interval data (Fleiss 1986). Level of agreement based on the kappa statistic is not intuitive and is provided in Table 16.3 (Landis and Koch 1977). Variables with accepted kappa measurements are considered for inclusion in the prediction rule, whereas those variables with unacceptable kappa measurement are not. Variables are acceptable if the kappa value is greater than 0.5 or 0.6 and have a lower bound of the 95% confidence interval greater than 0.4. Such a requirement assures at least moderate agreement for variables considered in the prediction rule.

Dichotomous variables (e.g., loss of consciousness) are reported with a simple kappa statistic, whereas ordinal variables (e.g., GCS score) are calculated with a weighted kappa (Cohen 1968). As the degrees of disagreement vary with ordinal data, a weighted kappa gives more credit the closer the rater's responses are to agreement. For example, two raters scoring a subject's mental status as GCS=13 and GCS=12 receive more "reliability credit" as compared to scores of GCS=13 and GCS=6. In neither instance do the raters perfectly agree but more penalty is applied the greater the disagreement. In cases of extremely rare variables (e.g., peritoneal irritation), the kappa statistic becomes limited due to the high likelihood of agreement by chance, and the calculated kappa results are biased against agreement. An alternative method, prevalence-adjusted bias-adjusted kappa (PABAK), exists for rare variables, but this estimate likely overestimates agreement and is not well accepted (Byrt et al. 1993).

Sample size. As with any rigorous scientific study, appropriate sample size is critical. Adequate sample size is necessary for prediction rule development to avoid a limited number of outcomes and over-fitting the data. Over-fitted data are an all-too-frequent problem in prediction rule generation.

The first rule of sample size calculation is collecting an appropriate number of subjects with the outcome of interest for the number of variables to be analyzed. As previously discussed, ten subjects with the outcome of interest are required for every predictor variable entered into a regression model (Concato et al. 1993; Harrell et al. 1985). If the investigators wish to enter 12 variables into the regression model, then 120 subjects with the outcome of interest are required. If the outcome of interest occurs in only 5% of eligible subjects, 2,400 subjects are required. Failure to include an appropriate number of subjects with the outcome of interest may result in either over-fitting or under-fitting a model and challenges the validity of the results. Such a rigorous sample size requirement (ten outcomes of interest for every variable considered in the model) is not necessary with binary recursive partitioning, but most investigators still follow this rule.

The second step in sample size calculation is determining an acceptable level of confidence interval precision around the rule's test characteristics. Investigators determine the acceptable lower bound of the 95% confidence interval for the rule's sensitivity or specificity. If the sensitivity of the rule is expected to be 100% but the investigators will tolerate a sensitivity of 98%, the sample size calculation is based on a 95% confidence interval with a lower bound of 98%. In this example, the study requires 149 subjects with the outcome of interest (anticipated sensitivity=100, 95% CI 98, 100%). Increasing the acceptable lower bound of the 95% confidence interval to 99% would increase the sample size to 300 subjects with the outcome of interest. Determining the acceptable lower bound of the 95% confidence interval is difficult but includes factoring in variables such as complications from the treatment/diagnosis and costs.

Finally, if preliminary data are available, more detailed sample size calculations are possible. Sample size calculations for regression modeling based on preliminary evidence are available (Flack and Eudey 1993; Hsieh 1989).

Similar issues arise when determining sample size calculations for validation studies. Validation of a regression model that results in a patient "score" requires enrolling 100 subjects with the outcome of interest (Vergouwe et al. 2005). Validation of prediction rules giving a simple binary outcome (yes/no disease), rely upon determining the acceptable lower bound of the 95% confidence interval as discussed above.

Standard formulas for sample size calculation for inter-rater reliability studies (kappa measurements) are available (Donner and Zou 2002; Fleiss 1981; Walter et al. 1998; Zou and Donner 2004). Basic principles for sample size calculation in a reliability study include (1) preliminary data are needed, (2) variables with lower kappa values require larger samples, and (3) uncommon/rare variables require larger samples. Table 16.4 presents an example of a sample size calculation for an inter-rater reliability study.

Table 16.4 Sample size calculation for kappa calculations

Data in table from pilot study (177 observations)	Observed kappa (assumed true)	Number of subjects with predictor present (out of 177)	Sample size needed to show $\kappa > 0.4$
GCS (<14 versus 14–15)	0.81	154	31
Distracting injury	0.56	47	166
Right costal margin tenderness	0.65	19	106
Left costal margin tenderness	0.64	20	112
Abdominal tenderness	0.62	72	73
Seat belt sign	0.58	13	293
Flank tenderness	0.44	20	4,583

The calculations above are only interested in testing whether the true kappa is >0.4 , not whether it is different from 0.4 (i.e., possibly smaller than 0.4). Therefore, the calculations are based on one-sided hypothesis testing with 80% power. As the observed kappa value gets lower and the variable becomes rarer, a larger sample size is necessary to demonstrate that the lower bound of the 95% confidence interval is >0.4 .

Validation of the Clinical Prediction Rules

Appropriate validation of prediction rules is necessary prior to implementation. Several derived prediction rules did not perform adequately in the validation phase and refinement or new derivation was required (Stiell et al. 1993; Holmes et al. 2009b; DeSmet et al. 1979). Reasons the derived rule may not be valid include over-fitting or under-fitting the model (Holmes et al. 2009b) or differences in the prevalence of disease in the validation population (Poses et al. 1986; Vergouwe et al. 2010b). Successful validation without refinement also occurs if the derivation sample is very large (Holmes et al. 2009a; Kuppermann et al. 2009; Stiell et al. 1996). Unfortunately, prediction rule validation encompasses many different techniques ranging from a statistical validation within the derivation cohort (discussed previously) to the gold standard of external validation in a separate cohort by “new” investigators.

Regardless of the technique, validation is performed with the same variables described in the derivation study. Attempting to validate a prediction rule by altering variable definitions is ill conceived and adds confusion to the clinical dilemma. Changing variables to more restrictive definitions increases rule specificity but decreases sensitivity, whereas changing to a broader variable definition increases rule sensitivity but sacrifices specificity. An attempt to validate a previously derived pediatric head CT prediction rule was made by changing several variables as follows: “any headache” was changed to “severe headache” and “any vomiting” was changed to “recurrent, projectile, or forceful vomiting” (Sun et al. 2007). The prediction rule sensitivity decreased from 99 to 90%, whereas specificity increased from 26 to 43% (Sun et al. 2007). Although very minor changes in variable terminology may have no impact on the performance of the prediction rule, substantial changes, as previously described, significantly alter the rule’s test performance and are to be avoided.

Types of Validation. Statistical validation encompasses statistical testing to determine model stability within the derivation dataset. Techniques include bootstrap validation (Steyerberg et al. 2003) with regression analyses and tenfold cross validation with binary recursive partitioning. These analyses are routinely performed in the derivation model and strengthen the validity of the derivation model. These statistical techniques do not, however, replace separate validation of the prediction rule in a new cohort as external validation is required (Bleeker et al. 2003).

Appropriate validation is performed in a separate cohort collected by one of two methods. Split sample validation is a method whereby the investigators collect a large sample and then split the sample into two cohorts (Haydel et al. 2000; Kuppermann et al. 2009; Holmes et al. 2009b). Typically, two-thirds of the initial sample is used to derive the rule and one-third is used to validate the rule, but the actual amount in each cohort is better determined by appropriate sample size calculations

(as described above). In these instances, the validation sample is collected before the actual rule is created from the derivation sample. Such a method has the advantage of continuous data collection at participating study sites such that “enrollment momentum” is maintained, but falls short of a gold standard validation.

The highest grade of prediction rule validation involves enrolling a separate sample of subjects *after the rule is derived*. The subject is determined to be positive or negative for the prediction rule and an assessment of the inter-rater reliability of the rule as a whole is also calculated. The investigators may provide the variables on the data collection form and collect data on each variable, but *the key step in the validation study is collecting “yes/no” for the prediction rule as a whole*. This method of validation and measurement of the inter-rater reliability of the rule can be performed only after the rule is derived. Furthermore, the validation is of highest quality if the investigators performing the validation are different from the investigators who completed the derivation, as potential bias is removed. In these instances, the investigators attempting to validate the prediction rule should contact the original investigators and discuss their plans prior to designing their study and enrolling any patients. In addition, the investigators should assess clinician acceptance of the rule during the validation.

Data collection in the validation phase follows steps similar to those in the derivation phase. Patient inclusion and exclusion criteria should be similar. Clinicians should be informed of the validation study and trained on the variables in the prediction rule (Stiell and Wells 1999). Failure to do such may result in failure to validate the rule (Kelly et al. 1994). In addition, eligible patients not enrolled must be assessed to demonstrate that the validation sample is representative of the targeted population.

Prediction rule accuracy (rule characteristics). Important prediction rule test characteristics include not only the sensitivity, specificity, positive and negative predictive values but also the positive and negative likelihood ratios. Although investigators often report the sensitivity and specificity in the derivation sample, determining the true characteristics of the prediction rule are reliant upon a validation sample, as the rule’s test characteristics in the derivation sample are likely overestimated.

The prediction rule’s sensitivity essentially “rules out” the outcome, whereas the specificity is associated with “ruling in” the outcome. The sensitivity provides the clinician with an estimate of the percentage of patients with the outcome of interest who are identified if the prediction rule is implemented. If identification of patients with the outcome of interest is critical, this is the most important test characteristic. The prediction rule’s specificity provides an estimate of the percentage of patients without disease who test negative when the rule is applied. This value provides an estimate of those who may benefit from test reduction if applied. If the prediction rule is designed to decrease inappropriate diagnostic testing (e.g., cranial CT scanning), and the rule is derived from a population undergoing CT scanning, the specificity provides an approximate reduction in CT scanning realized if the rule was implemented.

The rule’s positive and negative predictive value have less importance as they are highly dependent on the prevalence of the outcome in the population studied. The negative predictive value does, however, provide important information in those instances where the goal is to provide a prediction rule identifying a very low-risk population.

Calculating the prediction rule error rate. The error rate is the proportion of patients misclassified. The acceptable error rate varies by disease process and the severity resulting from the misclassification. As the risks of misclassification are often substantially different, calculating and reporting the actual error rate is not routinely performed. Miss-classifying a patient as positive for the outcome of needing a craniotomy is much different than miss-classifying a patient as negative for needing hospital admission. In one instance, the patient undergoes an unnecessary surgery and in the other instance the patient is discharged from the emergency department when he or she is better served by hospitalization. Assuming the patient is able to return to the hospital in case of deterioration, little

morbidity is likely to result from such misclassification. Investigators should be more focused on avoiding serious errors as opposed to the overall error rate. In designing the rule, the investigators must weigh the risks of misclassification and determine the acceptable rates of misclassification. Such determination guides the investigators in their analysis.

Implementation and Impact of the Rule

Unfortunately, the clinical impact of the prediction rule is rarely studied. Formal studies measuring the impact of the rule should be performed to ultimately determine the utility of the rule. These types of studies assess the rule's impact on resource utilization and include a cost effectiveness component. Determining the prediction rule's impact is the ultimate measure of a prediction rule.

Unless the prediction rule is successfully implemented and followed, the prediction rule has no use. Recently, investigators studied implementation of the Canadian cranial CT rule trial by randomizing sites to receive the prediction rule and measuring CT use before and after receiving this information. Unfortunately, this attempt at implementation failed to reduce CT scanning and perhaps *increased* CT utilization (Stiell et al. 2010).

Regrettably, the most appropriate method to distribute and implement prediction rules is currently unknown. Rapid technological advances have increased the use of electronic health records and electronic devices capable of providing clinicians with immediate access to previously derived and validated prediction rules. Decision-support web sites exist (MDcalc). The utility of these and their impact on care are unknown but likely are limited, as they require the clinician to "seek out" the information. Providing the clinician with the prediction rule at the time of decision making has the greatest impact on decision making and researchers should focus on this aspect. With increased reliance on electronic technology and the electronic health record, successful implementation of prediction rules into everyday use will likely involve designing validated prediction rules into the electronic health record.

References

- Altman, D. G., & Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8, 771–783.
- Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, 19, 453–473.
- Bandiera, G., Stiell, I. G., Wells, G. A., Clement, C., De Maio, V., Vandemheen, K. L., Greenberg, G. H., Lesiuk, H., Brison, R., Cass, D., Dreyer, J., Eisenhauer, M. A., Macphail, I., McKnight, R. D., Morrison, L., Reardon, M., Schull, M., & Worthington, J. (2003). The Canadian C-spine rule performs better than unstructured physician judgment. *Annals of Emergency Medicine*, 42, 395–402.
- Baxt, W. G. (1995). Application of artificial neural networks to clinical medicine. *The Lancet*, 346, 1135–1138.
- Blackmore, C. C. (2005). Clinical prediction rules in trauma imaging: Who, how, and why? *Radiology*, 235, 371–374.
- Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R., Derksen-Lubsen, G., Grobbee, D. E., & Moons, K. G. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, 56, 826–832.
- Bowman, S. M., Zimmerman, F. J., Christakis, D. A., Sharar, S. R., & Martin, D. P. (2005). Hospital characteristics associated with the management of pediatric splenic injuries. *JAMA*, 294, 2611–2617.
- Breiman, L., Friedman, J. H., Olshen, R. A., et al. (1984). *Classification and regression trees*. Washington, DC: Chapman & Hall.
- Brenner, D. J., & Hall, E. J. (2007). Computed tomography – an increasing source of radiation exposure. *The New England Journal of Medicine*, 357, 2277–2284.
- Birt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46, 423–429.

- Chen, C. H., & George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine*, 4, 39–46.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220.
- Concato, J., Feinstein, A. R., & Holford, T. R. (1993). The risk of determining risk with multivariable models. *Annals of Internal Medicine*, 118, 201–210.
- Culica, D., & Aday, L. A. (2008). Factors associated with hospital mortality in traumatic injuries: Incentive for trauma care integration. *Public Health*, 122, 285–296.
- DeSmet, A. A., Fryback, D. G., & Thornbury, J. R. (1979). A second look at the utility of radiographic skull examination for trauma. *AJR. American Journal of Roentgenology*, 132, 95–99.
- Donders, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59, 1087–1091.
- Donner, A., & Zou, G. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics*, 58, 209–215.
- Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science*, 253, 390–395.
- Feinstein, A. R. (1977). Clinical biostatistics. XLI. Hard science, soft data, and the challenges of choosing clinical variables in research. *Clinical Pharmacology and Therapeutics*, 22, 485–498.
- Feinstein, A. R. (1987). *Clinimetrics*. New Haven, CT: Yale University Press.
- Fine, M. J., Auble, T. E., Yealy, D. M., Hanusa, B. H., Weissfeld, L. A., Singer, D. E., Coley, C. M., Marrie, T. J., & Kapoor, W. N. (1997). A prediction rule to identify low-risk patients with community-acquired pneumonia. *The New England Journal of Medicine*, 336, 243–250.
- Flack, V. F., & Eudey, T. L. (1993). Sample size determinations using logistic regression with pilot data. *Statistics in Medicine*, 12, 1079–1084.
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: John Wiley and Sons.
- Fleiss, J. L. (1986). The design and analysis of chemical experiments. In V. Barnett, R. A. Bradley, & J. S. Hunter (Eds.), *Reliability of measurement* (pp. 1–31). New York: John Wiley & Sons.
- Garber, B. G., Bigelow, E., Yelle, J. D., & Pagliarello, G. (2000). Use of abdominal computed tomography in blunt trauma: Do we scan too much? *Canadian Journal of Surgery*, 43, 16–21.
- Glance, L. G., Osler, T. M., Dick, A. W., Mukamel, D. B., & Meredith, W. (2010). The Survival Measurement and Reporting Trial for Trauma (SMARTT): Background and study design. *The Journal of Trauma*, 68, 1491–1497.
- Gorelick, M. H., Atabaki, S. M., Hoyle, J., Dayan, P. S., Holmes, J. F., Holubkov, R., Monroe, D., Callahan, J. M., & Kuppermann, N. (2008). Interobserver agreement in assessment of clinical variables in children with blunt head trauma. *Academic Emergency Medicine*, 15, 812–818.
- Graham, I. D., Stiell, I. G., Laupacis, A., O'Connor, A. M., & Wells, G. A. (1998). Emergency physicians' attitudes toward and use of clinical decision rules for radiography. *Academic Emergency Medicine*, 5, 134–140.
- Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, 142, 1255–1264.
- Guldner, G., Babbitt, J., Boulton, M., O'Callaghan, T., Feleke, R., & Hargrove, J. (2004). Deferral of the rectal examination in blunt trauma patients: A clinical decision rule. *Academic Emergency Medicine*, 11, 635–641.
- Guyatt, G., Walter, S., Shannon, H., Cook, D., Jaeschke, R., & Heddle, N. (1995). Basic statistics for clinicians: 4. Correlation and regression. *CMAJ*, 152, 497–504.
- Harrell, F. E., Jr., Lee, K. L., Matchar, D. B., & Reichert, T. A. (1985). Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69, 1071–1077.
- Haydel, M. J., Preston, C. A., Mills, T. J., Luber, S., Blaudeau, E., & DeBlieux, P. M. (2000). Indications for computed tomography in patients with minor head injury. *The New England Journal of Medicine*, 343, 100–105.
- Hoffman, J. R., Mower, W. R., Wolfson, A. B., Todd, K. H., & Zucker, M. I. (2000). Validity of a set of clinical criteria to rule out injury to the cervical spine in patients with blunt trauma. National Emergency X-Radiography Utilization Study Group. *The New England Journal of Medicine*, 343, 94–99.
- Hollander, J. E., Go, S., Lowery, D. W., Wolfson, A. B., Pollack, C. V., Herbert, M., Mower, W. R., & Hoffman, J. R. (2003). Interrater reliability of criteria used in assessing blunt head injury patients for intracranial injuries. *Academic Emergency Medicine*, 10, 830–835.
- Holmes, J. F., Sokolove, P. E., Land, C., & Kuppermann, N. (1999). Identification of intra-abdominal injuries in children hospitalized following blunt torso trauma. *Academic Emergency Medicine*, 6, 799–806.
- Holmes, J. F., Sokolove, P. E., Brant, W. E., Palchak, M. J., Vance, C. W., Owings, J. T., & Kuppermann, N. (2002a). Identification of children with intra-abdominal injuries after blunt trauma. *Annals of Emergency Medicine*, 39, 500–509.
- Holmes, J. F., Sokolove, P. E., Brant, W. E., & Kuppermann, N. (2002b). A clinical decision rule for identifying children with thoracic injuries after blunt torso trauma. *Annals of Emergency Medicine*, 39, 492–499.

- Holmes, J. F., Wisner, D. H., McGahan, J. P., Mower, W. R., & Kuppermann, N. (2009a). Clinical prediction rules for identifying adults at very low risk for intra-abdominal injuries after blunt trauma. *Annals of Emergency Medicine*, *54*, 575–584.
- Holmes, J. F., Mao, A., Awasthi, S., McGahan, J. P., Wisner, D. H., & Kuppermann, N. (2009b). Validation of a prediction rule for the identification of children with intra-abdominal injuries after blunt torso trauma. *Annals of Emergency Medicine*, *54*, 528–533.
- Hoyt, D. B., Hollingsworth-Fridlund, P., Winchell, R. J., Simons, R. K., Holbrook, T., & Fortlage, D. (1994). Analysis of recurrent process errors leading to provider-related complications on an organized trauma service: Directions for care improvement. *The Journal of Trauma*, *36*, 377–384.
- Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, *8*, 795–802.
- Hunink, M. G. (2005). Decision making in the face of uncertainty and resource constraints: Examples from trauma imaging. *Radiology*, *235*, 375–383.
- Janssen, K. J., Donders, A. R., Harrell, F. E., Jr., Vergouwe, Y., Chen, Q., Grobbee, D. E., & Moons, K. G. (2010). Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, *63*, 721–727.
- Kelly, A. M., Richards, D., Kerr, L., Grant, J., O'Donovan, P., Basire, K., & Graham, R. (1994). Failed validation of a clinical decision rule for the use of radiography in acute ankle injury. *The New Zealand Medical Journal*, *107*, 294–295.
- Klassen, T. P., Reed, M. H., Stiell, I. G., Nijssen-Jordan, C., Tenenbein, M., Joubert, G., Jarvis, A., Baldwin, G., St-Vil, D., Pitters, C., Belanger, F., McConnell, D., Vandemheen, K., Hamilton, M. G., Sutcliffe, T., & Colbourne, M. (2000). Variation in utilization of computed tomography scanning for the investigation of minor head trauma in children: A Canadian experience. *Academic Emergency Medicine*, *7*, 739–744.
- Kuppermann, N., Holmes, J. F., Dayan, P. S., Hoyle, J. D., Jr., Atabaki, S. M., Holubkov, R., Nadel, F. M., Monroe, D., Stanley, R. M., Borgialli, D. A., Badawy, M. K., Schunk, J. E., Quayle, K. S., Mahajan, P., Lichenstein, R., Lillis, K. A., Tunik, M. G., Jacobs, E. S., Callahan, J. M., Gorelick, M. H., Glass, T. F., Lee, L. K., Bachman, M. C., Cooper, A., Powell, E. C., Gerardi, M. J., Melville, K. A., Muizelaar, J. P., Wisner, D. H., Zuspan, S. J., Dean, J. M., & Wootton-Gorges, S. L. (2009). Identification of children at very low risk of clinically-important brain injuries after head trauma: A prospective cohort study. *The Lancet*, *374*, 1160–1170.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.
- Langdorf, M. I., Rudkin, S. E., Dellota, K., Fox, J. C., & Munden, S. (2002). Decision rule and utility of routine urine toxicology screening of trauma patients. *European Journal of Emergency Medicine*, *9*, 115–121.
- Laupacis, A., Sekar, N., & Stiell, I. G. (1997). Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA*, *277*, 488–494.
- Little, R. J. (1988). Some statistical analysis issues at the World Fertility Survey. *The American Statistician*, *42*, 31–36.
- McGinn, T. G., Guyatt, G. H., Wyer, P. C., Naylor, C. D., Stiell, I. G., & Richardson, W. S. (2000). Users' guides to the medical literature: XXII: How to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA*, *284*, 79–84.
- MD + calc. <http://www.mdcalc.com/>.
- Minai, J. P., Schmicker, R. H., Kerby, J. D., Stiell, I. G., Schreiber, M. A., Bulger, E., Tisherman, S., Hoyt, D. B., & Nichol, G. (2010). Severe traumatic injury: Regional variation in incidence and outcome. *Annals of Surgery*, *252*, 149–157.
- Moons, K. G., Harrell, F. E., & Steyerberg, E. W. (2002). Should scoring rules be based on odds ratios or regression coefficients? *Journal of Clinical Epidemiology*, *55*, 1054–1055.
- Mower, W. R., Hoffman, J. R., Herbert, M., Wolfson, A. B., Pollack, C. V., Jr., & Zucker, M. I. (2005). Developing a decision instrument to guide computed tomographic imaging of blunt head injury patients. *The Journal of Trauma*, *59*, 954–959.
- Newgard, C. D., Hedges, J. R., Stone, J. V., Lenfesty, B., Diggs, B., Arthur, M., & Mullins, R. J. (2005a). Derivation of a clinical decision rule to guide the interhospital transfer of patients with blunt traumatic brain injury. *Emergency Medicine Journal*, *22*, 855–860.
- Newgard, C. D., Hui, S. H., Griffin, A., Wuerstle, M., Pratt, F., & Lewis, R. J. (2005b). Prospective validation of an out-of-hospital decision rule to identify seriously injured children involved in motor vehicle crashes. *Academic Emergency Medicine*, *12*, 679–687.
- Nishijima, D. K., Sena, M. J., & Holmes, J. F. (2010). Identification of low-risk patients with traumatic brain injury and intracranial hemorrhage who do not need intensive care unit admission. *J Trauma*, *70*(6), E101–E107.
- Nuss, K. E., Dietrich, A. M., & Smith, G. A. (2001). Effectiveness of a pediatric trauma team protocol. *Pediatric Emergency Care*, *17*, 96–100.
- Palchak, M. J., Holmes, J. F., Vance, C. W., Gelber, R. E., Schauer, B. A., Harrison, M. J., Willis-Shore, J., Wootton-Gorges, S. L., Derlet, R. W., & Kuppermann, N. (2003). A decision rule for identifying children at low risk for brain injuries after blunt head trauma. *Annals of Emergency Medicine*, *42*, 492–506.

- Palmer, S., Bader, M. K., Qureshi, A., Palmer, J., Shaver, T., Borzatta, M., & Stalcup, C. (2001). The impact on outcomes in a community hospital setting of using the AANS traumatic brain injury guidelines. *American Association for Neurologic Surgeons. The Journal of Trauma, 50*, 657–664.
- Perry, J. J., & Stiell, I. G. (2006). Impact of clinical decision rules on clinical care of traumatic injuries to the foot and ankle, knee, cervical spine, and head. *Injury, 37*, 1157–1165.
- Poses, R. M., Cebul, R. D., Collins, M., & Fager, S. S. (1986). The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. *Annals of Internal Medicine, 105*, 586–591.
- Rodriguez, R. M., Hendey, G. W., Mower, W., Kea, B., Fortman, J., Merchant, G., & Hoffman, J. R. (2011). Derivation of a decision instrument for selective chest radiography in blunt trauma. *J Trauma, 71*(3), 549–553.
- Rose, J. S., Richards, J. R., Battistella, F., Bair, A. E., McGahan, J. P., & Kuppermann, N. (2005). The fast is positive, now what? Derivation of a clinical decision rule to determine the need for therapeutic laparotomy in adults with blunt torso trauma and a positive trauma ultrasound. *The Journal of Emergency Medicine, 29*, 15–21.
- Rudy, T. E., Kubinski, J. A., & Boston, J. R. (1992). Multivariate analysis and repeated measurements: A premier. *Journal of Critical Care, 7*, 30–41.
- Sariego, J. (2000). Impact of a formal trauma program on a small rural hospital in Mississippi. *Southern Medical Journal, 93*, 182–185.
- Sokolove, P. E., Kuppermann, N., & Holmes, J. F. (2005). Association between the “seat belt sign” and intra-abdominal injury in children with blunt torso trauma. *Academic Emergency Medicine, 12*, 808–813.
- Steele, R., Green, S. M., Gill, M., Coba, V., & Oh, B. (2006). Clinical decision rules for secondary trauma triage: Predictors of emergency operative management. *Annals of Emergency Medicine, 47*, 135.
- Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., & Moons, K. G. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology, 56*, 441–447.
- Stiell, I. G., & Wells, G. A. (1999). Methodologic standards for the development of clinical decision rules in emergency medicine. *Annals of Emergency Medicine, 33*, 437–447.
- Stiell, I. G., Greenberg, G. H., McKnight, R. D., Nair, R. C., McDowell, I., Reardon, M., Stewart, J. P., & Maloney, J. (1993). Decision rules for the use of radiography in acute ankle injuries. Refinement and prospective validation. *JAMA, 269*, 1127–1132.
- Stiell, I. G., Greenberg, G. H., Wells, G. A., McKnight, R. D., Cwinn, A. A., Cacciotti, T., McDowell, I., & Smith, N. A. (1995a). Derivation of a decision rule for the use of radiography in acute knee injuries. *Annals of Emergency Medicine, 26*, 405–413.
- Stiell, I. G., Wells, G. A., McDowell, I., Greenberg, G. H., McKnight, R. D., Cwinn, A. A., Quinn, J. V., & Yeats, A. (1995b). Use of radiography in acute knee injuries: Need for clinical decision rules. *Academic Emergency Medicine, 2*, 966–973.
- Stiell, I. G., Greenberg, G. H., Wells, G. A., McDowell, I., Cwinn, A. A., Smith, N. A., Cacciotti, T. F., & Sivilotti, M. L. (1996). Prospective validation of a decision rule for the use of radiography in acute knee injuries. *JAMA, 275*, 611–615.
- Stiell, I. G., Wells, G. A., Vandemheen, K., Laupacis, A., Brison, R., Eisenhauer, M. A., Greenberg, G. H., MacPhail, I., McKnight, R. D., Reardon, M., Verbeek, R., Worthington, J., & Lesiuk, H. (1997). Variation in emergency department use of cervical spine radiography for alert, stable trauma patients. *CMAJ, 156*, 1537–1544.
- Stiell, I. G., Wells, G. A., Vandemheen, K., Clement, C., Lesluk, H., Laupacis, A., McKnight, R. D., Verbeek, R., Brison, R., Cass, D., Eisenhauer, M. A., Greenberg, G. H., & Worthington, J. (2001a). The Canadian CT head rule for patients with minor head injury. *The Lancet, 357*, 1391–1396.
- Stiell, I. G., Wells, G. A., Vandemheen, K. L., Clement, C. M., Lesiuk, H., De Maio, V. J., Laupacis, A., Schull, M., McKnight, R. D., Verbeek, R., Brison, R., Cass, D., Dreyer, J., Eisenhauer, M. A., Greenberg, G. H., MacPhail, I., Morrison, L., Reardon, M., & Worthington, J. (2001b). The Canadian C-spine rule for radiography in alert and stable trauma patients. *JAMA, 286*, 1841–1848.
- Stiell, I. G., Clement, C. M., Grimshaw, J. M., Brison, R. J., Rowe, B. H., Lee, J. S., Shah, A., Brehaut, J., Holroyd, B. R., Schull, M. J., McKnight, R. D., Eisenhauer, M. A., Dreyer, J., Letovsky, E., Rutledge, T., Macphail, I., Ross, S., Perry, J. J., Ip, U., Lesiuk, H., Bennett, C., & Wells, G. A. (2010). A prospective cluster-randomized trial to implement the Canadian CT Head Rule in emergency departments. *CMAJ, 182*, 1527–1532.
- Sun, B. C., Hoffman, J. R., & Mower, W. R. (2007). Evaluation of a modified prediction instrument to identify significant pediatric intracranial injury after blunt head trauma. *Annals of Emergency Medicine, 49*, 325–332. 332.e1.
- Tinkoff, G. H., O’Connor, R. E., & Fulda, G. J. (1996). Impact of a two-tiered trauma response in the emergency department: Promoting efficient resource utilization. *The Journal of Trauma, 41*, 735–740.
- Ungar, T. C., Wolf, S. J., Haukoos, J. S., Dyer, D. S., & Moore, E. E. (2006). Derivation of a clinical decision rule to exclude thoracic aortic imaging in patients with blunt chest trauma after motor vehicle collisions. *The Journal of Trauma, 61*, 1150–1155.
- van der Heijden, G. J., Donders, A. R., Stijnen, T., & Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology, 59*, 1102–1109.

- Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J., & Habbema, J. D. (2005). Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology*, *58*, 475–483.
- Vergouwe, Y., Royston, P., Moons, K. G., & Altman, D. G. (2010a). Development and validation of a prediction model with missing predictor data: A practical approach. *Journal of Clinical Epidemiology*, *63*, 205–214.
- Vergouwe, Y., Moons, K. G., & Steyerberg, E. W. (2010b). External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology*, *172*, 971–980.
- Walter, S. D., Eliasziw, M., & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, *17*, 101–110.
- Wasson, J. H., Sox, H. C., Neff, R. K., & Goldman, L. (1985). Clinical prediction rules. Applications and methodological standards. *The New England Journal of Medicine*, *313*, 793–799.
- Zou, G., & Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics*, *60*, 807–811.

Chapter 17

Biomarkers of Traumatic Injury

Cameron B. Jeter, John B. Redell, Anthony N. Moore, Georgene W. Hergenroeder, Jing Zhao, Daniel R. Johnson, Michael J. Hylin, and Pramod K. Dash

Introduction

Traumatic injuries claim more lives in those under the age of 45 than any other cause. Throughout the lifespan, traumatic injuries are surpassed in mortality only by cancer and atherosclerosis (Todd 2004). A traumatic injury due to mechanical energy can result from the shearing forces caused by a sudden deceleration (e.g., vehicular accidents), from impacts (e.g., falls, blunt force trauma), rapid changes in pressure (e.g., blast exposure), or penetration by objects (e.g., knives) or high-velocity projectiles (e.g., bullets, shrapnel). An individual patient can present with a combination of injuries to a single body organ (e.g., both shearing and impact injuries from a motor vehicle accident or an acute-on-chronic hematoma in an elderly patient who has repeated falls) or multiple injuries simultaneously to different organ systems (multi-trauma or polytrauma). Identifying the mechanism of injury assists in directing the clinical assessment; augmenting clinical assessment with sensitive and specific biomarkers that can be rapidly assessed provides additional tools to direct interventions.

A biomarker is a physiological indicator of a biological or disease state that can be used, for example, to diagnose a condition, monitor disease progression, or evaluate treatment efficacy. While clinical measurements, such as blood pressure and body temperature, and results from diagnostic imaging can be considered biomarkers, the term is generally reserved for the measure of biological molecules in samples taken from either the affected tissue or body fluids. Significant injuries to internal organs can occur in the absence of overt external signs of physical injury and may not initially present with obvious clinical signs or symptoms. While computed tomography (CT) and magnetic resonance imagery (MRI) are standards in the diagnosis of traumatic injury, augmenting imaging tests with biomarkers provides multiple advantages. First, many biomarkers are found in bodily fluids that can be accessed using minimally invasive techniques. The normal ranges of these molecules are typically well defined, with values outside this range indicative of abnormality.

C.B. Jeter, PhD (✉) • J.B. Redell, PhD • A.N. Moore, BS • J. Zhao, MD, PhD
• D.R. Johnson, PhD • M.J. Hylin, PhD • P.K. Dash, PhD

Department of Neurobiology & Anatomy, The University of Texas Medical School at Houston,
6431 Fannin Street, Houston, TX 77030, USA

e-mail: Cameron.B.Jeter@uth.tmc.edu; John.B.Redell@uth.tmc.edu; Anthony.N.Moore@uth.tmc.edu;

Jing.Zhao@uth.tmc.edu; Daniel.Johnson@uth.tmc.edu; Michael.J.Hylin@uth.tmc.edu; P. Dash@uth.tmc.edu

G.W. Hergenroeder, RN, MHA

The Vivian L. Smith Department of Neurosurgery, The University of Texas Medical School at Houston,
6431 Fannin Street, Houston, TX 77030, USA

e-mail: Georgene.W.Hergenroeder@uth.tmc.edu

Second, the high costs associated with CT and MRI instrumentation and the specialized technical expertise of the personnel necessary to provide these services are not within reach of all clinics. In many cases, biomarkers can provide an objective measurement at a relatively low cost, with the results used as part of the decision regarding whether more invasive or costly measures are warranted. Third, prognostic biomarkers can be used to identify a patient's likely clinical progression or long-term outcome. However, even though valuable information can be obtained from clinically useful biomarkers, they are typically not used in isolation. Rather, optimal utility is obtained when biomarkers are used in conjunction with multifarious clinical data, including imaging results.

The methodologies used in initial biomarker discovery can be classified as either biased or unbiased. A biased approach (also called "top-down" approach) is a guided search aided by an hypothesis or prior knowledge of the disease process. For instance, it is known that after traumatic brain injury (TBI), the cell death in the core of the primary injury is largely necrotic in nature. Because necrosis is associated with calpain activation, specific protein degradation profiles will be generated (both in terms of resultant peptide fragments and the time course of their appearance), giving rise to putative biomarkers of necrosis. Ideally, these biomarkers would be distinct from biomarkers generated as a result of apoptotic cell death, because apoptosis is associated with the activation of caspases and is typically more delayed in progression. The biased approach is often executed in the injured tissue first, then resultant candidate biomarkers are examined in more easily accessible sample sources. An unbiased approach (also called "bottom-up" approach) examines changes in the composition of a tissue or body fluid, and then attempts to establish links between these changes and pathology. It should be noted, however, that a biomarker is a surrogate that assists in the identification/classification of a condition, and does not necessarily have to be directly linked to the condition's pathology to be clinically informative. Although an unbiased approach using an easily accessible sample source, such as plasma, may overcome the potential problem of peripheral biomarker detectability, this approach often identifies general markers that may not be unique to the condition of interest. Thus, whereas a plasma biomarker may be capable of distinguishing traumatic injury patients from healthy volunteers (i.e., is a sensitive marker of injury), it may be a general indicator of injury and therefore less likely to distinguish among different types of traumatic injury (i.e., not a specific marker of brain injury).

A challenge when using blood as the source material is its extraordinary protein concentration range: the 12 most abundant proteins comprise >95% of its total protein content (~60–80 mg/ml) and can mask detection of low abundant proteins that can be present at concentrations 9–10 orders of magnitude lower (pg/ml). Further complicating detection and accurate quantification, many circulating proteins are often modified (e.g., glycosylated) or bound to carrier proteins. These complications often require that blood be fractionated prior to its use for biomarker discovery.

Advances in high throughput screening, improved detection sensitivity, and increased precision in measurements have all contributed to the recent large increase in both the ability and capacity of researchers to identify putative candidate biomarkers. However, identifying potential candidates is just the first step in a long process involving the validation and development of clinically useful diagnostic biomarkers. Once a candidate biomarker has been identified, it then needs to be validated in subsequent experiments. Biomarker validation can be viewed as having at least two major components (1) validating the detection assay itself and (2) validating that the identified candidate is accurate and predictive for the condition under study. These validations are critical for determining the diagnostic accuracy of an identified biomarker. Two of the most commonly reported diagnostic accuracy measures are sensitivity and specificity, which are also used in the calculation of several other useful metrics. Sensitivity is a measure of the probability of a positive test to identify the condition when it is actually present. (i.e., it does not miss a diagnosis). Specificity is defined as the probability that a negative test will correctly identify healthy individuals (i.e., it does not falsely diagnose patients). Sensitivity and specificity are not affected by the condition's prevalence within the study population, but can be highly dependent on the condition's spectrum. Thus, while a test may have high sensitivity and specificity for diagnosing severe trauma, for example, it may not have

adequate sensitivity to detect mild trauma. As a single biomarker may not have the desired level of sensitivity and specificity for diagnostic purposes, a combination of biomarkers, often termed a “biomarker signature,” can be generated to improve diagnostic accuracy.

Current and Prospective Biomarkers of Traumatic Injury

Brain

Although the brain is protected by the neocranium of the skull and is functionally isolated from most disease-causing agents by the blood–brain barrier (BBB) and blood–cerebrospinal fluid (CSF) barrier, its large size and relatively fragile composition make it highly vulnerable to trauma. According to the Centers for Disease Control and Prevention, a head injury occurs every 15 s yielding over two million new brain trauma cases each year. Most commonly, trauma to the brain occurs as either the result of a closed head injury (blunt force trauma), penetration of the skull by a projectile, or, more recently, as the result of exposure to blast overpressure. Depending on the severity and location of the injury, non-fatal brain trauma can result in deficits ranging from difficulties in fine motor control to lasting impairments in learning and memory and personality change.

Neurons and glia, two of the major cellular components of the brain, can be distinguished not only by their functional properties but also by the unique proteins they express. Damage to these cells as a result of trauma can cause release of these proteins into the CSF and/or blood circulation where they can be detected and used as biomarkers of brain injury. Although none of the markers described below is currently approved for the diagnosis of brain trauma or evaluation of treatment, numerous clinical studies have tested the diagnostic accuracy of many of these putative TBI biomarkers.

Protein Biomarkers of TBI

Neuron-specific enolase (NSE). NSE is one of five isozymes of the glycolytic enzyme enolase, and it has a normal serum level of less than 12.5 ng/ml. Elevations in serum levels of NSE above 21.7 ng/ml have been correlated to injury severity and poor outcome in human TBI patients with a good sensitivity, although it is only moderately specific. Serum NSE levels have also been correlated with performance in neuropsychological exams (sensitivity 55% and specificity 78%) and have been indicated to predict intracranial lesions (sensitivity 77% and specificity 52%) (Fridriksson et al. 2000; Herrmann et al. 2001; Vos et al. 2004). However, its poor specificity hampers its use as a diagnostic test of brain trauma, likely due to the expression of NSE by other organs such as the lung and gut and by non-neuronal cells such as thrombocytes and erythrocytes.

Cleaved tau (C-Tau). Tau is a microtubule-associated protein that is enriched in the axons of neurons. Brain trauma often results in the proteolysis of tau, producing a cleaved product called C-tau. In patients with severe brain trauma, initial post-injury CSF C-tau levels correlate with clinical outcome (sensitivity 92% and specificity 94%) and may also predict the occurrence of elevated intracranial pressure (Zemlan et al. 2002). However, in patients with mild brain injury, C-tau levels have been shown to be poor predictors of post-concussion syndrome.

Glial fibrillary acidic protein (GFAP). GFAP is a monomeric intermediate filament protein expressed by astrocytes that is released after TBI. Elevated serum GFAP levels following severe TBI are predictive of poorer outcome, and serum GFAP levels correlate with intracranial pressure, mean arterial pressure, cerebral perfusion pressure, Glasgow Outcome Score, and mortality (Nylen et al. 2006; Pelinka et al. 2004). GFAP serum concentrations above 1.5 ng/ml are predictive of death (85% sensitivity and 52% specificity) or poor neurological outcome (80% sensitivity and 59% specificity) (Vos et al. 2004).

S100 β . S100 β is a low molecular weight (10.5 kDa) calcium-binding protein that is primarily expressed and secreted by astrocytes. S100 β is typically found in very low levels in the CSF and serum, and normal levels of this protein have been strongly correlated with the absence of intracranial injury (Uden and Romner 2010). However, after severe brain injury, the serum levels of S100 β increase, with concentrations over 1.13 ng/ml associated with increased mortality and morbidity (Vos et al. 2004). Since S100 β does not cross the intact BBB, its serum levels are also thought to reflect the degree of BBB disruption.

Ubiquitin carboxyl-terminal esterase-L1 (UCHL1). UCHL1 is a small, approximately 25 kDa cysteine protease that hydrolyzes the bond at the C-terminal between ubiquitin and small adducts or unfolded polypeptides (Setsuie and Wada 2007). This enzyme comprises approximately 1–2% of total soluble protein in the brain where it is expressed exclusively in neurons, with very low levels also expressed in some neuroendocrine cells. Mutations in *UCHL1* may be associated with Parkinson's disease and other neurodegenerative disorders (Belin and Westerlund 2008; Setsuie and Wada 2007). Recently, elevated levels of CSF and serum UCHL1 were found to be correlated with poor outcome in severe TBI patients (Brophy et al. 2011; Papa et al. 2010).

Myelin basic protein (MBP). MBP is the major protein component of myelin expressed by oligodendrocytes. Shearing of brain white matter leading to diffuse axonal injury results in the release of MBP into the CSF and serum, where it has been found to remain elevated for up to 2 weeks post-injury (Kochanek et al. 2008). Interestingly, MBP can cause opening of the BBB, thereby facilitating its own entry (and possibly other central nervous system (CNS)-derived biomarkers) into the circulation. A MBP test is often used to assess the levels of this protein in the CSF in neurological conditions where demyelination is suspected. Normally, MBP levels are less than 4 ng/ml in the CSF, with levels exceeding 9 ng/ml indicative of active myelin degradation.

In addition to the predominately CNS-derived biomarkers above, other circulating proteins have been evaluated for their ability to diagnose TBI and predict outcome. As TBI, like most bodily injuries, is associated with acute inflammation, several cytokines, acute phase reactant proteins, and chemokines have been examined for their diagnostic and prognostic properties.

Interleukin-1 (IL-1). IL-1 is a cytokine that when bound to the IL-1 receptor regulates several physiological, metabolic, and hematopoietic activities. In particular, IL-1 β has a well-described history of activating the immune system in response to infection and injury. Following TBI, IL-1 β has been proposed to play a predominate role in the development of astroglial scars (Giulian and Lachman 1985), a pathology thought to contribute to the prevention of axonal extension and poor functional recovery. Consistent with this, high-CSF levels of IL-1 β have been demonstrated to be predictive of poor outcome at 3 months following TBI (Chiaretti et al. 2005; Singhal et al. 2002).

Interleukin-6 (IL-6). IL-6 is an important mediator of fever and the acute phase response following injury. A number of studies, both experimental and clinical, have reported that IL-6 levels are dramatically increased in the CSF and serum following TBI. Interestingly, these studies suggest that while high initial serum levels of IL-6 are associated with poor outcome (Venetsanou et al. 2007) and secondary pathologies such as elevated intracranial pressure (Hergenroeder et al. 2010), parenchymal and CSF IL-6 levels have been correlated with enhanced survival and improved outcome (Chiaretti et al. 2008; Winter et al. 2004).

Transforming growth factor beta (TGF- β). TGF- β is a member of a large superfamily of growth factors that is important for a number of biological functions including cell growth and differentiation, angiogenesis, immune function, extracellular matrix production, cell chemotaxis, and apoptosis. Following TBI, the CSF levels of TGF- β peak by 24-h post-injury and remain elevated for 3 weeks (Morganti-Kossmann et al. 1999).

Acute phase proteins. Acute phase proteins are induced by the liver in response to elevated serum levels of proinflammatory cytokines such as IL-1 and IL-6. Using an unbiased screen, it was found that the serum levels of C-reactive protein and serum amyloid A are rapidly increased after severe TBI, and remain elevated for at least 5 days after the injury (Hergenroeder et al. 2008). Significantly, a large percentage of patients had significant elevations in these proteins (up to 300-fold) as early as 10-h post-injury, making them highly sensitive indicators of brain injury. However, similar to IL-1 and IL-6, their diagnostic value for TBI may be limited if the patient has suffered injury to other organs. Ceruloplasmin, also known as ferroxidase or iron(II):oxygen oxidoreductase, is the major copper carrier protein in the blood and plays an important role in copper and iron metabolism. As an acute phase protein, the serum level of ceruloplasmin has been reported to be increased in response to injury and infection, where it is thought to play a protective role due to its antioxidant properties (Goldstein et al. 1982; Rebhun et al. 1991; Tomas and Toparceanu 1986). Consistent with this protective role, it has been found that serum ceruloplasmin levels are significantly reduced within the first 24 h after injury in TBI patients who subsequently develop elevated intracranial pressure (Dash et al. 2010).

Chemokines. Inflammatory chemokines are small proteins secreted by cells in response to inflammatory cytokines such as IL-1. These proteins function primarily as chemoattractants, recruiting peripheral immune cells (leukocytes, monocytes, and neutrophils) into the injured brain. For example, chemokine CC ligand-2 (CCL2, also known as monocyte chemoattractant protein-1) is important for the recruitment of macrophages into the brain following injury or infection. This is thought to be due to its influence not only as a chemoattractant, but also its ability to directly enhance BBB permeability (Song and Pachter 2004). Sustained elevations of CCL2 have been found in the CSF of severe TBI patients for up to 10 days following trauma (Semple et al. 2010). When mice lacking the CCL2 gene were subjected to experimental brain trauma, it was found that these animals had reduced lesion size and attenuated macrophage accumulation, suggesting that this chemokine may play a critical role in exacerbating brain damage (Semple et al. 2010). However, although high serum levels of CCL2 have been associated with reduced survival following severe TBI, they are not predictive of contusion expansion (Rhodes et al. 2009). In addition to CCL2, it has been reported that CXCL8 (chemokine (C-X-C motif) ligand 8; also known as IL-8) has been found to be elevated in the CSF of patients after severe TBI and correlates with severe BBB dysfunction and increased mortality (Kossmann et al. 1997; Whalen et al. 2000).

Metabolites as TBI Biomarkers

Metabolites of lipids, neurotransmitters, and glycolytic intermediates have shown potential as biomarkers for TBI.

F2-Isoprostane. F2-isoprostanes are produced from arachidonic acid as a result of peroxidation and have been found to be increased in both the CSF and the serum of TBI patients. In children with severe TBI, the levels of F2-isoprostane are elevated by sixfold over that observed in uninjured controls (Bayir et al. 2009). In adults, males have been found to have approximately twice the levels of CSF F2-isoprostane of females (Bayir et al. 2004). This disparity in lipid peroxidation has been proposed to be a potential reason for the greater neuroprotection seen in females than in males.

4-Hydroxynonenal (4-HNE). 4-HNE, an endogenous end product of lipid peroxidation, is another well-accepted lipid peroxidation marker. Elevated 4-HNE level has been observed in both cortical and hippocampal tissues early after TBI in animal models (Hall et al. 2004; Singh et al. 2006) and reduced 4-HNE level has been found in parallel with neuroprotective effects in animals treated with antioxidants (Singh et al. 2007). Although CSF and plasma 4-HNE levels have not been examined following TBI in humans, its levels have been reported to increase in other neurologic disorders including Alzheimer's disease, amyotrophic lateral sclerosis, and stroke (Markesbery and Carney 1999; Smith et al. 1998).

Catecholamines and their metabolites. The neurotransmitters dopamine and norepinephrine are synthesized at high levels in the brain where they act as modulatory neurotransmitters to regulate numerous brain functions. Altered levels of these catecholamines have been linked to a number of neurological disorders such as Parkinson's disease and brain injury. The serum concentration of norepinephrine is markedly increased in comatose, but not mild brain injury, patients (Clifton et al. 1981). Interestingly, in polytrauma patients, norepinephrine was increased regardless of brain injury severity, suggesting the release of norepinephrine in response to other bodily organs. The concentration of metabolites of dopamine (homovanillic acid, HVA) and serotonin (5-hydroxy indol acetic acid, 5-HIAA) have also been reported to increase in the CSF of TBI patients (Inagawa et al. 1980; Porta et al. 1975), with elevations in 5-HIAA being reliable indicators of brain trauma.

N-acetylaspartate (NAA). NAA, a derivative of aspartate, is the second most abundant chemical in the brain. NAA is synthesized by neurons and is thought to be involved in a variety of processes including fluid balance in the brain, energy production, and myelin synthesis. NAA also serves as the precursor for the generation of the neurotransmitter NAAG (*N*-acetylaspartylglutamate). Global decreases in the levels of NAA, as detected by proton magnetic resonance spectroscopy, have been associated with poorer outcome following mild TBI (Govind et al. 2010). Furthermore, the post-injury ratios of NAA to creatine in various brain regions have been demonstrated to predict aspects of neuropsychological dysfunction, indicating the utility of measuring regional biomarkers (Yeo et al. 2006). However, the specificity of NAA for brain trauma is very poor since decreased NAA levels have been observed in virtually all neuropathological conditions.

Glycolytic intermediates. A number of studies have shown that TBI alters the levels of glucose, lactate, pyruvate, glycerol, and glutamate both in the CSF and in microdialysates obtained from the injured brain. Lactate accumulation in the CSF after craniocerebral injury has been shown to correlate with injury severity (Czernicki and Grochowski 1976). Furthermore, clinical hypothermia, which has been tested as a therapy for brain injury, reduces lactate levels (Soukup et al. 2002), suggesting that this small molecule biomarker may be able to serve as a surrogate marker of treatment effectiveness.

Genetic Polymorphisms

The varied pathophysiology of TBI leads to a patient-unique constellation of symptoms including memory deficits, working memory deficits, attention deficits, depression, increased risk of Alzheimer's disease, epilepsy, and anxiety. Beyond the myriad causes of TBI and environmental influences that produce a diversity of patient outcomes, it is unclear why two patients with similar injuries can experience divergent clinical progression and outcomes. Because individuals with certain genetic polymorphisms have increased susceptibility to neurological conditions associated with these nucleotide mutations, if these people go on to suffer TBI they may have an added, or even multiplicative, risk of these conditions. One type of genetic polymorphism results from an inter-individual difference in the DNA sequence coding for a gene, sometimes resulting in a change in amino acid sequence and protein functionality. Table 17.1 lists genetic polymorphisms linked to neurological conditions also appearing as post-TBI symptoms. Future research may establish direct connections between these genetic polymorphisms and TBI outcome.

Interestingly, genetic polymorphisms of the protein ApoE have been associated with an increased risk of Alzheimer's disease in TBI patients. ApoE promotes maintenance of neuronal membranes and is involved in neural repair and remodeling and the formation of new synapses (Cedazo-Minguez 2007). The $\epsilon 4$ isoform of ApoE is associated with reduced stability of the protein. This isoform has been demonstrated to have an increased binding affinity to amyloid β peptides and promotes rapid aggregation of amyloid fibrils, which may lead to neurofibrillary tangles characteristic of Alzheimer's disease through a dysfunctional interaction with tau proteins (Strittmatter et al. 1993, 1994). Further,

Table 17.1 Genetic polymorphisms of symptoms common in traumatic brain injury

Symptom	Protein linked	Site of polymorphism	References
Reduced plasticity/ regeneration	ApoE	ϵ 4 Arg at amino acid positions 112 and 158	Arendt et al. (1997, 1998) and Ji et al. (2003)
	BDNF	Val to Met substitution at codon 66	Kleim et al. (2006), Krueger et al. (2011), and McHughen et al. (2010)
Declarative memory deficits	BDNF	Val to Met substitution at codon 66	Egan et al. (2003) and Hariri et al. (2003)
Working memory deficits	BDNF	Val to Met substitution at codon 66	Richter-Schmidinger et al. (2011)
	COMT	G to A (Val to Met at codon 158)	Egan et al. (2001) and Meyer-Lindenberg et al. (2006)
Memory recall (verbal)	NTSR1	rs4334545 and rs6090453	Li et al. (2011)
	HTR2A (5-HT _{2a} receptor)	His to Tyr substitution at residue 452	de Quervain et al. (2003)
Attention deficits	COMT	G to A (Val to Met at codon 158)	Blasi et al. (2005)
	CHRNA7 (α 7 nicotinic acetylcholine receptor)	15q13–q14 linkage region	Leonard et al. (2002)
Depression Post-traumatic stress disorder	SLC6A4 (SERT)	Short version of the allele	Canli and Lesch (2007)
	SLC6A4 (SERT)	Short version of the allele	Bryant et al. (2010), Koenen et al. (2009), and Kolassa et al. (2010)
	ADCYAP1R1 (receptor for pituitary adenylate cyclase-activating polypeptide)	rs2267735	Ressler et al. (2011)
Increased risk of Alzheimer's	ApoE	ϵ 4 Arg at amino acid positions 112 and 158	Lambert et al. (2002)
Epilepsy	SCN1A (sodium channel)	IVS5N+5 G to A, rs3812718	Schlachter et al. (2009)
	Copy number variants	15q13.3, 16p13.11 and 15q11.2	Sisodiya and Mefford (2011)

A adenine, *ApoE* Apolipoprotein E, *Arg* arginine, *BDNF* brain-derived neurotrophic factor, *COMT* catechol *O*-methyltransferase, *G* guanine, *His* histidine, *Met* methionine, *SERT* serotonin receptor, *Tyr* tyrosine, *Val* valine

the presence of ϵ 4 isoform is associated with increased deposits of amyloid β peptides in the brains of TBI patients (Nicoll et al. 1995) and also affects memory function following TBI (Crawford et al. 2002). Therefore, genetic polymorphisms, such as those found in the gene for ApoE, may be useful for determining whether a TBI patient has increased risk of other potential neurological disorders subsequent to his or her injury.

Genetic polymorphisms have implications for the treatment of TBI as well. Many enzymes involved in drug metabolism contain functionally important genetic polymorphisms that result in stratified levels of catalytic activity and even inactivity. The cytochrome P450 superfamily of mixed-function oxidative enzymes, which participate in the phase I metabolism of most current drugs, is an enzyme family that harbors many such polymorphisms. This may partially explain why some patients respond considerably differently to the same dose of a drug than other patients. For example, variations in the *VKORC1* and *CYP2C9* genes produce broad warfarin metabolic rates among individuals (Tan et al. 2010). This narrow therapeutic window leads to excessive bleeding or cerebrovascular

clotting and stroke in some patients. Thus, knowing a patient's genetic polymorphisms may better guide drug dosing, ultimately increasing drug effectiveness and decreasing risk. It may also guide researchers and clinicians in understanding the potential risks in the development of subsequent neurological and psychological impairments in TBI patients. Ultimately, this could result in TBI patients receiving more individualized and effective treatment, and improved outcomes.

mRNA and miRNA

An exciting area that has potential for identifying novel molecular TBI biomarkers or biomarker signatures is the genomic changes that occur in response to the primary and/or secondary injury. Whether the injury magnitude is mild, moderate, or severe, it is thought to trigger a unique pattern of changes in TBI-related molecular mechanisms and signaling pathways in the different affected CNS cell types, leading to potentially predictable changes in messenger RNA (mRNA) and micro RNA (miRNA) expression levels. To date, the most heavily investigated genomic changes after TBI have been at the level of altered expression levels of mRNAs isolated from CNS tissues. Although standard molecular biology techniques have been used for decades to examine changes in expression (spatial, temporal, and level) of an individual gene, since the turn of the century there has been an increased emphasis on the global analysis of genomic responses to more fully appreciate the complex pathophysiology of TBI (Crawford et al. 2007; Di Pietro et al. 2010; Kobori et al. 2002; Matzilevich et al. 2002; Rall et al. 2003). These and other studies have found that up to several hundred genes belonging to many different functional groups respond to TBI including, among others, genes involved in signal transduction, transcription/translation, cell cycle, inflammation, proliferation, and plasticity.

It has been reported that expression of 353 and 305 genes were significantly altered in the hippocampus at 3- and 24-h post-injury, respectively, after a moderate/severe controlled cortical impact injury (Matzilevich et al. 2002). Some gene functional classifications, such as cell cycle, growth factors, inflammation, and neuropeptides were found to be exclusively upregulated, likely related to the types of pathophysiological processes activated by TBI (e.g., inflammation). In the cortex, 403 genes showed altered expression 24 h after injury (Rall et al. 2003). Similar to the hippocampal data, functional classifications including transcription/translation, signal transduction, and metabolism showed the largest numbers of mRNAs with altered expression after injury. Interestingly, there appeared to be a temporal separation between TBI-related gene regulation in the hippocampus and cortex. Approximately 25% of the genes that were transiently upregulated at 3 h in the hippocampus (i.e., had returned to control levels at 24 h) were found to be upregulated in the cortex at 24 h (Dash et al. 2004). Controlled cortical impact in rats has also been shown to increase the expression of 146 mRNA transcripts whose expression levels vary according to time after injury, and include chemokines, adhesion molecules, regulators of growth, and cellular signaling (Israelsson et al. 2008). Microarray analysis of mRNA extracted from 10 or 50% mechanically stretched organotypic hippocampal slice cultures yielded 908 downregulated genes, 307 of which were shared across treatments, while 341 genes were upregulated, with only 30 genes shared across treatments (Di Pietro et al. 2010). Some laboratories have even explored the changes in mRNA expression profiles of single TUNEL-positive neurons from the cortex of fluid percussion brain-injured rats and demonstrated that there was differential expression of such genes as CREB, NR2A, NR2C, GluR2, and RED1 at 12- and 24-h post-injury (O'Dell et al. 2000). These and other related studies have greatly increased our understanding of the signaling pathways and gene networks activated after TBI, and identified novel genes that could potentially be exploited in future studies (Dash et al. 2004).

There is mounting evidence that clinically informative diagnostic and prognostic mRNA signatures may shortly be forthcoming for some diseases and pathologies such as cancer, but their effective application in the case of TBI is problematic. While it is relatively safe and easy to obtain tissue or biopsy samples from most cancer patients, it is both ethically dubious and technically difficult to

obtain comparable traumatized CNS tissue samples from patients for mRNA analysis. Furthermore, unprotected mRNAs released into the extracellular space or circulation from dead and dying cells are rapidly degraded by RNases and would therefore quickly become useless for reliable, quantitative analysis of biofluids. Given the bidirectionality of communication between the brain and periphery, however, it is possible that cells in the periphery, in particular circulating immune and/or mobilized stem cells, would contain characteristic changes in gene expression that reflect the extent of CNS damage and/or repair processes. To explore this, a study by Moore et al. examined changes in peripheral blood mononuclear cells in response to acute ischemic stroke and found a distinct pattern of largely upregulated genes. The authors were further able to identify a panel of 22 genes associated with white blood cell activation, hypoxic stress, and vascular repair that comprised a predictable pattern of expression that was 78% sensitive and 80% specific in correctly identifying stroke patients from an independent patient cohort (Moore et al. 2005). Subsequent analyses of whole blood samples have identified differential expression of several additional mRNA products unique to stroke patients including ARG1, CA4, LY96, MMP9, and S100A12, which may also be useful in classifying patients (Barr et al. 2010; Tang et al. 2006). Grond-Ginsbach et al. compared mononuclear leukocyte mRNA expression patterns between acute ischemic stroke patients, stroke survivors, acute TBI patients, and healthy control subjects, and identified differences in PDE4D, FPRL1, C3AR1, and IL1RN expression levels between stroke and healthy subjects (Grond-Ginsbach et al. 2008). Unfortunately, there are currently no comparable published studies examining peripheral blood for potential mRNA molecular biomarkers in TBI patients. These stroke studies, however, indicate that future studies examining the genomic response of peripheral whole blood following TBI may be useful for identifying molecular biomarkers of brain injury.

In contrast to mRNAs, which are unprotected and rapidly degraded upon exposure to the extracellular environment, miRNAs have been reliably detected in many cell-free preparations of bodily fluids, including serum, plasma, urine, saliva, and CSF, making miRNAs attractive candidates for exploitation as molecular biomarkers. Circulating miRNAs are thought to be resistant to extracellular RNase degradation because they are encapsulated within protective microparticles. Microparticles, which include exosomes, microvesicles, and apoptotic bodies, are very small (approximately 0.05–3.0 μm) abundant lipoprotein vesicles present in the circulation in normal conditions, while increased circulating microparticle levels are associated with several pathologic conditions (Doeuvre et al. 2009; Mause and Weber 2010; Orozco and Lewis 2010). Microparticles are produced by most cell types, express characteristic surface proteins derived from their cell of origin, and contain a complex mixture of intracellular components including mRNAs, miRNAs, and cytosolic proteins (Ai et al. 2010; Laterza et al. 2009; Mathivanan and Simpson 2009; Simpson et al. 2009). Interestingly, the relative abundance of the contents within exosomes appears to be distinct from that of the generating cell type, suggesting that there may be selective accumulation of specific mRNA, miRNA, and protein species (Al-Nedawi et al. 2009; Skog et al. 2008; Valadi et al. 2007). Several studies have demonstrated that secreted exosomes and other microparticles are able to transfer their biologically active cargo into distant recipient cells, resulting in modified gene expression, protein translation, and signaling (Al-Nedawi et al. 2008; Kosaka et al. 2010; Skog et al. 2008; Smalheiser 2007; Valadi et al. 2007; Yuan et al. 2009; Zernecke et al. 2009).

While new miRNAs are still being identified and their specific functions have yet to be fully elucidated, they are gaining increased attention as potential biomarkers for the detection and classification of multiple pathologies including cancer, lupus, hepatitis, and traumatic injury (Dai et al. 2010; Laterza et al. 2009; Li et al. 2004; Redell et al. 2010, 2011). miRNAs possess multiple characteristics that make them attractive as potentially useful molecular biomarkers for identifying tissue- and disease-specific pathologies, including (a) cell- and tissue-specific expression patterns (Cordes and Srivastava 2009; Rogelj and Giese 2004; Avnit-Sagi et al. 2009), (b) expression levels altered in disease-specific patterns (Bartels and Tsongalis 2009; Lu et al. 2005), and (c) stable and detectable in serum, plasma, CSF, and other bodily fluids (Gilad et al. 2008; Michael et al. 2010; Mitchell et al. 2008). Many recent studies have demonstrated characteristic changes in serum and plasma miRNA profiles associated

with several different types of cancer or other pathologic conditions, indicating their potential clinical utility as molecular biomarkers of disease (Gilad et al. 2008; Mitchell et al. 2008; Taylor and Gercel-Taylor 2008; Wang et al. 2009). For example, elevated plasma levels of miR-17-3p and miR-92 were detected in colorectal cancer patients, and miR-92 was 89% sensitive and 70% specific for differentiating colorectal cancer patients from healthy volunteers (relative expression cut-off 240; miR-92 normalized against RNU6B snoRNA) (Ng et al. 2009), and the plasma miR-92a:miR-638 ratio is a sensitive marker for the detection of acute leukemia (Tanaka et al. 2009). Damage-induced release of tissue-specific miRNAs into the circulation have also demonstrated diagnostic utility as biomarkers of myocardial infarction, stroke, liver damage, and neurotrauma (Ai et al. 2010; Laterza et al. 2009; Wang et al. 2009, 2010). Alterations in the plasma levels of miRNAs miR-16, miR-92a, and miR-765 in severe TBI patients (GCS score >8), which when combined were 100% specific and sensitive for identifying severe TBI patients in a small cohort of patient samples, further demonstrating the potential clinical utility of miRNAs for the diagnosis and treatment of individuals with TBI.

Biomarkers of Injury to Other Bodily Organs

It is important to recognize that brain injury often occurs within the context of other bodily injuries, a condition known as polytrauma. Thus, the diagnostic value of a biomarker can be confounded if the occurrence of injuries to other organs also affects its levels. This is particularly important in cases where damage to an organ may be difficult to ascertain by current imaging and diagnostic procedures.

Recent work has identified several proteins that have utility in identifying damage to a particular organ. Used in combination with available TBI biomarkers, these proteins could be used to determine the reliability of recorded levels of TBI biomarkers and/or assist in the selection of a biomarker signature that provides sufficient specificity.

Kidney

Kidney injury molecule-1 (KIM-1). KIM-1 is a receptor that recognizes apoptotic cells exposing phosphatidyl serine on their outer plasma membrane. After a kidney injury, the mRNA for KIM-1 is increased more than any other gene, and the extracellular domain of KIM-1 protein is cleaved and released into the urine. In murine models of kidney injury, urine KIM-1 levels out-perform serum creatinine and blood urea nitrogen as a biomarker for kidney injury. Furthermore, the urine levels of KIM-1 correlate with pathological lesions of the kidney (Vaidya et al. 2010).

Neutrophil gelatinase-associated lipocalin (NGAL). NGAL is a small (25 kDa) secreted protein that binds and transports small lipophilic molecules such as retinol and prostaglandins. NGAL is expressed in multiple cell types in the body and can be found in the loop of Henle and distal tubule of nephrons. Urine and plasma NGAL levels increase within a few hours of acute kidney injury or ischemic injury to the kidney (Al-Ismaili et al. 2011; Grigoryev et al. 2008).

Clusterin. Clusterin is secreted from kidney cells and is involved in cell aggregation, cell attachment, and cell protection. Enhanced levels of clusterin can be detected in the urine of animals and humans with kidney injury. Clusterin is more sensitive than serum creatinine and blood urea nitrogen as a diagnostic marker for acute kidney injury.

Trefoil factor 3 (TEF3). TEF3 is a small peptide hormone secreted by a number of cell types in the body including mucus-producing cells and epithelial cells. A drop in the level of urinary TEF3 has both good specificity and sensitivity for the diagnosis of acute kidney injury.

Liver-type fatty acid binding protein (L-FABP). Liver-type fatty acid binding proteins are 14–15 kDa proteins that bind to and transport long-chain fatty acids. L-FABP is expressed in the proximal tubule of nephrons, and its expression is increased within minutes to hours after kidney injury. A phase II study reported that increased L-FABP is an excellent biomarker of kidney injury and higher urinary L-FABP levels correlated with poor outcome (Ferguson et al. 2010).

Liver

Alanine aminotransferase (ALT). ALT is currently regarded as the gold standard for determining cellular hepatotoxicity. ALT activity is normally low in the bloodstream, but is released into the circulation from damaged hepatocytes.

Aspartate aminotransferase (AST). AST activity is less liver-specific than ALT. It is expressed predominantly in the heart and liver, and to a lesser degree in the kidney and muscle.

Gamma glutamyl transferase (GGT). GGT is relatively highly expressed in the liver, especially in cells lining the bile ducts, but is also found in the spleen, pancreas, and kidney. GGT is a relatively sensitive hepatobiliary injury marker, especially for cholestasis, but is not particularly specific.

Alkaline phosphatase (ALP). The highest concentrations of ALP activity are found in the liver (ALP1) and bone (ALP2), with lesser activity expressed in the intestine, kidney, and the placenta of pregnant women. Within the liver, ALP activity is predominantly expressed in cells lining the bile duct, and its activity is a relatively sensitive hepatobiliary injury marker.

Prothrombin time, albumin, and globulins. Unlike the liver enzymes mentioned above, these proteins are constitutively synthesized and released by the liver. Therefore, low serum levels are indicative of impaired liver function, but are not specific to liver trauma.

Bilirubin. Bilirubin is released from the liver following the breakdown of hemoglobin. Bilirubin tests usually examine the total and direct (conjugated) bilirubin forms in circulation to gauge the detoxification capacity of the liver.

Gut

Intestinal-type fatty acid binding protein (I-FABP). I-FABP is not detectable in the plasma of healthy volunteers (Fakhry et al. 2003; Pelters et al. 2003), but has been shown to be elevated in the blood and urine after shock, sepsis, and systemic inflammatory response syndrome (de Haan et al. 2009; Derikx et al. 2010; Lieberman et al. 1998). Plasma I-FABP levels have been demonstrated to correlate with the extent of abdominal trauma (de Haan et al. 2009).

Procalcitonin (PCT). Early post-injury PCT levels are positively associated with severity of abdominal injury (Maier et al. 2009; Sauerland et al. 2003) and serum PCT levels are also effective at predicting sepsis and inflammation.

D-Lactate. D-lactate is produced by bacterial metabolism in the gastrointestinal tract. Plasma D-lactate is used to measure gastrointestinal perfusion related to gut barrier function (Jessen and Mirsky 1985), and serum levels are typically elevated above the normal nanomolar levels in cases involving infection, ischemia, and trauma (Ewaschuk et al. 2005).

Polymorphonuclear elastase. Polymorphonuclear elastase is a neutrophilic inflammatory marker. Elevations in polymorphonuclear elastase levels have been associated with sepsis in abdominal surgery patients (Duswald et al. 1985) and have been found to increase 12–24 h after abdominal trauma (Hensler et al. 2002).

Heart

Creatinine kinase-muscle/brain isozyme (CK-MB). CK-MB is one of the most studied biomarkers being evaluated clinically for cardiac trauma patients. CK-MB is highly expressed in myocardial tissue and is rapidly released from necrotic myocardial cells into the circulation after injury (4–6 h). However, CK-MB is also expressed in non-cardiac tissues such as skeletal muscle, brain, kidney, liver, and small intestine, making the specificity of CK-MB for cardiac trauma rather poor (Bansal et al. 2005; McLean et al. 2008; Saenger 2010).

Troponin. Troponins (T, I, and C subunits) are cytosolic proteins involved in Ca²⁺-mediated contraction of skeletal and cardiac muscle. Small amounts of cardiac troponin T (cTnT, 6%) and I (cTnI, 3%) stay in free/unbound forms and are responsible for the early increase in troponin released into the blood after acute cardiac events such as infarction, trauma, and toxic damage. Increased cTnT level is reported in approximately one-third of patients with cardiac injury suggested by echocardiography, while cTnI has been reported to have a sensitivity of only 63% for cardiac contusion diagnosis (Edouard et al. 2004; Gupta and de Lemos 2007).

Heart-type fatty acid binding protein (H-FABP). H-FABP is also rapidly released into the circulation after necrotic cardiac injuries. H-FABP has higher sensitivity (93%) for the diagnosis of myocardial infarction than both CK-MB and cTn in the acute phase of injury (less than 3 h). However, H-FABP is present in multiple organs and tissues, and therefore has a low specificity for cardiac injury (McLean et al. 2008; Pelsers et al. 2005).

Lung

Krebs von den Lungen-6 (KL-6). KL-6 is a high-molecular-weight glycoprotein originally classified as mucin, a major component of the mucus layer covering the airway epithelium. KL-6 is an excellent biomarker for type II alveolar cell injury since these cells express KL-6 at very high levels. At a 550 U/ml cutoff, KL-6 has also been reported to serve as a serum marker for interstitial pneumonia, with a sensitivity and diagnostic accuracy of 95% (Kobayashi and Kitamura 1995).

Receptor for advanced glycation end-products (RAGE). RAGE is a transmembrane receptor of the immunoglobulin superfamily that is constitutively expressed at low levels in virtually all cell types, but is expressed in abundance in lung tissue (Buckley and Ehrhardt 2010; Leclerc et al. 2009). Recently, it has been reported that plasma RAGE levels correlate with clinical outcome following acute lung injury, especially in patients ventilated with higher tidal volumes (Calfee et al. 2008).

Surfactant protein D (SP-D). SP-D is a high-molecular-weight protein that is synthesized by non-ciliated epithelial cells in the peripheral airway. Elevated plasma SP-D levels are associated with a greater risk of death, fewer ventilator-free days, and fewer organ failure-free days (Eisner et al. 2003). At a serum cut off value of 110 ng/ml, SP-D shows a sensitivity of 85% and diagnostic accuracy of 95% for idiopathic pulmonary fibrosis (Takahashi et al. 2000).

Intercellular adhesion molecule-1 (ICAM-1). A glycoprotein primarily synthesized by endothelial cells, lymphocytes, and macrophages, ICAM-1 expression is markedly increased by exposure to proinflammatory cytokines such IL-1 and TNF α . Serum levels of ICAM-1 have been shown to increase with acute lung injury, although it is not a specific marker for lung injury.

Summary

The search for clinically useful biomarkers, biological molecules whose expression/activity changes in association with an injury or disease, has become an intense area of preclinical and clinical study. Both hypothesis-driven and unbiased approaches have been used to identify potential biomarkers. With additional research, promising TBI biomarkers can be clinically validated with large-scale clinical studies. Biomarkers may help stratify the risk of secondary injury, including increased intracranial pressure, and may provide particular utility in the diagnosis of concussion/mild TBI where imaging results are inconclusive. Metabolites and microRNAs offer promising new avenues for identifying informative/prognostic biomarkers. Genetic polymorphisms also have the potential to improve risk assessment and likely outcomes. Biomarkers of trauma will be useful surrogates whose monitoring could play a role in improving clinical care by assessing effectiveness of therapeutic interventions and improving outcome prediction.

Acknowledgments This chapter was made possible by funding support from the Department of Defense (W81XWH-08-2-0134 and W81XWH-11-2-0056), the Health Resources and Services Administration (R38OT10585), The Institute for Rehabilitation and Research/Mission Connect, The Vivian L. Smith Foundation, and the Gilson-Longenbaugh Foundation.

References

- Ai, J., Zhang, R., Li, Y., Pu, J., Lu, Y., Jiao, J., Li, K., Yu, B., Li, Z., Wang, R., Wang, L., Li, Q., Wang, N., Shan, H., Li, Z., & Yang, B. (2010). Circulating microRNA-1 as a potential novel biomarker for acute myocardial infarction. *Biochemical and Biophysical Research Communications*, *391*(1), 73–77.
- Al-Ismaili, Z., Palijan, A., & Zappitelli, M. (2011). Biomarkers of acute kidney injury in children: Discovery, evaluation, and clinical application. *Pediatric Nephrology*, *26*(1), 29–40.
- Al-Nedawi, K., Meehan, B., Micallef, J., Lhotak, V., May, L., Guha, A., & Rak, J. (2008). Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells. *Nature Cell Biology*, *10*(5), 619–624.
- Al-Nedawi, K., Meehan, B., & Rak, J. (2009). Microvesicles: Messengers and mediators of tumor progression. *Cell Cycle*, *8*(13), 2014–2018.
- Arendt, T., Schindler, C., Bruckner, M. K., Eschrich, K., Bigl, V., Zedlick, D., & Marcova, L. (1997). Plastic neuronal remodeling is impaired in patients with Alzheimer's disease carrying apolipoprotein epsilon 4 allele. *Journal of Neuroscience*, *17*(2), 516–529.
- Arendt, T., Bruckner, M. K., Gertz, H. J., & Marcova, L. (1998). Cortical distribution of neurofibrillary tangles in Alzheimer's disease matches the pattern of neurons that retain their capacity of plastic remodelling in the adult brain. *Neuroscience*, *83*(4), 991–1002.
- Avnit-Sagi, T., Kantorovich, L., Kredon-Russo, S., Hornstein, E., & Walker, M. D. (2009). The promoter of the pri-miR-375 gene directs expression selectively to the endocrine pancreas. *PLoS One*, *4*(4), e5033.
- Bansal, M. K., Maraj, S., Chewaproug, D., & Amanullah, A. (2005). Myocardial contusion injury: Redefining the diagnostic algorithm. *Emergency Medicine Journal*, *22*(7), 465–469.
- Barr, T. L., Conley, Y., Ding, J., Dillman, A., Warach, S., Singleton, A., & Matarin, M. (2010). Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. *Neurology*, *75*(11), 1009–1014.
- Bartels, C. L., & Tsongalis, G. J. (2009). MicroRNAs: Novel biomarkers for human cancer. *Clinical Chemistry*, *55*(4), 623–631.
- Bayir, H., Marion, D. W., Puccio, A. M., Wisniewski, S. R., Janesko, K. L., Clark, R. S., & Kochanek, P. M. (2004). Marked gender effect on lipid peroxidation after severe traumatic brain injury in adult patients. *Journal of Neurotrauma*, *21*(1), 1–8.

- Bayir, H., Adelson, P. D., Wisniewski, S. R., Shore, P., Lai, Y., Brown, D., Janesko-Feldman, K. L., Kagan, V. E., & Kochanek, P. M. (2009). Therapeutic hypothermia preserves antioxidant defenses after severe traumatic brain injury in infants and children. *Critical Care Medicine*, *37*(2), 689–695.
- Belin, A. C., & Westerlund, M. (2008). Parkinson's disease: A genetic perspective. *The FEBS Journal*, *275*(7), 1377–1383.
- Blasi, G., Mattay, V. S., Bertolino, A., Elvevag, B., Callicott, J. H., Das, S., Kolachana, B. S., Egan, M. F., Goldberg, T. E., & Weinberger, D. R. (2005). Effect of catechol-O-methyltransferase val158met genotype on attentional control. *Journal of Neuroscience*, *25*(20), 5038–5045.
- Brophy, G., Mondello, S., Papa, L., Robicsek, S., Gabrielli, A., Tepas, I. J., Buki, A., Robertson, C., Tortella, F. C., & Wang, K. K. (2011). Biokinetic analysis of ubiquitin C-terminal hydroxylase-L1 (UCH-L1) in severe traumatic brain injury patient biofluids. *Journal of Neurotrauma*, *28*(6), 861–870.
- Bryant, R. A., Felmingham, K. L., Falconer, E. M., Pe, B. L., Dobson-Stone, C., Pierce, K. D., & Schofield, P. R. (2010). Preliminary evidence of the short allele of the serotonin transporter gene predicting poor response to cognitive behavior therapy in posttraumatic stress disorder. *Biological Psychiatry*, *67*(12), 1217–1219.
- Buckley, S. T., & Ehrhardt, C. (2010). The receptor for advanced glycation end products (RAGE) and the lung. *Journal of Biomedicine and Biotechnology*, *2010*, 917108.
- Calfee, C. S., Ware, L. B., Eisner, M. D., Parsons, P. E., Thompson, B. T., Wickersham, N., & Matthay, M. A. (2008). Plasma receptor for advanced glycation end products and clinical outcomes in acute lung injury. *Thorax*, *63*(12), 1083–1089.
- Canli, T., & Lesch, K. P. (2007). Long story short: The serotonin transporter in emotion regulation and social cognition. *Nature Neuroscience*, *10*(9), 1103–1109.
- Cedazo-Minguez, A. (2007). Apolipoprotein E and Alzheimer's disease: Molecular mechanisms and therapeutic opportunities. *Journal of Cellular and Molecular Medicine*, *11*(6), 1227–1238.
- Chiaretti, A., Genovese, O., Aloe, L., Antonelli, A., Piastra, M., Polidori, G., & Di, R. C. (2005). Interleukin 1beta and interleukin 6 relationship with paediatric head trauma severity and outcome. *Child's Nervous System*, *21*(3), 185–193.
- Chiaretti, A., Antonelli, A., Mastrangelo, A., Pezzotti, P., Tortorolo, L., Tosi, F., & Genovese, O. (2008). Interleukin-6 and nerve growth factor upregulation correlates with improved outcome in children with severe traumatic brain injury. *Journal of Neurotrauma*, *25*(3), 225–234.
- Clifton, G. L., Ziegler, M. G., & Grossman, R. G. (1981). Circulating catecholamines and sympathetic activity after head injury. *Neurosurgery*, *8*(1), 10–14.
- Cordes, K. R., & Srivastava, D. (2009). MicroRNA regulation of cardiovascular development. *Circulation Research*, *104*(6), 724–732.
- Crawford, F. C., Vanderploeg, R. D., Freeman, M. J., Singh, S., Waisman, M., Michaels, L., Abdullah, L., Warden, D., Lipsky, R., Salazar, A., & Mullan, M. J. (2002). APOE genotype influences acquisition and recall following traumatic brain injury. *Neurology*, *58*(7), 1115–1118.
- Crawford, F. C., Wood, M., Ferguson, S., Mathura, V. S., Faza, B., Wilson, S., Fan, T., O'Steen, B., it-Ghezala, G., Hayes, R., & Mullan, M. J. (2007). Genomic analysis of response to traumatic brain injury in a mouse model of Alzheimer's disease (APPsw). *Brain Research*, *1185*, 45–58.
- Czernicki, Z., & Grochowski, W. (1976). Cerebrospinal fluid lactates following craniocerebral injuries. *Neurologia i Neurochirurgia Polska*, *10*(5), 651–653.
- Dai, R., Zhang, Y., Khan, D., Heid, B., Caudell, D., Crasta, O., & Ahmed, S. A. (2010). Identification of a common lupus disease-associated microRNA expression pattern in three different murine models of lupus. *PLoS One*, *5*(12), e14302.
- Dash, P. K., Kober, N., & Moore, A. N. (2004). A molecular description of brain trauma pathophysiology using microarray technology: An overview. *Neurochemical Research*, *29*(6), 1275–1286.
- Dash, P. K., Redell, J. B., Hergenroeder, G., Zhao, J., Clifton, G. L., & Moore, A. (2010). Serum ceruloplasmin and copper are early biomarkers for traumatic brain injury-associated elevated intracranial pressure. *Journal of Neuroscience Research*, *88*(8), 1719–1726.
- de Haan, J. J., Lubbers, T., Derikx, J. P., Relja, B., Henrich, D., Greve, J. W., Marzi, I., & Buurman, W. A. (2009). Rapid development of intestinal cell damage following severe trauma: A prospective observational cohort study. *Critical Care*, *13*(3), R86.
- de Quervain, D. J., Henke, K., Aerni, A., Coluccia, D., Wollmer, M. A., Hock, C., Nitsch, R. M., & Papassotiropoulos, A. (2003). A functional genetic variation of the 5-HT_{2a} receptor affects human memory. *Nature Neuroscience*, *6*(11), 1141–1142.
- Derikx, J. P., Bijker, E. M., Vos, G. D., van Bijnen, A. A., Heineman, E., Buurman, W. A., & van Waardenburg, D. A. (2010). Gut mucosal cell damage in meningococcal sepsis in children: Relation with clinical outcome. *Critical Care Medicine*, *38*(1), 133–137.
- Di Pietro, V., Amin, D., Pernagallo, S., Lazzarino, G., Tavazzi, B., Vagnozzi, R., Pringle, A., & Belli, A. (2010). Transcriptomics of traumatic brain injury: Gene expression and molecular pathways of different grades of insult in a rat organotypic hippocampal culture model. *Journal of Neurotrauma*, *27*(2), 349–359.

- Doeuvre, L., Plawinski, L., Toti, F., & ngles-Cano, E. (2009). Cell-derived microparticles: A new challenge in neuroscience. *Journal of Neurochemistry*, *110*(2), 457–468.
- Duswald, K. H., Jochum, M., Schramm, W., & Fritz, H. (1985). Released granulocytic elastase: An indicator of pathobiochemical alterations in septicemia after abdominal surgery. *Surgery*, *98*(5), 892–899.
- Edouard, A. R., Felten, M. L., Hebert, J. L., Cosson, C., Martin, L., & Benhamou, D. (2004). Incidence and significance of cardiac troponin I release in severe trauma patients. *Anesthesiology*, *101*(6), 1262–1268.
- Egan, M. F., Goldberg, T. E., Kolachana, B. S., Callicott, J. H., Mazzanti, C. M., Straub, R. E., Goldman, D., & Weinberger, D. R. (2001). Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(12), 6917–6922.
- Egan, M. F., Kojima, M., Callicott, J. H., Goldberg, T. E., Kolachana, B. S., Bertolino, A., Zaitsev, E., Gold, B., Goldman, D., Dean, M., Lu, B., & Weinberger, D. R. (2003). The BDNF val66met polymorphism affects activity-dependent secretion of BDNF and human memory and hippocampal function. *Cell*, *112*(2), 257–269.
- Eisner, M. D., Parsons, P., Matthay, M. A., Ware, L., & Greene, K. (2003). Plasma surfactant protein levels and clinical outcomes in patients with acute lung injury. *Thorax*, *58*(11), 983–988.
- Ewaschuk, J. B., Naylor, J. M., & Zello, G. A. (2005). D-lactate in human and ruminant metabolism. *Journal of Nutrition*, *135*(7), 1619–1625.
- Fakhry, S. M., Watts, D. D., & Luchette, F. A. (2003). Current diagnostic approaches lack sensitivity in the diagnosis of perforated blunt small bowel injury: Analysis from 275,557 trauma admissions from the EAST multi-institutional HVI trial. *The Journal of Trauma*, *54*(2), 295–306.
- Ferguson, M. A., Vaidya, V. S., Waikar, S. S., Collings, F. B., Sunderland, K. E., Gioules, C. J., & Bonventre, J. V. (2010). Urinary liver-type fatty acid-binding protein predicts adverse outcomes in acute kidney injury. *Kidney International*, *77*(8), 708–714.
- Fridriksson, T., Kini, N., Walsh-Kelly, C., & Hennes, H. (2000). Serum neuron-specific enolase as a predictor of intracranial lesions in children with head trauma: A pilot study. *Academic Emergency Medicine*, *7*(7), 816–820.
- Gilad, S., Meiri, E., Yogev, Y., Benjamin, S., Lebanony, D., Yerushalmi, N., Benjamin, H., Kushnir, M., Cholak, H., Melamed, N., Bentwich, Z., Hod, M., Goren, Y., & Chajut, A. (2008). Serum microRNAs are promising novel biomarkers. *PLoS One*, *3*(9), e3148.
- Giulian, D., & Lachman, L. B. (1985). Interleukin-1 stimulation of astroglial proliferation after brain injury. *Science*, *228*(4698), 497–499.
- Goldstein, I. M., Kaplan, H. B., Edelson, H. S., & Weissmann, G. (1982). Ceruloplasmin: An acute phase reactant that scavenges oxygen-derived free radicals. *Annals of the New York Academy of Sciences*, *389*, 368–379.
- Govind, V., Gold, S., Kaliannan, K., Saigal, G., Falcone, S., Arheart, K. L., Harris, L., Jagid, J., & Maudsley, A. A. (2010). Whole-brain proton MR spectroscopic imaging of mild-to-moderate traumatic brain injury and correlation with neuropsychological deficits. *Journal of Neurotrauma*, *27*(3), 483–496.
- Grigoryev, D. N., Liu, M., Hassoun, H. T., Cheadle, C., Barnes, K. C., & Rabb, H. (2008). The local and systemic inflammatory transcriptome after acute kidney injury. *Journal of the American Society of Nephrology*, *19*(3), 547–558.
- Grond-Ginsbach, C., Hummel, M., Wiest, T., Horstmann, S., Pflieger, K., Hergenbahn, M., Hollstein, M., Mansmann, U., Grau, A. J., & Wagner, S. (2008). Gene expression in human peripheral blood mononuclear cells upon acute ischemic stroke. *Journal of Neurology*, *255*(5), 723–731.
- Gupta, S., & de Lemos, J. A. (2007). Use and misuse of cardiac troponins in clinical practice. *Progress in Cardiovascular Diseases*, *50*(2), 151–165.
- Hall, E. D., Detloff, M. R., Johnson, K., & Kupina, N. C. (2004). Peroxynitrite-mediated protein nitration and lipid peroxidation in a mouse model of traumatic brain injury. *Journal of Neurotrauma*, *21*(1), 9–20.
- Hariri, A. R., Goldberg, T. E., Mattay, V. S., Kolachana, B. S., Callicott, J. H., Egan, M. F., & Weinberger, D. R. (2003). Brain-derived neurotrophic factor val66met polymorphism affects human memory-related hippocampal activity and predicts memory performance. *Journal of Neuroscience*, *23*(17), 6690–6694.
- Hensler, T., Sauerland, S., Bouillon, B., Raum, M., Rixen, D., Helling, H. J., Andermahr, J., & Neugebauer, E. A. (2002). Association between injury pattern of patients with multiple injuries and circulating levels of soluble tumor necrosis factor receptors, interleukin-6 and interleukin-10, and polymorphonuclear neutrophil elastase. *The Journal of Trauma*, *52*(5), 962–970.
- Hergenroeder, G., Redell, J. B., Moore, A. N., Dubinsky, W. P., Funk, R. T., Crommett, J., Clifton, G. L., Levine, R., Valadka, A., & Dash, P. K. (2008). Identification of serum biomarkers in brain-injured adults: Potential for predicting elevated intracranial pressure. *Journal of Neurotrauma*, *25*(2), 79–93.
- Hergenroeder, G. W., Moore, A. N., McCoy, J. P., Jr., Samsel, L., Ward, N. H., III, Clifton, G. L., & Dash, P. K. (2010). Serum IL-6: A candidate biomarker for intracranial pressure elevation following isolated traumatic brain injury. *Journal of Neuroinflammation*, *7*, 19.
- Herrmann, M., Curio, N., Jost, S., Grubich, C., Ebert, A. D., Fork, M. L., & Synowitz, H. (2001). Release of biochemical markers of damage to neuronal and glial brain tissue is associated with short and long term neuropsychological outcome after traumatic brain injury. *Journal of Neurology, Neurosurgery, and Psychiatry*, *70*(1), 95–100.

- Inagawa, T., Ishikawa, S., & Uozumi, T. (1980). Homovanillic acid and 5-hydroxyindoleacetic acid in the ventricular CSF of comatose patients with obstructive hydrocephalus. *Journal of Neurosurgery*, *52*(5), 635–641.
- Israelsson, C., Bengtsson, H., Kylberg, A., Kullander, K., Lewen, A., Hillered, L., & Ebendal, T. (2008). Distinct cellular patterns of upregulated chemokine expression supporting a prominent inflammatory role in traumatic brain injury. *Journal of Neurotrauma*, *25*(8), 959–974.
- Jessen, K. R., & Mirsky, R. (1985). Glial fibrillary acidic polypeptides in peripheral glia. Molecular weight, heterogeneity and distribution. *Journal of Neuroimmunology*, *8*(4–6), 377–393.
- Ji, Y., Gong, Y., Gan, W., Beach, T., Holtzman, D. M., & Wisniewski, T. (2003). Apolipoprotein E isoform-specific regulation of dendritic spine morphology in apolipoprotein E transgenic mice and Alzheimer's disease patients. *Neuroscience*, *122*(2), 305–315.
- Kleim, J. A., Chan, S., Pringle, E., Schallert, K., Procaccio, V., Jimenez, R., & Cramer, S. C. (2006). BDNF val66met polymorphism is associated with modified experience-dependent plasticity in human motor cortex. *Nature Neuroscience*, *9*(6), 735–737.
- Kobayashi, J., & Kitamura, S. (1995). KL-6: A serum marker for interstitial pneumonia. *Chest*, *108*(2), 311–315.
- Kobori, N., Clifton, G. L., & Dash, P. (2002). Altered expression of novel genes in the cerebral cortex following experimental brain injury. *Brain Research. Molecular Brain Research*, *104*(2), 148–158.
- Kochanek, P. M., Berger, R. P., Bayir, H., Wagner, A. K., Jenkins, L. W., & Clark, R. S. (2008). Biomarkers of primary and evolving damage in traumatic and ischemic brain injury: Diagnosis, prognosis, probing mechanisms, and therapeutic decision making. *Current Opinion in Critical Care*, *14*(2), 135–141.
- Koenen, K. C., Aiello, A. E., Bakshis, E., Amstadter, A. B., Ruggiero, K. J., Acierno, R., Kilpatrick, D. G., Gelernter, J., & Galea, S. (2009). Modification of the association between serotonin transporter genotype and risk of posttraumatic stress disorder in adults by county-level social environment. *American Journal of Epidemiology*, *169*(6), 704–711.
- Kolassa, I. T., Ertl, V., Eckart, C., Glockner, F., Kolassa, S., Papassotiropoulos, A., de Quervain, D. J., & Elbert, T. (2010). Association study of trauma load and SLC6A4 promoter polymorphism in posttraumatic stress disorder: Evidence from survivors of the Rwandan genocide. *The Journal of Clinical Psychiatry*, *71*(5), 543–547.
- Kosaka, N., Iguchi, H., Yoshioka, Y., Takeshita, F., Matsuki, Y., & Ochiya, T. (2010). Secretory mechanisms and intercellular transfer of microRNAs in living cells. *Journal of Biological Chemistry*, *285*(23), 17442–17452.
- Kossmann, T., Stahel, P. F., Lenzlinger, P. M., Redl, H., Dubs, R. W., Trentz, O., Schlag, G., & Morganti-Kossmann, M. C. (1997). Interleukin-8 released into the cerebrospinal fluid after brain injury is associated with blood-brain barrier dysfunction and nerve growth factor production. *Journal of Cerebral Blood Flow and Metabolism*, *17*(3), 280–289.
- Krueger, F., Pardini, M., Huey, E. D., Raymont, V., Solomon, J., Lipsky, R. H., Hodgkinson, C. A., Goldman, D., & Grafman, J. (2011). The role of the Met66 brain-derived neurotrophic factor allele in the recovery of executive functioning after combat-related traumatic brain injury. *Journal of Neuroscience*, *31*(2), 598–606.
- Lambert, J. C., Araria-Goumidi, L., Myllykangas, L., Ellis, C., Wang, J. C., Bullido, M. J., Harris, J. M., Artiga, M. J., Hernandez, D., Kwon, J. M., Frigard, B., Petersen, R. C., Cumming, A. M., Pasquier, F., Sastre, I., Tienari, P. J., Frank, A., Sulkava, R., Morris, J. C., St Clair, D., Mann, D. M., Wavrant-DeVrieze, F., Ezquerro-Trabalon, M., Amouyel, P., Hardy, J., Haltia, M., Valdivieso, F., Goate, A. M., Perez-Tur, J., Lendon, C. L., & Chartier-Harlin, M. C. (2002). Contribution of APOE promoter polymorphisms to Alzheimer's disease risk. *Neurology*, *59*(1), 59–66.
- Laterza, O. F., Lim, L., Garrett-Engele, P. W., Vlasakova, K., Muniappa, N., Tanaka, W. K., Johnson, J. M., Sina, J. F., Fare, T. L., Sistare, F. D., & Glaab, W. E. (2009). Plasma MicroRNAs as sensitive and specific biomarkers of tissue injury. *Clinical Chemistry*, *55*(11), 1977–1983.
- Leclerc, E., Fritz, G., Vetter, S. W., & Heizmann, C. W. (2009). Binding of S100 proteins to RAGE: An update. *Biochimica et Biophysica Acta*, *1793*(6), 993–1007.
- Leonard, S., Gault, J., Hopkins, J., Logel, J., Vianzon, R., Short, M., Drebing, C., Berger, R., Venn, D., Sirota, P., Zerbe, G., Olincy, A., Ross, R. G., Adler, L. E., & Freedman, R. (2002). Association of promoter variants in the alpha7 nicotinic acetylcholine receptor subunit gene with an inhibitory deficit found in schizophrenia. *Archives of General Psychiatry*, *59*(12), 1085–1096.
- Li, H. H., Lee, S. M., Cai, Y., Sutton, R. L., & Hovda, D. A. (2004). Differential gene expression in hippocampus following experimental brain trauma reveals distinct features of moderate and severe injuries. *Journal of Neurotrauma*, *21*(9), 1141–1153.
- Li, J., Chen, C., Chen, C., He, Q., Li, H., Li, J., Moyzis, R. K., Xue, G., & Dong, Q. (2011). Neurotensin Receptor 1 Gene (NTSR1) polymorphism is associated with working memory. *PloS One*, *6*(3), e17365.
- Lieberman, J. M., Marks, W. H., Cohn, S., Jaicks, R., Woode, L., Sacchetti, J., Fischer, B., Moller, B., & Burns, G. (1998). Organ failure, infection, and the systemic inflammatory response syndrome are associated with elevated levels of urinary intestinal fatty acid binding protein: Study of 100 consecutive patients in a surgical intensive care unit. *The Journal of Trauma*, *45*(5), 900–906.
- Lu, J., Getz, G., Miska, E. A., varez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., Downing, J. R., Jacks, T., Horvitz, H. R., & Golub, T. R. (2005). MicroRNA expression profiles classify human cancers. *Nature*, *435*(7043), 834–838.

- Maier, M., Wutzler, S., Lehnert, M., Szermutzky, M., Wyen, H., Bingold, T., Henrich, D., Walcher, F., & Marzi, I. (2009). Serum procalcitonin levels in patients with multiple injuries including visceral trauma. *The Journal of Trauma*, *66*(1), 243–249.
- Markesbery, W. R., & Carney, J. M. (1999). Oxidative alterations in Alzheimer's disease. *Brain Pathology*, *9*(1), 133–146.
- Mathivanan, S., & Simpson, R. J. (2009). ExoCarta: A compendium of exosomal proteins and RNA. *Proteomics*, *9*(21), 4997–5000.
- Matzilevich, D. A., Rall, J. M., Moore, A. N., Grill, R. J., & Dash, P. K. (2002). High-density microarray analysis of hippocampal gene expression following experimental brain injury. *Journal of Neuroscience Research*, *67*(5), 646–663.
- Mause, S. F., & Weber, C. (2010). Microparticles: Protagonists of a novel communication network for intercellular information exchange. *Circulation Research*, *107*(9), 1047–1057.
- McHUGHEN, S. A., RODRIGUEZ, P. F., KLEIM, J. A., KLEIM, E. D., MARCHAL, C. L., PROCACCIO, V., & CRAMER, S. C. (2010). BDNF val66met polymorphism influences motor system function in the human brain. *Cerebral Cortex*, *20*(5), 1254–1262.
- McLean, A. S., Huang, S. J., & Salter, M. (2008). Bench-to-bedside review: The value of cardiac biomarkers in the intensive care patient. *Critical Care*, *12*(3), 215.
- Meyer-Lindenberg, A., Nichols, T., Callicott, J. H., Ding, J., Kolachana, B., Buckholtz, J., Mattay, V. S., Egan, M., & Weinberger, D. R. (2006). Impact of complex genetic variation in COMT on human brain function. *Molecular Psychiatry*, *11*(9), 867–877.
- Michael, A., Bajracharya, S. D., Yuen, P. S., Zhou, H., Star, R. A., Illei, G. G., & Alevizos, I. (2010). Exosomes from human saliva as a source of microRNA biomarkers. *Oral Diseases*, *16*(1), 34–38.
- Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanian, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., & Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(30), 10513–10518.
- Moore, D. F., Li, H., Jeffries, N., Wright, V., Cooper, R. A., Jr., Elkahoul, A., Gelderman, M. P., Zudaire, E., Blevins, G., Yu, H., Goldin, E., & Baird, A. E. (2005). Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: A pilot investigation. *Circulation*, *111*(2), 212–221.
- Morganti-Kossmann, M. C., Hans, V. H., Lenzlinger, P. M., Dubs, R., Ludwig, E., Trentz, O., & Kossmann, T. (1999). TGF-beta is elevated in the CSF of patients with severe traumatic brain injuries and parallels blood-brain barrier function. *Journal of Neurotrauma*, *16*(7), 617–628.
- Ng, E. K., Chong, W. W., Jin, H., Lam, E. K., Shin, V. Y., Yu, J., Poon, T. C., Ng, S. S., & Sung, J. J. (2009). Differential expression of microRNAs in plasma of patients with colorectal cancer: A potential marker for colorectal cancer screening. *Gut*, *58*(10), 1375–1381.
- Nicoll, J. A., Roberts, G. W., & Graham, D. I. (1995). Apolipoprotein E epsilon 4 allele is associated with deposition of amyloid beta-protein following head injury. *Nature Medicine*, *1*(2), 135–137.
- Nylen, K., Ost, M., Csajbok, L. Z., Nilsson, I., Blennow, K., Nellgard, B., & Rosengren, L. (2006). Increased serum-GFAP in patients with severe traumatic brain injury is related to outcome. *Journal of Neurological Sciences*, *240*(1–2), 85–91.
- O'Dell, D. M., Raghupathi, R., Crino, P. B., Eberwine, J. H., & McIntosh, T. K. (2000). Traumatic brain injury alters the molecular fingerprint of TUNEL-positive cortical neurons in vivo: A single-cell analysis. *Journal of Neuroscience*, *20*(13), 4821–4828.
- Orozco, A. F., & Lewis, D. E. (2010). Flow cytometric analysis of circulating microparticles in plasma. *Cytometry. Part A*, *77*(6), 502–514.
- Papa, L., Akinyi, L., Liu, M. C., Pineda, J. A., Tepas, J. J., III, Oli, M. W., Zheng, W., Robinson, G., Robicsek, S. A., Gabrielli, A., Heaton, S. C., Hannay, H. J., Demery, J. A., Brophy, G. M., Layon, J., Robertson, C. S., Hayes, R. L., & Wang, K. K. (2010). Ubiquitin C-terminal hydrolase is a novel biomarker in humans for severe traumatic brain injury. *Critical Care Medicine*, *38*(1), 138–144.
- Pelinka, L. E., Kroepfl, A., Schmidhammer, R., Krenn, M., Buchinger, W., Redl, H., & Raabe, A. (2004). Glial fibrillary acidic protein in serum after traumatic brain injury and multiple trauma. *The Journal of Trauma*, *57*(5), 1006–1012.
- Pelsters, M. M., Namiot, Z., Kisielewski, W., Namiot, A., Januszkiewicz, M., Hermens, W. T., & Glatz, J. F. (2003). Intestinal-type and liver-type fatty acid-binding protein in the intestine. Tissue distribution and clinical utility. *Clinical Biochemistry*, *36*(7), 529–535.
- Pelsters, M. M., Hermens, W. T., & Glatz, J. F. (2005). Fatty acid-binding proteins as plasma markers of tissue injury. *Clinica Chimica Acta*, *352*(1–2), 15–35.
- Porta, M., Bareggi, S. R., Collice, M., Assael, B. M., Selenati, A., Calderini, G., Rossanda, M., & Morselli, P. L. (1975). Homovanillic acid and 5-hydroxyindole-acetic acid in the csf of patients after a severe head injury. II. Ventricular csf concentrations in acute brain post-traumatic syndromes. *European Neurology*, *13*(6), 545–554.
- Rall, J. M., Matzilevich, D. A., & Dash, P. K. (2003). Comparative analysis of mRNA levels in the frontal cortex and the hippocampus in the basal state and in response to experimental brain injury. *Neuropathology and Applied Neurobiology*, *29*(2), 118–131.

- Rebhun, J., Madorsky, J. G., & Glovsky, M. M. (1991). Proteins of the complement system and acute phase reactants in sera of patients with spinal cord injury. *Annals of Allergy*, *66*(4), 335–338.
- Redell, J. B., Moore, A. N., Ward, N. H., III, Hergenroeder, G. W., & Dash, P. K. (2010). Human traumatic brain injury alters plasma microRNA levels. *Journal of Neurotrauma*, *27*(12), 2147–2156.
- Redell, J. B., Zhao, J., & Dash, P. K. (2011). Altered expression of miRNA-21 and its targets in the hippocampus after traumatic brain injury. *Journal of Neuroscience Research*, *89*(2), 212–221.
- Ressler, K. J., Mercer, K. B., Bradley, B., Jovanovic, T., Mahan, A., Kerley, K., Norrholm, S. D., Kilaru, V., Smith, A. K., Myers, A. J., Ramirez, M., Engel, A., Hammack, S. E., Toufexis, D., Braas, K. M., Binder, E. B., & May, V. (2011). Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. *Nature*, *470*(7335), 492–497.
- Rhodes, J., Sharkey, J., & Andrews, P. (2009). Serum IL-8 and MCP-1 concentration do not identify patients with enlarging contusions after traumatic brain injury. *The Journal of Trauma*, *66*(6), 1591–1597.
- Richter-Schmidinger, T., Alexopoulos, P., Horn, M., Maus, S., Reichel, M., Rhein, C., Lewczuk, P., Sidiropoulos, C., Kneib, T., Pernecky, R., Doerfler, A., & Kornhuber, J. (2011). Influence of brain-derived neurotrophic-factor and apolipoprotein E genetic variants on hippocampal volume and memory performance in healthy young adults. *Journal of Neural Transmission*, *118*(2), 249–257.
- Rogelj, B., & Giese, K. P. (2004). Expression and function of brain specific small RNAs. *Reviews in the Neurosciences*, *15*(3), 185–198.
- Saenger, A. K. (2010). A tale of two biomarkers: The use of troponin and CK-MB in contemporary practice. *Clinical Laboratory Science*, *23*(3), 134–140.
- Sauerland, S., Hensler, T., Bouillon, B., Rixen, D., Raum, M. R., Andermahr, J., & Neugebauer, E. A. (2003). Plasma levels of procalcitonin and neopterin in multiple trauma patients with or without brain injury. *Journal of Neurotrauma*, *20*(10), 953–960.
- Schlachter, K., Gruber-Sedlmayr, U., Stogmann, E., Lausecker, M., Hotzy, C., Balzar, J., Schuh, E., Baumgartner, C., Mueller, J. C., Illig, T., Wichmann, H. E., Lichtner, P., Meitinger, T., Strom, T. M., Zimprich, A., & Zimprich, F. (2009). A splice site variant in the sodium channel gene SCN1A confers risk of febrile seizures. *Neurology*, *72*(11), 974–978.
- Semple, B. D., Bye, N., Rancan, M., Ziebell, J. M., & Morganti-Kossmann, M. C. (2010). Role of CCL2 (MCP-1) in traumatic brain injury (TBI): Evidence from severe TBI patients and CCL2^{-/-} mice. *Journal of Cerebral Blood Flow and Metabolism*, *30*(4), 769–782.
- Setuie, R., & Wada, K. (2007). The functions of UCH-L1 and its relation to neurodegenerative diseases. *Neurochemistry International*, *51*(2–4), 105–111.
- Simpson, R. J., Lim, J. W., Moritz, R. L., & Mathivanan, S. (2009). Exosomes: Proteomic insights and diagnostic potential. *Expert Review of Proteomics*, *6*(3), 267–283.
- Singh, I. N., Sullivan, P. G., Deng, Y., Mbye, L. H., & Hall, E. D. (2006). Time course of post-traumatic mitochondrial oxidative damage and dysfunction in a mouse model of focal traumatic brain injury: Implications for neuroprotective therapy. *Journal of Cerebral Blood Flow and Metabolism*, *26*(11), 1407–1418.
- Singh, I. N., Sullivan, P. G., & Hall, E. D. (2007). Peroxynitrite-mediated oxidative damage to brain mitochondria: Protective effects of peroxynitrite scavengers. *Journal of Neuroscience Research*, *85*(10), 2216–2223.
- Singhal, A., Baker, A. J., Hare, G. M., Reinders, F. X., Schlichter, L. C., & Moulton, R. J. (2002). Association between cerebrospinal fluid interleukin-6 concentrations and outcome after severe human traumatic brain injury. *Journal of Neurotrauma*, *19*(8), 929–937.
- Sisodiya, S. M., & Mefford, H. C. (2011). Genetic contribution to common epilepsies. *Current Opinion in Neurology*, *24*(2), 140–145.
- Skog, J., Wurdinger, T., Van, R. S., Meijer, D. H., Gainche, L., Sena-Esteves, M., Curry, W. T., Jr., Carter, B. S., Krichevsky, A. M., & Breakefield, X. O. (2008). Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. *Nature Cell Biology*, *10*(12), 1470–1476.
- Smalheiser, N. R. (2007). Exosomal transfer of proteins and RNAs at synapses in the nervous system. *Biology Direct*, *2*, 35.
- Smith, R. G., Henry, Y. K., Mattson, M. P., & Appel, S. H. (1998). Presence of 4-hydroxynonenal in cerebrospinal fluid of patients with sporadic amyotrophic lateral sclerosis. *Annals of Neurology*, *44*(4), 696–699.
- Song, L., & Pachter, J. S. (2004). Monocyte chemoattractant protein-1 alters expression of tight junction-associated proteins in brain microvascular endothelial cells. *Microvascular Research*, *67*(1), 78–89.
- Soukup, J., Zauner, A., Doppenberg, E. M., Menzel, M., Gilman, C., Bullock, R., & Young, H. F. (2002). Relationship between brain temperature, brain chemistry and oxygen delivery after severe human head injury: The effect of mild hypothermia. *Neurological Research*, *24*(2), 161–168.
- Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., & Roses, A. D. (1993). Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(5), 1977–1981.

- Strittmatter, W. J., Saunders, A. M., Goedert, M., Weisgraber, K. H., Dong, L. M., Jakes, R., Huang, D. Y., Pericak-Vance, M., Schmechel, D., & Roses, A. D. (1994). Isoform-specific interactions of apolipoprotein E with microtubule-associated protein tau: Implications for Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(23), 11183–11186.
- Takahashi, H., Fujishima, T., Koba, H., Murakami, S., Kurokawa, K., Shibuya, Y., Shiratori, M., Kuroki, Y., & Abe, S. (2000). Serum surfactant proteins A and D as prognostic factors in idiopathic pulmonary fibrosis and their relationship to disease extent. *American Journal of Respiratory and Critical Care Medicine*, *162*(3 Pt 1), 1109–1114.
- Tan, G. M., Wu, E., Lam, Y. Y., & Yan, B. P. (2010). Role of warfarin pharmacogenetic testing in clinical practice. *Pharmacogenomics*, *11*(3), 439–448.
- Tanaka, M., Oikawa, K., Takanashi, M., Kudo, M., Ohyashiki, J., Ohyashiki, K., & Kuroda, M. (2009). Down-regulation of miR-92 in human plasma is a novel marker for acute leukemia patients. *PLoS One*, *4*(5), e5532.
- Tang, Y., Xu, H., Du, X., Lit, L., Walker, W., Lu, A., Ran, R., Gregg, J. P., Reilly, M., Pancioli, A., Khoury, J. C., Sauerbeck, L. R., Carrozzella, J. A., Spilker, J., Clark, J., Wagner, K. R., Jauch, E. C., Chang, D. J., Verro, P., Broderick, J. P., & Sharp, F. R. (2006). Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: A microarray study. *Journal of Cerebral Blood Flow and Metabolism*, *26*(8), 1089–1102.
- Taylor, D. D., & Gercel-Taylor, C. (2008). MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. *Gynecologic Oncology*, *110*(1), 13–21.
- Todd, S. R. (2004). Critical concepts in abdominal injury. *Critical Care Clinics*, *20*(1), 119–134.
- Tomas, E., & Toparceanu, F. (1986). Considerations about the possible function of ceruloplasmin in influenza and parainfluenza virus infections. *Virologie*, *37*(4), 279–287.
- Uden, J., & Romner, B. (2010). Can low serum levels of S100B predict normal CT findings after minor head injury in adults? An evidence-based review and meta-analysis. *The Journal of Head Trauma Rehabilitation*, *25*(4), 228–240.
- Vaidya, V. S., Ozer, J. S., Dieterle, F., Collings, F. B., Ramirez, V., Troth, S., Muniappa, N., Thudium, D., Gerhold, D., Holder, D. J., Bobadilla, N. A., Marrer, E., Perentes, E., Cordier, A., Vonderscher, J., Maurer, G., Goering, P. L., Sistare, F. D., & Bonventre, J. V. (2010). Kidney injury molecule-1 outperforms traditional biomarkers of kidney injury in preclinical biomarker qualification studies. *Nature Biotechnology*, *28*(5), 478–485.
- Valadi, H., Ekstrom, K., Bossios, A., Sjostrand, M., Lee, J. J., & Lotvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature Cell Biology*, *9*(6), 654–659.
- Venetsanou, K., Vlachos, K., Moles, A., Fragakis, G., Fildissis, G., & Baltopoulos, G. (2007). Hypolipoproteinemia and hyperinflammatory cytokines in serum of severe and moderate traumatic brain injury (TBI) patients. *European Cytokine Network*, *18*(4), 206–209.
- Vos, P. E., Lamers, K. J., Hendriks, J. C., van, H. M., Beems, T., Zimmerman, C., van, G. W., de, R. H., Biert, J., & Verbeek, M. M. (2004). Glial and neuronal proteins in serum predict outcome after severe traumatic brain injury. *Neurology*, *62*(8), 1303–1310.
- Wang, K., Zhang, S., Marzolf, B., Troisch, P., Brightman, A., Hu, Z., Hood, L. E., & Galas, D. J. (2009). Circulating microRNAs, potential biomarkers for drug-induced liver injury. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(11), 4402–4407.
- Wang, G. K., Zhu, J. Q., Zhang, J. T., Li, Q., Li, Y., He, J., Qin, Y. W., & Jing, Q. (2010). Circulating microRNA: A novel potential biomarker for early diagnosis of acute myocardial infarction in humans. *European Heart Journal*, *31*(6), 659–666.
- Whalen, M. J., Carlos, T. M., Wisniewski, S. R., Clark, R. S., Mellick, J. A., Marion, D. W., & Kochanek, P. M. (2000). Effect of neutropenia and granulocyte colony stimulating factor-induced neutrophilia on blood-brain barrier permeability and brain edema after traumatic brain injury in rats. *Critical Care Medicine*, *28*(11), 3710–3717.
- Winter, C. D., Pringle, A. K., Clough, G. F., & Church, M. K. (2004). Raised parenchymal interleukin-6 levels correlate with improved outcome after traumatic brain injury. *Brain*, *127*(Pt 2), 315–320.
- Yeo, R. A., Phillips, J. P., Jung, R. E., Brown, A. J., Campbell, R. C., & Brooks, W. M. (2006). Magnetic resonance spectroscopy detects brain injury and predicts cognitive functioning in children with brain injuries. *Journal of Neurotrauma*, *23*(10), 1427–1435.
- Yuan, A., Farber, E. L., Rapoport, A. L., Tejada, D., Deniskin, R., Akhmedov, N. B., & Farber, D. B. (2009). Transfer of microRNAs by embryonic stem cell microvesicles. *PLoS One*, *4*(3), e4722.
- Zemlan, F. P., Jauch, E. C., Mulchahey, J. J., Gabbita, S. P., Rosenberg, W. S., Speciale, S. G., Zuccarello, M. (2002). C-tau biomarker of neuronal damage in severe brain injured patients: Association with elevated intracranial pressure and clinical outcome. *Brain Res*, *947*(1), 131–139.
- Zernecke, A., Bidzhekov, K., Noels, H., Shagdarsuren, E., Gan, L., Denecke, B., Hristov, M., Koppel, T., Jahantigh, M. N., Lutgens, E., Wang, S., Olson, E. N., Schober, A., & Weber, C. (2009). Delivery of microRNA-126 by apoptotic bodies induces CXCL12-dependent vascular protection. *Science Signaling*, *2*(100), ra81.

Chapter 18

Functional Outcomes

Renan C. Castillo

Introduction

In 1966, Donobedian stated that “Outcomes, by and large, remain the ultimate validation of the effectiveness and quality of medical care.” Since then, the outcomes movement has made a profound impact on medical care (Bourne et al. 2004). At the same time, it is increasingly clear that decision makers at all levels want to see evidence that their investment in policies and resources is going to be effective. This movement, which some call evidence-based public health (EBPH), challenges injury researchers to use state-of-the-art technology to demonstrate the value of injury control programs and policies. Measurement of functional outcome following injuries is a central tool in the assessment of the human and economic costs of injury, and is critical to the development and evaluation of programs and policies to improve outcomes following injury. A well-designed outcome measurement plan improves the quality of injury control research, minimizes study participant burden, and maximizes opportunities for future secondary data analyses. A key challenge in the development of a study measurement plan is the identification of appropriate, practical, well-validated measures.

In this chapter, we review the major theoretical and practical issues in choosing functional outcome measures and review some of the more widely used instruments. The first two sections of this chapter focus on concepts and tools used to assess the quality and usefulness of outcome measures. In section “Goals of Outcome Measurement” of this chapter, we review the main goals of outcome measurement and how these goals affect our choice of instruments and procedures. In section “Outcome Measurement Concepts,” we review the concepts of validity, reliability, and responsiveness for the measurement of functional outcomes following injury. Any discussion of the measurement of functional outcomes must begin with a review of the International Classification of Function (ICF), which has been created to provide a universal language for disability research. Section “The International Classification of Function” of this chapter focuses on a review of the ICF and the challenges we still face in operationalizing this tool. Finally, in section “The Future of Outcome Measurement,” we discuss the future of functional outcome measurement with a focus on new technologies such as electronic monitoring and computer adaptive testing. This chapter will not attempt,

R.C. Castillo, PhD (✉)

Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health,
624 North Broadway, HH 544, Baltimore, MD 21205, USA
e-mail: rcastill@jhsph.edu

as several authors have done before, to catalog the numerous outcome measurement instruments available to the injury researcher. Several excellent efforts already in the literature do a far more comprehensive job of this, and the universe of available instruments has not changed enough to warrant revisiting at this time. The one exception is the WHODAS 2.0, which is discussed as part of our review of the ICF. Readers looking for reviews of specific outcome measurement instruments are likely to find this in Ian McDowell's *Measuring Health* (Oxford University Press, 2006), which describes in great detail a wide range of health status measurements. For more concise reviews of only the major instruments, readers are referred to Ellen MacKenzie's chapter on *Measuring Disability and Quality of Life Post Injury*, in *Injury Control* (Cambridge University Press, 2001).

Goals of Outcome Measurement

A key consideration in the development of an outcome measurement plan is the specific reason, or goal, of measurement. McDowell et al. (2004) describe the four main goals of overall health measurement as description, explanation, prediction, and evaluation. A similar classification may be used for the measurement of functional outcomes in injury research. This classification is summarized in Table 18.1.

When the function of measurement is description, the goal is to characterize the impact of a disease, condition, or intervention across a population. The measurement tool must be practical to implement across large samples, and must be either a "natural measure" – a measure that has obvious relevance to a natural or social process – or one that has been so well described as to have direct correlation with a natural measure. In injury research, a number of such natural measures are used. The most widely used is mortality. Another important and widely used descriptive measure in this field is return to work following injury. This is particularly true in occupational injury research (where lost work days is the main driver of workers compensation costs) and trauma outcomes research (where a key goal of the rehabilitation process is returning people back to productive employment). Another widely used descriptive measure is independence, which is particularly relevant to older populations for whom functional independence is a direct correlate of being able to reside outside of nursing or long-term care facilities.

When the function of measurement is prediction, the goal is to characterize the future impact of a disease, condition, or intervention. A widely used predictive measure is the Injury Severity Score (Baker et al. 1974), which measures injury severity as a function of the risk of death due to that injury. Another such measure is the Charlson Comorbidity Index (Charlson et al. 1987), which summarizes a person's comorbid conditions into a single score meant to reflect the increased risk of death due to these conditions.

When the function of measurement is explanation, the goal is to measure a number of domains in order to shed light on a process or mechanism. Explanatory measures may be used to identify the mechanism of action of an intervention, or understand the process by which a disease or condition results in restrictions in social participation. An explanatory measure is generally well rooted in a theoretical framework that will lend backing to an explanatory approach. For example, the WHODAS V2 may be considered an explanatory measure in that it reflects many of the domains in the ICF. Similarly, a measure like the Pain Stages of Change Questionnaire (Kerns et al. 1997), which is rooted in Prochaska's transtheoretical stages of change framework (Prochaska and DiClemente 1984), may serve as a useful explanatory measure when investigating the etiology of a successful injury recovery intervention.

When the function of measurement is evaluation, the goal is to assess the effect of programs or policies in *changing* the health or functional status of a population. Because the goal is to assess change, the key characteristic of these instruments is responsiveness: the ability to record

Table 18.1 Key qualities of measurement instruments

Goal of measurement	Key quality of measurement instrument	Sample construct or instrument
Description	Natural measures	Mortality Independence
Prediction	High predictive validity	Injury Severity Scale ^a Charlson Comorbidity Index ^b
Explanation	Multidimensional measures rooted in a theoretical framework	WHODAS V2 Pain Stages of Change Questionnaire ^c
Evaluation	High responsiveness	SF-36 ^d Sickness Impact Profile ^e

^aBaker et al. (1974)^bCharlson et al. (1987)^cKerns et al. (1997)^dWare and Sherbourne (1992)^eJurkovich et al. (1995)

measurable change when a change has occurred in the observed population. This is not a trivial matter. While a great deal of emphasis is correctly placed on the validity and reliability of a measure, responsiveness is not uniformly reported in psychometric validation papers. In the next section of this chapter, we briefly review some of the criteria used for the evaluation of instrument responsiveness as part of a general discussion of three basic measurement concepts: validity, reliability, and responsiveness.

Outcome Measurement Concepts

A key challenge in the development of a study measurement plan is the identification of appropriate, practical, well-validated measures. While the literature on functional outcomes is enormous and there are literally hundreds of candidate instruments, the subset that might be useful to a researcher for a particular study is generally much smaller. Thus, it is a substantial challenge to identify useful, relevant measures. Further complicating matters is the fact that validation approaches vary substantially from study to study and instrument to instrument.

When identifying a measurement instrument, the two primary questions a researcher must ask are (a) does this instrument measure the construct, outcome, or condition I am trying to measure, and (b) how consistent is this instrument. If the goal of the research is to evaluate change, a researcher might also ask a third question: (c) can I expect this instrument to change if my program or intervention causes the health or function of my population to change. The first question is generally answered by examining the validity of the instrument, while the second question is referred to as reliability. As we discussed earlier, the third question relates to instrument responsiveness.

Measurement validity means that the instrument in question succeeds in describing or quantifying what it is designed to measure. A measure of physical functional recovery following ankle injuries should measure physical functioning, particularly around the lower extremity. It should contain questions focusing on mobility and ambulation, stair climbing, balance, gait, etc. At the same time, it should ideally not focus on, for example, anxious distress as a result of the injury. Measurement validity is never fully demonstrated, and in some cases may be dependent on the population being studied. For example, a valid and reliable measure of patient satisfaction with care might be measuring entirely different constructs in a primary care setting, a chronic disease setting, and a trauma care setting.

One challenge when assessing instrument validation studies is the fact that a wide range of validation approaches are often used. Generally, there are four main classes of evidence of validity: face validity, content validity, criterion validity, and construct validity.

Face and content validity are most often established during the development phase of an instrument. Face validity is often used when there is evidence that the instrument is accepted by “experts” as being valid “on the face of it.” Experts in this setting might not just refer to researchers and clinical specialists, but also front-line clinicians, patients, family members, or policy makers. Content validity refers to whether the items included in the instrument cover the universe of items that encompass the construct under consideration. This is generally accomplished during the instrument development phase by compiling an “item bank” from a number of previously validated instruments measuring closely related constructs. The developers use this item bank to assess the extent to which their own items are representative of the times covered by the larger set of related measures. It is extremely common during the development of an instrument to convene an expert or stakeholder panel to assist in the assessment of face and content validity.

Criterion and construct validity are most commonly examined when instruments are being evaluated within the target population as part of specific validation studies. As such, understanding these two types of evidence of validity is critical when evaluating functional outcome measures. Criterion validity assesses whether the instrument agrees with an objective external criterion. This objective external criterion can take one of two forms: a future event or outcome, preferably a natural outcome such as death or loss of independence, or a “gold standard” instrument. When comparing against a future outcome, researchers refer to this type of validity as predictive evidence. When comparing against a “gold standard” outcome, researchers refer to this type of validity as concurrent evidence. In practice, these types of studies are not routinely available. Predictive evidence requires expensive longitudinal studies and the availability of a good natural outcome to examine in the future. Concurrent evidence requires the availability of a gold standard. Gold standard measures, when available, may also be expensive to administer. More common applications of concurrent evidence are studies which show the validity of a shortened version of a full scale.

Construct validity relates to whether the instrument is consistent with the theoretical concept being assessed. All tests of validity are ultimately designed to provide evidence of an instrument’s construct validity. As with criterion validity, there are several types of evidence of construct validity. Convergent evidence is demonstrated when a measure correlates highly with measures of the same construct. For example, a measure of self-efficacy should ideally be highly correlated with other measures of self-efficacy. At the same time, that same measure of self-efficacy should have low correlation with instruments measuring a different construct, such as catastrophizing or hope. This type of evidence of construct validity is called discriminant evidence. Researchers will often combine convergent and discriminant evidence of construct validity into a single table showing the correlation coefficients of the new measure with multiple other measures.

Two other forms of evidence of construct validity are “known group differences” and factor structure. Known group differences are demonstrated when a measure is shown to score differently between groups where different scores would be expected. For example, a post-traumatic stress disorder (PTSD) measure such as the PTSD Checklist (PCL) (Weathers et al. 1993) would be expected to yield higher scores among individuals who have recently experienced a traumatic event than people who have not, and even higher among people who have experienced traumatic events in which their life was in danger. Factorial evidence of construct validity is shown when the clustering of the items in the instrument supports the theory-based grouping. For example, items in a measure of readiness to engage in injury prevention behaviors based on Prochaska’s transtheoretical model of stages of change would be expected to cluster around the five stages (Prochaska and DiClemente 1984).

Generally, this clustering is investigated using psychometric techniques such as principal components and exploratory factor analyses.

Reliability of a measure or instrument refers to the degree to which the measurement technique produces consistent results upon repeated application. As such, reliability is best seen as a component of validity. In other words, a measure cannot be valid if it is not reliable. However, a measure can certainly be reliable in that it yields consistent results, but not valid in that it is consistently measuring a construct other than that which the researcher is trying to measure. Further, a measure may be valid and reliable in a particular setting, (such as for phone administration by a trained interviewer) but not reliable (and thus also not valid) in a different setting, such as self-administration using paper and pencil. This is a critical aspect of reliability that must be considered when assessing measurement instruments: reliability must be evaluated using circumstances that are as close as possible to the conditions under which the instrument will be used in your study. These conditions include mode of administration, interviewer training, setting, respondent population, timing, and other potential sources of variability. Thus, reliability is not a property of an instrument per se, but rather an instrument has a certain degree of reliability when applied to certain populations under certain conditions.

Reliability refers to a number of characteristics, the most common of which are consistency when measured at different times (test–retest reliability), consistency when measured by different observers (inter-rater reliability), and consistency when measured using different subsets of items within the same scale (internal consistency). The most commonly used measure of reliability across time and raters is the kappa statistic.

Internal consistency, the extent to which the items in a scale “belong or hang together,” is often measured using Cronbach’s coefficient alpha. Regardless of the method used to measure reliability, it is important to remember that reliability is defined mathematically as the ratio of the variance of the true score to the variance of the observed score. If the ratio of true variance to observed variance is high, it makes sense that the reliability of the measure is high. In general, reliabilities above 0.7 are considered acceptable, and above 0.8 are considered good. Note that exceedingly high reliabilities (say, over 0.95) may indicate the measure is too repetitive, and might reasonably be shortened.

The final criterion used to evaluate outcome instruments is responsiveness: the ability to record measurable change when a change has occurred in the observed population. This is critical to assess if the purpose of measurement is evaluation of a policy or intervention using longitudinal data. Unfortunately, this is one of the least well-reported characteristics of measurement instruments. In this section, we briefly review some of the criteria used for the evaluation of instrument responsiveness. Ultimately, responsiveness hinges on evidence of a difference in a measure between a group that has changed or has undergone an intervention known to be effective, compared with a control group. A control group can be a baseline measurement, prior to the intervention, or a separate group that did not undergo the intervention. The usual concerns about selection and historical biases must of course be addressed depending on the study design. There are three widely used indices of responsiveness. Most commonly used is effect size, which is obtained by dividing the average change score by the standard deviation of the “baseline” score for a longitudinal design, or of the control group baseline for experimental or quasi-experimental designs. The standardized response mean (SRM) is obtained by dividing the average change score by the standard deviation of the change. Calculation of the SRM is only possible in longitudinal designs. Guyatt’s responsiveness index (GRI) is obtained by dividing the average change score by the standard deviation of the change in the control group (Guyatt et al. 1993). Thus, the GRI is only possible in longitudinal designs with a control group. The GRI thus has the advantage of adjusting for systematic change for both the changed and stable groups.

The International Classification of Function

The key theoretical concept in the measurement of functioning and disability following injuries is the International Classification of Functioning, Disability and Health (ICF), released by the World Health Organization in 2001 (WHO 2001). The publication of the ICF has been described as a “landmark” event in the fields of rehabilitation and disability research (Stamm and Machold 2007; Stucki et al. 2003). The ICF has been endorsed by the Institute of Medicine’s Committee on Disability in America, which used the ICF framework in its most recent report, *The Future of Disability in America* (IOM 2007). In order to examine the components and development of the ICF and related frameworks, it is important to review the history of the concepts and definitions that make up these models.

A critical definition when discussing disability is the concept of disablement. Because disability has been variously defined by different authors and institutions, and has been incorporated into different models over time, it is important to distinguish it from the broader concept of disablement. Disablement, as defined by Jette, is:

... global term that reflects all the diverse consequences that disease, injury, or congenital abnormalities may have on human functioning at many different levels.

(Jette 1994)

As such, disablement includes all of the “various impacts” of trauma or disease on both “specific body systems,” “basic human performance,” and “people’s functioning in necessary, usual, expected, and personally desired roles in society” (Verbrugge and Jette 1993). It is straightforward to see that this concept includes all the different levels of functioning that are typically discussed in the context of recovery following a major illness or injury: physical impairments, pain, psychologic distress, functioning, activity limitations, role participation, and return to work. Following Jette’s terminology, we refer to a model that attempts to link these concepts as a “disablement framework” (Jette 1994).

The two dominant disablement frameworks over the past 30 years have been derived from the work of Phillip Wood, which was adopted by the World Health Organization (WHO 1980), and the work of Saad Nagi, utilized by the Institute of Medicine (IOM 1991). Jette (1994) argues that most disablement frameworks developed since then are “modifications and extensions of these two early formulations.” While there are clear similarities between the two approaches, it is important to recognize that there are some important distinctions between the two frameworks.

Nagi’s Disablement Framework

The IOM published “Disability in America” in 1991, using the work of Saad Nagi in the 1960s as the basis for its definition of disability. Nagi’s (1991) disablement framework is based on three related but distinct concepts: (1) impairment, (2) functional limitations, and (3) disability. Impairment refers to a deficit or an abnormality at the anatomical, physiological, or mental level. A functional limitation is the behavioral manifestation of this deficit, or an abnormality that occurs at the level of the person. Both impairment and functional limitations involve function. Impairment is a loss of function at the organ or physiologic system level, while functional limitations refer to a loss of function at the level of the whole person. Nagi argued that disability is the result of a process in which a disease or injury leads to an impairment (reduced joint movement or reduced cognition). This impairment in turn can lead to a functional limitation (for example, the patient is unable to handle small objects, or unable to perform specific mental tasks). This functional limitation may result in disability if the functional limitation leads to a restriction in role functioning (such as being able to work or participate in a recreational activity). This model of disablement differed from previous conceptualizations in which

there was no distinction made between different levels of human functioning. Nagi (1991) defined disability as “the inability or limitation in performing socially defined roles and tasks expected of an individual within a sociocultural and physical environment.” While Nagi’s framework was not explicitly etiological, it was evident from the model that functional limitations provided at least one key pathway through which impairments resulted in disability.

Wood’s Disablement Framework

A decade before “Disability in America” was published, the WHO published the “International Classification of Impairments, Disabilities, and Handicaps (ICIDH)” based on the work of Phillip Wood (WHO 1980). Like Nagi’s work, Wood’s disablement framework is also built on three distinct concepts: (1) impairment, (2) disability, and (3) handicap. Impairment is “any loss or abnormality of psychological, physiological, or anatomical structure or function at the organ level” (WHO 1980). Thus, Nagi’s and Wood’s disablement frameworks build upon very similar concepts of impairment, but differ on the definition of disability. Wood defined disability as “any restriction or lack of ability to perform an activity in the manner considered normal for a human being” (WHO 1980). Thus, Wood is not defining disability as only a limitation in socially expected roles, but a limitation in any activity. A critical result from Nagi’s model is that a person with a functional limitation may adapt around that limitation and thus avoid disability. In Wood’s model, that same limitation constitutes a disability regardless of the level of adaptation, and a handicap results from any discrimination or exclusion that may arise due to this disability. This discrimination may in turn limit that person’s ability to participate in the roles that are normal for them. Thus, Wood’s model placed the final stage of disability within a framework of disability rights.

The International Classification of Function

The ICIDH was revised in 2001 with the release of the ICF, which expanded and clarified the definition of disability to include “the compound of integrated tasks, skills, and activities expected of a person or of the body as a whole” (Jette 1994). In addition, the ICF incorporates Nagi’s work by clarifying the concept of functional limitations as distinct from the broader concept of disability. The new ICF framework represents a substantial step forward in addressing two major criticisms of previous models: the failure to use positive language that could be equally used to describe excellent or poor function and a lack of emphasis on the role of environmental factors (De Kleijn-De Vrankrijker 2003; Hurst 2003). The taxonomy proposed in the ICF is becoming widely accepted in the field of disability research (McNaughton et al. 2001) and increasingly as a tool for policy and program development and evaluation (Ustün et al. 2003). As stated by Jette:

The ICF framework holds great promise to provide a synthesis of earlier models of disablement and to provide a universal language with which to discuss disability and related phenomena.

(Jette 2006)

The new terminology in the ICF specifies three levels of function rooted in the concept of positive health: body functions and structures, activity, and participation. The disablements associated with these three levels are thus: impairments in body functions and structures, activity limitations, and participation restrictions. The relationships between the levels of functioning in the ICF are mediated and moderated by environmental and personal contextual factors. The official WHO definitions of the components of the ICF are provided in Table 18.2.

Table 18.2 Components of the International Classification of Function (WHO 2001)

Body functions	Are the physiological functions of body systems (including psychological functions)
Body structures	Are anatomical parts of the body such as organs, limbs and their components
Impairments	Are problems in body function or structure such as a significant deviation or loss
Activity	Is the execution of a task or action by an individual
Activity limitations	Are difficulties an individual may have in executing activities
Participation	Is involvement in a life situation
Participation restrictions	Are problems an individual may experience in involvement in life situations
Environmental factors	Make up physical, social, and attitudinal environment in which people live and conduct their lives

Causal Relationships Within the ICF

While both Nagi and Wood believed that the disablement components were sequential in nature, the ICF “takes a neutral stand with regard to etiology” (WHO 2001). Nagi stated that limitations at lower levels of organization may be reflected at higher levels but the reverse is not true. For example, functional limitations may affect the level of disability but the opposite is not true. In the ICIDH, the arrows connecting the levels of functioning were unidirectional, suggesting causal relationships. In the ICF framework, unidirectional arrows that relate impairment with disability and handicap have been replaced with double-headed arrows between impairment, activity, and participation. Figure 18.1 shows the relationships between components of the ICF as illustrated by the WHO. The ICF thus encourages the scientific community to “collect data on these constructs independently and thereafter explore associations and causal links between them” (WHO 2001).

Operationalizing the Domains of Activity and Participation in the ICF, and the WHODAS 2.0

A current limitation of the ICF framework is the fact that it does not sufficiently distinguish between activity and participation. While the ICF provides separate definitions of activity and participation, it provides only a single list of domains, stating only that users may wish to “differentiate activities and participation in their own operational ways” (WHO 2001). As these constructs have become more widely used and accepted, the need to better define and measure them has become increasingly apparent (Jette et al. 2003).

A number of projects have been initiated to define and differentiate the activity limitation and participation domains of the ICF. Some of these studies have explored the development of condition-specific instruments or core sets for osteoarthritis (Pollard et al. 2006), developmental disabilities in children (McConachie et al. 2006), stroke (Schepers et al. 2007), rheumatoid arthritis (Coenen et al. 2006), spinal cord injury (Biering-Sørensen et al. 2006), and chronic pain (Dixon et al. 2007). The NIH-sponsored Patient-Reported Outcomes Measurement Information System (PROMIS) is working to develop a computer adaptive social role participation domain (PROMIS 2006). The participation team, led by researchers at the Toronto Rehabilitation Institute, is working on the “development of the construct of participation in the field of rehabilitation” (Participation Team Working Report 2005). Researchers at WHO have attempted to develop a set of candidate categories for an ICF generic core set, specifically for studies that cross multiple disease conditions (Cieza et al. 2006). Finally, researchers at the Rehabilitation Institute of Chicago have attempted to catalog

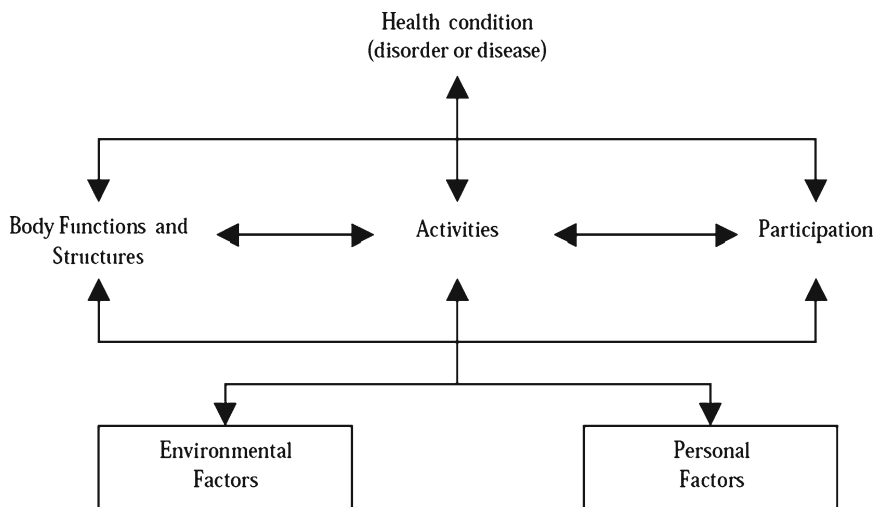


Fig. 18.1 Schematic representation of the relationships between components of the International Classification of Function (WHO 2001)

beliefs about participation from persons living with disabilities using qualitative interviewing techniques (Hammel et al. 2008).

Despite these efforts, no consensus exists yet about how to measure activity and participation. In addition, even if gold standard instruments for the measurement of participation are developed, a substantial body of research and collected data already utilizes functional outcome measures that do not include activity and participation domains per se. Thus, in parallel to the development of new instruments, it will be important to explore the feasibility of adapting previously developed measures to the new ICF framework. This will make it possible to utilize existing datasets to conduct research that is still relevant to the current language of the field of rehabilitation.

A major step forward in operationalizing the ICF is the development of the World Health Organization Disability Assessment Schedule, Version 2 (WHODAS 2.0). The WHODAS 2.0 differs from existing functional outcome measurements in that it was designed specifically to reflect the content and domains of the ICF. It has undergone extensive psychometric validation (Manual for WHODAS 2.0 2010) and is available in two forms: a 36-item questionnaire and a 12-item short form that captures 81% of the variance in the longer form. The WHODAS 2.0 captures information across six domains of function: cognition, mobility, self-care, getting along (interacting with other people), life activities, and participation.

The Future of Outcome Measurement

As new technologies emerge, we are increasingly able to utilize them to expand the ways in which functional outcomes are developed, delivered, and analyzed. These technologies will not only make outcome measurement easier in the future but also will in some ways revolutionize the quality and breadth of the data that can be captured. Two major examples are the increased tendency to utilize electronic devices in order to measure outcomes directly rather than rely on participant questionnaires, and the emergence of computer adaptive testing instruments.

Table 18.3 Sample applications of new technologies for direct observation of functional outcomes

Domain	Computer-assisted alternative	Sample application
Ambulation	Wireless-enabled pedometer	Numerous devices available in the market
Mobility	GPS technology, widely available in modern mobile devices, can be used to document extent to which participants travel to specific locations	Wolf and Jacobs (2010)
Physical exercise	Wireless-enabled accelerometer	http://www.polar.fi/en/ http://www.goherman.com/
Medication adherence	Pill bottle electronic monitoring devices	MEMS Monitor Glowcap
Gait	Gait monitoring system built into a watch can be used to assess gait patterns in real-world settings	Barth et al. (2010)
Instrumental activities of daily living	Flexible remote sensor technologies throughout the home can be used to discreetly monitor activities like cooking, bathroom use, medication use, etc.	http://www.grandcare.com

Increasingly, electronic monitoring devices make it possible to directly measure constructs that until now have relied on the research participant's recall and willingness to provide accurate information. Many examples are actively being used in research projects today. MEMS caps and related technologies make it possible to obtain accurate, real-time counts of patients' adherence to drug regimens. Some of these technologies come with built-in communication devices that make it possible to alert researchers if the patient has missed a dose, or if the pill bottle is being opened at the wrong time. Other products come with built-in adherence interventions, including lights, sounds, and the capacity to automatically send text reminders to participants' phones. Electronic step-counting devices also promise to revolutionize the way functional outcomes are evaluated following injuries. The ability to walk is a major determinant of success in recovery following injuries. Many functional outcome measurement tools include questions about walking, such as difficulty walking a certain distance or an estimate of the amount of walking that has been done recently. However, these rely on the participants' recall and are subject to many biases. Step monitoring devices provide accurate data on the exact number of steps the participant has taken over a precise period of time. Advanced monitoring devices can also examine gait deviations, and thus provide data on not only the amount but also the quality of walking. Similarly, mobile devices may soon become research tools. Patient diaries, real-time functional assessments and screeners, and mobility monitoring may all be delivered via devices people already carry around with them. While some of these technologies may appear daunting because of the technical expertise required to set up, customize, and deploy them, it is important to remember that technology becomes more accessible every day. And while cutting-edge technology always comes at a cost, these costs must be balanced against potential savings, particularly when the alternatives use relatively expensive survey methods. Table 18.3 lists just a few examples of functional outcomes domains that may be assessed using direct observation rather than by relying on patient questionnaires.

A second technology that is likely to revolutionize the measurement of functional outcomes following injury over the next decade is Computer Adaptive Testing, or CAT. CAT technology allows for dynamic, or adaptive, outcome measures. Rather than giving respondents a long battery of questions, only those questions that are most suitable for that individual are used. For example, if the construct in question is ambulation, a classical outcome measure might include multiple questions about walking 1 mile, 2 miles, many miles, in uneven terrain, etc. But none of these questions may be necessary if the respondent cannot walk at all. One can readily assume that a single question asking whether the respondent can walk a block, if answered negatively, would make any further questions about walking several blocks or several miles unnecessary. At the heart of CAT is item response theory, or IRT. IRT differs from classical testing, which is the basis for most current paper- and

pencil-based outcome measures. In classical testing, the key assumption is that the observed score is equal to the true score, plus bias, plus unsystematic error:

$$\text{Observed score} = \text{true score} + \text{bias} + \text{unsystematic error.}$$

Ideally, bias is small, and unsystematic error is centered at zero. As the number of items increases, unsystematic errors cancel out and the observed score approaches the true score. Thus, in classical testing, a larger number of items measuring the construct contribute to greater reliability. In IRT, there is a different central assumption, that the probability of a response is equal to the ability of the respondent minus the difficulty of the item:

$$P(\text{response}) = \text{ability} - \text{item difficulty.}$$

In IRT, any question in an item bank can be described by its item difficulty. Thus, any one question is replaceable by any other question from that same item bank, thus freeing the researcher from having to provide the same set of questions each time. A respondent who is performing at a high level of function can be given questions that have a high item difficulty, while a respondent at a low level of function can be given low-difficulty questions. The result is a greater level of accuracy with fewer questions.

CAT technology was prohibitive in the past for two main reasons: the need for computing power and software in order to perform the background calculations required to pick the best next question from the item bank and the lack of well-developed CAT outcome measures. Both of these limitations may disappear in the near future. The Patient Reported Outcomes Measurement Information System (PROMIS), an NIH-funded consortium that began in 2004, is in the process of developing CAT-ready item banks for dozens of constructs. Many of these are relevant to injury research, including domains for physical functioning, participation, and pain. In addition to developing these domains, the PROMIS network is making available the technology behind their CAT engine, which can be used via their website, as desktop software, and even within third-party data capture systems.

Thus, it is now possible for many injury researchers to access CAT technology and item banks. Widespread use of these banks will advance the field of injury research by reducing study participant burden, increasing the accuracy of outcome measurement, facilitating the capture of multiple domains of functional recovery, and increasing the comparability of results across studies. Combined with the widespread access to electronic devices, there is a real possibility for a qualitative change in the nature of outcome measurement over the coming decade.

References

- Baker, S. P., O'Neill, B., Haddon, W., Jr., & Long, W. B. (1974). The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *The Journal of Trauma*, *14*(3), 187–196.
- Barth, A. T., Bennett, B. C., Boudaoud, B., Brantley, J. S., Chen, S., Cunningham, C. L., et al. (2010). Longitudinal high-fidelity gait analysis with wireless inertial body sensors. *Wireless Health*, 192–193.
- Biering-Sørensen, F., Scheuringer, M., Baumberger, M., Charlifue, S. W., Post, M. W., Montero, F., et al. (2006). Developing core sets for persons with spinal cord injuries based on the International Classification of Functioning, Disability and Health as a way to specify functioning. *Spinal Cord*, *44*(9), 541–546.
- Bourne, R. B., Maloney, W. J., & Wright, J. G. (2004). An AOA critical issue: the outcome of the outcomes movement. *The Journal of Bone and Joint Surgery*, *86*, 633–640.
- Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*, *40*(5), 373–383.
- Cieza, A., Geyh, S., Chatterji, S., Kostanjsek, N., Ustün, B. T., & Stucki, G. (2006). Identification of candidate categories of the International Classification of Functioning Disability and Health (ICF) for a Generic ICF Core Set based on regression modelling. *BMC Medical Research Methodology*, *6*, 36.

- Coenen, M., Cieza, A., Stamm, T. A., Amann, E., Kollerits, B., & Stucki, G. (2006). Validation of the International Classification of Functioning, Disability and Health (ICF) Core Set for rheumatoid arthritis from the patient perspective using focus groups. *Arthritis Research & Therapy*, 8(4), R84.
- De Kleijn-De Vrankrijker, M. W. (2003). The long way from the International Classification of Impairments, Disabilities and Handicaps (ICIDH) to the International Classification of Functioning, Disability and Health (ICF). *Disability and Rehabilitation*, 25(11–12), 561–564.
- Dixon, D., Pollard, B., & Johnston, M. (2007). What does the chronic pain grade questionnaire measure? *Pain*, 130(3), 249–253.
- Donabedian, A. (1966). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44, 166–206.
- Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine*, 118, 622–629.
- Hammel, J., Magasi, S., Heinemann, A., Whiteneck, G., Bogner, J., & Rodriguez, E. (2008). What does participation mean? An insider perspective from people with disabilities. *Disability and Rehabilitation*, 30(19), 1445–1460.
- Hurst, R. (2003). The international disability rights movement and the ICF. *Disability and Rehabilitation*, 24, 572–576.
- Institute of Medicine (IOM); Pope, A. M., & Tarlov, A. R. (Eds.). (1991). *Disability in America: a national agenda for prevention*. Washington, DC: National Academy Press.
- Institute of Medicine's Committee on Disability in America; Field, M. J., & Jette, A. M. (Eds.). (2007). *The future of disability in America*. Washington, DC: National Academy Press.
- Jette, A. M. (1994). Physical disablement concepts for physical therapy research and practice. *Physical Therapy*, 74(5), 380–386.
- Jette, A. M. (2006). Toward a common language for function, disability, and health. *Physical Therapy*, 86(5), 726–734.
- Jette, A. M., Haley, S. M., & Kooyoomjian, J. T. (2003). Are the ICF activity and participation dimensions distinct? *Journal of Rehabilitation Medicine*, 35, 145–149.
- Jurkovich, G., Mock, C., MacKenzie, E., Burgess, A., Cushing, B., deLateur, B., et al. (1995). The Sickness Impact Profile as a tool to evaluate functional outcome in trauma patients. *The Journal of Trauma*, 39, 625–631.
- Kerns, R. D., Rosenberg, R., Jamison, R. N., Caudill, M. A., & Haythornthwaite, J. (1997). Readiness to adopt a self-management approach to chronic pain: the Pain Stages of Change Questionnaire (PSOCQ). *Pain*, 72(1–2), 227–234.
- McConachie, H., Colver, A. F., Forsyth, R. J., Jarvis, S. N., & Parkinson, K. N. (2006). Participation of disabled children: how should it be characterised and measured? *Disability and Rehabilitation*, 28(18), 1157–1164.
- McDowell, I., Spasoff, R. A., & Kristjansson, B. (2004). On the classification of population health measurements. *American Journal of Public Health*, 94(3), 388–393.
- McNaughton, H., McPherson, K., Falkner, E., & Taylor, W. (2001). Impairment, disability, handicap and participation in post-polio myelitis subjects. *International Journal of Rehabilitation Research*, 24, 133–136.
- Nagi, S. (1991). Disability concepts revisited: implications for prevention. In A. Pope & A. Tarlov (Eds.), *Disability in America: toward a national agenda for prevention* (pp. 309–327). Washington, DC: National Academy Press.
- Participation Team Working Report. (2005, October). *Conceptualizing and measuring participation*. Toronto: Toronto Rehabilitation Institute.
- Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS Preliminary Item Banks. (2006). [http://www.nihpromis.org/WebPages/PreliminaryItemBanks\(2006\).aspx](http://www.nihpromis.org/WebPages/PreliminaryItemBanks(2006).aspx).
- Pollard, B., Johnston, M., & Dieppe, P. J. (2006). What do osteoarthritis health outcome instruments measure? Impairment, activity limitation, or participation restriction? *The Journal of Rheumatology*, 33(4), 757–763.
- Prochaska, J. O., & DiClemente, C. C. (1984). *The transtheoretical approach: crossing traditional boundaries of therapy*. Homewood, IL: Dow Jones-Irwin.
- Schepers, V. P., Ketelaar, M., van de Port, I. G., Visser-Meily, J. M., & Lindeman, E. (2007). Comparing contents of functional outcome measures in stroke rehabilitation using the International Classification of Functioning, Disability and Health. *Disability and Rehabilitation*, 29(3), 221–230.
- Stamm, T., & Machold, K. (2007). The International Classification of Functioning, Disability and Health in practice in rheumatological care and research. *Current Opinion in Rheumatology*, 19(2), 184–189.
- Stucki, G., Ewert, T., & Cieza, A. (2003). Value and application of the ICF in rehabilitation medicine. *Disability and Rehabilitation*, 25(11–12), 628–634.
- Ustün, T. B., Chatterji, S., Bickenbach, N., & Schneider, M. (2003). The International Classification of Functioning, Disability and Health: a new tool for understanding disability and health. *Disability and Rehabilitation*, 24, 565–571.
- Verbrugge, L., & Jette, A. (1993). The disablement process. *Social Science & Medicine*, 38(1), 1–14.
- Ware, J. J., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, 30, 473–483.

- Weathers, F., Litz, B., Herman, D., Huska, J., & Keane, T. (1993, October). *The PTSD Checklist (PCL): reliability, validity, and diagnostic utility*. Paper presented at the Annual Convention of the International Society for Traumatic Stress Studies, San Antonio, TX.
- WHODAS Home Page. (2010). <http://www.who.int/icidh/whodas/FAQ.html>.
- Wolf, P. S. A., & Jacobs, W. J. (2010). GPS technology and human psychological research: a methodological proposal. *Journal of Methods and Measurement in the Social Sciences*, 1(1), 1–15.
- World Health Organization. (WHO). (1980). *International Classification of Impairments, Disabilities, and Handicaps*. Geneva: World Health Organization.
- World Health Organization (WHO). (2001). *International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization.

Chapter 19

Injury Costing Frameworks

David Bishai and Abdulgafoor M. Bachani

Introduction

From the perspective of economics, the most appropriate rationale for measuring the cost of a disease or injury is to support a decision about how best to control it. Cost information assists choices by allowing decision-makers to explore the cost consequences of the various options that might be chosen. Knowing injury costs can help decision-makers compare the costs that can be prevented after choosing an intervention against the cost of that intervention. In other words, establishing the cost of an injury is just one step toward a full cost effectiveness analysis of a specific injury countermeasure. (The other step is knowing how effective the countermeasure will be). Where there is not yet a well-defined injury countermeasure, efforts to measure the cost of injuries can still be valuable from a “*what if*” perspective. In a “what if” analysis one answers, “*What* would the cost effectiveness be *if* a hypothetical intervention had a stipulated level of impact?” This can help engineers who are developing preventive strategies understand the breakpoint intervention cost beyond which an injury prevention strategy would be unlikely to be cost-effective.¹

Many studies that estimate the cost of injury or disease are divorced from countermeasures and outside of a decision-making framework. Advocacy is the stated rationale for these studies. The presumption is that if one can say “Injury X costs this society \$Y millions (or billions)” then policymakers will spend money on injury X. Authors of advocacy studies often resort to them after frustration with policymakers who have been unmotivated by descriptive epidemiological data about the numbers of people killed or maimed by injuries. There is (or ought to be) some ambivalence about this use of economic analysis. Unless there is evidence that spending money on solving or developing solutions for a problem will result in a cost effective reduction in the problem, the statement that a problem costs \$millions or \$billions is actually irrelevant.

¹ Knowing the cost of the prevented injuries would help one know the maximum one should spend on safety. If the cost of particular safety investments exceeds the comprehensive costs of the injuries to be prevented, then money spent on safety would be better spent on other things that can do more to improve the human experience.

D. Bishai, MD, PhD, MPH (✉)

Center for Injury Research and Policy, and International Injury Research Unit, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Suite E4622, Baltimore, MD 21205, USA
e-mail: dbishai@jhsph.edu

A.M. Bachani, PhD, MHS

International Injury Research Unit, Health Systems Program, Department of International Health, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Suite E8132, Baltimore, MD 21205, USA
e-mail: abachani@jhsph.edu

Using cost of injury studies for advocacy requires additional caution to guard against misuses of data. Investigators undertaking an advocacy study can suffer from bias that leads to overestimates of the cost of a problem. Decision-makers can potentially detect these upward biases and lose confidence in the results. As biased studies of the cost of injury accumulate, they could lead to decreased confidence in this type of investigation more generally. Advocacy projects can also tempt the investigator to prepare meaningless comparisons of the costs of one injury or disease to another. From a decision-making framework, the highest cost condition does not automatically merit the highest spending for prevention; what matters is the relationship between what is spent and how much could be saved because of that spending.

Basic Frameworks for Costing Disease and Injury

There are three related conceptual frameworks that could be used to guide a costing exercise. All three share a focus on depicting a counterfactual world without the injury. The three frameworks are *human capital*, *willingness to pay*, and the *general equilibrium* framework. A common concern for each approach is establishing the right time horizon for the analysis, which is related to the duration of injury sequelae (Drummond et al. 1997). It is also important for every study of the cost of injury to express an uncertainty range around the estimates. It is not as helpful to produce single point estimates of the cost of injury, as it is to help illuminate the distribution. Health costs are often skewed with a long right tail in which the severely injured have very high costs. Using averages to help predict expected future costs in a skewed sample can be misleading. An insurer who bears all of the injury costs for a population cannot ignore the rare cases with high costs. In a small sample, the statistical average is a poor guide to what could actually occur. Average cost only becomes a reliable guide to actual costs in a population that is very large – on the order of 10,000 s. It is often best to communicate the entire distribution of costs using histograms, deciles, or quartiles in addition to the mean.

Human Capital Framework

In the human capital approach, the analyst prepares a stylized model of the economic consequences of an injury (Rice 1967; Rice and MacKenzie 1989). This stylized model identifies the losses due to an injury as the sum of direct medical costs, indirect lost productivity costs, and intangible psychological costs of pain and suffering. Economists note that the human capital framework cannot be derived from a foundation in their utility theory. This objection has done little to deter human capital estimates from continuing to be produced and used. The approach appeals to many policy makers because it partitions injury costs into neat compartments.

The compartments can be seen in the following human capital equation:

$$\text{Cost of Illness} = \text{Cost}_{\text{Medical}} + \text{Cost}_{\text{Productivity}} + \text{Cost}_{\text{Property Loss}} + \text{Cost}_{\text{Pain and suffering}} \quad (19.1)$$

For injuries one might extend the model to include costs of police and fire services. The advantages of the human capital method are that it offers straightforward guidance for measurement of the first three ingredients of the cost model (in practice, the intangible costs are frequently ignored in the human capital model and the analyst notes this limitation). The medical care costs can be measured empirically in health systems where there are fee schedules for the various injury treatments. A fee schedule is a listing of medical services and procedures and the charges that the insurer has negotiated that it is willing to pay. Fee schedules usually code the services using Current Procedural Terminology

(CPT) or International Classification of Diseases Procedure Coding Systems (ICD-PCS). Utmost care should be devoted to understanding whether the fee schedule is listing “charges” or “allowable charges” or “costs”. Charges are amounts that health care providers ask for and offer a large overstatement of the true cost that the provider incurred. Allowable charges have been negotiated downward by a health insurance entity and may also overstate the true cost – it depends on how aggressive the negotiators were. In the USA, Medicaid’s fee schedules are considered to have been negotiated down to an approximate reflection of costs. In other high- and middle-income countries, large government payment systems have similar fee schedules. For low-income countries, fee schedules do not exist and one must resort to top down and bottom up costing methods. A full description of bottom up/top down costing is beyond our scope but good sources abound (Miller 2000; Waters 2000).

For example if one had fee schedules to calculate costs of various inpatient events (bed days, procedures) and various outpatient costs (visits, procedures) one could fill in the first term in (19.1) as:

$$\text{Cost}_{\text{Medical}} = \sum \beta^t (\text{Inpatient Costs}_t + \text{Outpatient Costs}_t), \quad (19.2)$$

where β (raised to the t th power) is a discount factor ranging between 0 and 1 and subscript t marks out the time period in which the costs are observed. Summation is taken across a time horizon from $t=0$ to $t=T$.

$$\text{Cost}_{\text{Productivity}} = \sum \beta^t [(\text{Full Disability Days}_t \times \text{Wage}_t) + (\text{Partial Disability Days}_t \times \text{Degree of Disability} \times \text{Wage}_t)]. \quad (19.3)$$

Lost productivity can be estimated with descriptive epidemiological data about the age at death, the duration of full and partial disability and the wages of the injury victims. Individuals whose productive work is child care or homemaking can also have lost productivity. One can apply the wage one would have to pay to acquire domestic services in the market or compute the average earnings of individuals with similar age, gender, and educations. One can still use (19.3) for people who die of their injuries by counting death as the equivalent of full disability and having a time horizon from the age at death until the life expectancy at the age of death.

$$\text{Cost}_{\text{Property Loss}} = \sum (\text{Item lost} \times \text{Replacement or Repair Cost}). \quad (19.4)$$

Injuries sometimes are accompanied by damage or destruction of property. Lost property occurs almost exclusively in the first year and should be valued at its replacement or repair cost. Property costs are not always relevant in evaluating an injury countermeasure because some measures prevent only human injury and do not prevent property damage. Examples are seatbelts and smoke detectors.

$$\text{Cost}_{\text{Pain and Suffering}} = \sum \beta^t [\text{Per Period Suffering Costs}_t]. \quad (19.5)$$

Strategies for measuring pain and suffering (also known as intangible costs) in the human capital model are not fully developed. Conceptually, the costs of pain and suffering belong in the human capital model because the prospect of an injury is something people want to avoid independent of medical bills and lost wages. In qualitative statements of why they avoid risks, people often refer to fear of pain; they also refer to concerns of how their injuries might impact suffering within their family.

Most studies of the cost of injury exclude costs of pain and suffering and simply note this limitation. Because some studies measure pain and suffering while most do not, it is important to bear these methodological differences in mind when comparing costs of injury across different studies. It would be unfortunate if advocacy groups paid additional attention to one type of injury as opposed to another simply because the methods used to estimate the cost of injury added on costs of pain and suffering for one and not the other.

For studies that include costs of pain and suffering, the approaches most commonly used involve a systematic study of jury awards or the use of willingness to pay studies where respondents are guided to reveal monetary values on only the pain and suffering involved in an injury. In studying jury award data, analysts examine these data and attempt to parse out how much of each award is designated in the “pain and suffering” category (Cohen and Miller 2003). One requires large sample sizes of awards for comparable injuries to solve the problem of high variability in juries and the emotional factors that could drive awards. Willingness-to-pay approaches require attention to ensuring that the respondents are envisioning only the pain and suffering when stating what they would pay. If they are also paying to avoid lost wages, then the estimate would double-count productivity in confounding it with pain and suffering.

An example of the human capital approach is worked out for the case of road crash injuries in Egypt in Appendix 1.

As one can see from Appendix 1, the human capital method can leave out a lot of the important factors that seem to many to be part of the cost of injuries. The calculations for Egypt left out the intangible costs of suffering and the costs of death for those who die. They left out the cost to the Egyptian government of responding to car crashes. One approach would be to shore up the human capital estimates for Egypt by attempting to add back these costs using heroic assumptions to obtain numbers where there is limited published information. Another approach would be to flag the limitations and note that the \$5.49 injury cost per Egyptian is an underestimate. There are pros and cons of either approach. Relying on unsupported assumptions threatens the validity of the estimate and could shake the confidence of policymakers. Leaving out injury costs might lead to insufficient attention being paid – especially if the costs that are left out are likely to be large.

Willingness to Pay Approach

An alternative to the human capital method is to consider what people would pay to live in a world with a lower risk of that injury. The advantage of this approach is that it places monetary values on injury that are grounded in the consumer’s own preferences; furthermore, it is one of the few options for including estimates of the value of pain and suffering. The disadvantages of the approach are the need for a thorough description of the prospect that the consumer is being asked to evaluate. There are two options in implementing this approach: *revealed preference* relies on observing people paying money to purchase measured decrements in the risk of injury; *stated preference* records stated willingness to purchase decrements to the risk of injury using survey methods.

Markets where people can pay extra money to lower their risk of injury include markets for consumer products that vary in their risk of injuring, and labor markets where laborers can and do negotiate higher wages for more dangerous jobs. The basic analytic strategy requires a large dataset where people are facing a variety of prices for a variety of injury risks. A classic example is a study of the market for safer automobiles where each sale price was regressed against each car model’s fatality rate, price, fuel efficiency, etc. The statistical correlation between the price paid and the fatality rate can inform economists on the monetary value users place on avoiding a fatal crash and, by extension, on avoiding the possible medical costs, lost productivity, pain and suffering and all other aspects of what goes along with a safer car (Miller 1990; Blomquist et al. 1996). Another genre of literature looks at labor markets, at what economists call “compensating variations” in the amount laborers are paid, and their risk of occupational fatality (Smith 1974; Thaler and Rosen 1976). Compensating variation studies bred a vibrant strain in economics, seeking to place a dollar value on small changes in the risk of occupational fatality. This value – of purchasing a small change in the probability of death – is referred to as the “value of statistical life” (Viscusi 1993;

Kniesner et al. 2010). If one will pay \$5.00 to purchase a 1 in a million decrement in death risk then one is said to have a value of statistical life of \$5 million.

The revealed preference approach to determining the willingness to pay to prevent injury is, however, limited in its ability to focus on the various aspects of injury, especially nonfatal injury. This is because of uncertainty in how the risks of consumer products in the market are actually connected to the multiple varieties of nonlethal injury. Although one can measure fatality risks and tie them to market choices such as buying a car or choosing an occupation, it is far more difficult to measure risks of the multiple nonlethal injury and other consequences that can ensue from a market choice. This limitation is corrected in survey approaches where one can parse out the monetary values surrounding various hypothetical consequences of an injury. Using a survey approach frees the investigator from being trapped by the cross-correlation of multiple related consequences from human choice. By setting up a series of hypothetical choices that focus more precisely on the consequences of interest the investigator can understand how people value just those consequences.

It is beyond our scope to offer details on designing survey-based research in this area. There are excellent texts offering an introduction (Champ et al. 2003; Hensher et al. 2005). Briefly, survey-based approaches to determining the value placed on outcomes from a choice consist of three parts: The investigator first uses a vignette to describe an aspect of injury that needs to be evaluated (Part 1); the respondent is offered a payment mechanism that offers a back-story to why they might be asked to pay for an altered risk of injury (Part 2); finally the respondent is presented various options that vary both in their risk of leading to this injury experience and in their price (Part 3). For example, if one wanted to evaluate what people were willing to pay to avoid visual loss stemming from injury one could use a setting such as that in Appendix 2.

Suppose a respondent encountering Appendix 2 selected the less dangerous prospect of an annual risk of blindness that was 1/1,000 lower, forsaking an additional \$1,000 in salary. This would imply a willingness to pay \$1,000,000 to avoid a 100% chance of blindness. They might pay even more than this, so by watching the applicant make this tradeoff for different values where s/he is asked to forsake more money for safety, one could potentially identify the person's maximal willingness to pay.

General Equilibrium Approach

Looking ahead for future developments in methods to understand the cost of injury, general equilibrium methods loom as the next big step. Societies are complex aggregates of individuals with emergent properties. Things happen in the macro perspective that cannot be accounted for by assessing the costs of injury one injured person at a time. From a macroeconomics perspective, one can ask questions about how a world with a lower rate of injury might be substantially different and how this might affect economic performance, migration patterns, and investments in precautions. In the general equilibrium approach, the analyst prepares a rich model of how the world adapts to the absence/presence of injury over a long horizon.

For example, suppose there were two low-income countries with an educated workforce, good telecommunications, and a stable legal infrastructure. Let us suppose that the two countries have a twofold difference in the probability of traffic crashes.

Now picture a boardroom in a multinational corporation where a firm is considering developing investments in one of these two countries. Because employees of the firm will have to travel to this country on a regular basis to conduct business, the injury rate of the country can become a deciding factor in a decision to invest. Multiply decisions like this across multiple corporations and one could envision substantial effects of injury rates on a country's access to foreign direct investment and, by extension, the overall rate of economic growth. These are very real costs of

injury and they are benefits of injury reduction that need to be considered.² Ignoring these aspects of policy could lead to underinvestment in safety as part of a low-income country's overall development strategy.

On a similar note, once a country begins to assign customers the right to expect safe products and develops regulatory capacity to ensure safe products, it develops a whole new economic sector – the safety sector. This sector is based on supplying safety, which is a commodity valued by citizens, and the creation of this new commodity involves the creation of new jobs and new goods and services that hitherto had not been produced. Many of these jobs might be in the government sector in the form of safety regulation, but some jobs could also be created in the private sector as firms hire safety officers. The new commodity produced by the safety officers and regulators is a contribution to the GDP and a very positive form of GDP growth.³

The strategies for actually measuring the macroeconomic and emergent costs of injury typically rely on a technique of simulation modeling called computable general equilibrium (CGE) models. Although these models have been produced to evaluate the macroeconomic impact of HIV/AIDS control strategies, they have not yet been applied to injury control. These simulations depict multiple economic sectors such as households, firms, government, health care, and transportation. Equations predict the outputs of each sector as a function of the outputs of every other sector. In macroeconomic policy analysis, the model is used to predict outcomes as a function of interest rates, free trade, or labor policy. In health policy analysis, one could undertake policy simulations of disease simulating impacts on labor supply and economic output as a function of changes in a prevalent disease or injury. Readers interested in this approach should consult WHO, *Guide To Identifying The Economic Consequences Of Disease And Injury* (World Health Organization 2009).

In short, traditional cost-of-illness studies employ a static, partial, and inconsistent approach to estimating the macroeconomic impact of disease and injury at the societal level. A more general and dynamic assessment of the present value of forgone consumption opportunities is required as a critical element of future research.

Summary

From advocacy to assisting decision-makers choose between different injury prevention strategies, injury costing is playing an increasingly important role. Methods for estimating the cost of injuries have thus also gained prominence in the field. This chapter describes three approaches commonly used to guide costing exercises for diseases or injury – the human capital, willingness to pay, and general equilibrium framework – and offers a brief guide to how one would go about costing injuries. The human capital framework identifies the losses due to an injury as the sum of direct medical costs, indirect lost productivity costs, and intangible psychological costs of pain and suffering.

²Countries could also attract foreign direct investment through a race to the bottom. If they offer an institutional infrastructure that offers workers no recompense for occupational injury, then they allow multinational firms to move dangerous jobs into these poorly regulated environments. The additional economic benefits of jobs created by the race to the bottom are perverse benefits from allowing a high rate of occupational fatality. They are artificial from the social perspective, which would have to balance the gains to the economy against the workers' losses from occupational injury costs that the international firm has offloaded to poor and vulnerable workers.

³Paradoxically, because medical services are a component of GDP, successful safety officers could (hopefully) lower GDP by replacing the high-priced services of trauma surgeons and emergency personnel with the lower-priced preventive services.

While this framework offers a straightforward way to measure the tangible costs (medical, productivity, and property losses), strategies for measuring pain and suffering are not fully developed. The willingness-to-pay approach, on the other hand, considers what people would pay to live in a world with a lower risk of that injury. The advantage of this approach is that it places monetary values on injury that are grounded in the consumer's own preferences and it is one of the few options for including estimates of the value of pain and suffering. While both of these approaches provide reasonable estimations of the cost of injuries, they do so one injured person at a time, and cannot take into account things that happen in the macro perspective. Societies are complex aggregates of individuals with emergent properties, and from a macroeconomics perspective one can ask questions about how a world with a lower rate of injury might be substantially different and how this might affect economic performance, migration patterns, and investments in precautions. The general equilibrium approach takes this into account, providing strategies for actually measuring the macroeconomic and emergent costs of injury through a technique of simulation modeling called CGE models. Although these models have been produced to evaluate the macroeconomic impact of HIV/AIDS control strategies, they have not yet been applied to injury control. A more general and dynamic assessment of the present value of forgone consumption opportunities would be a desirable element of future research on measuring the economic impact of injuries on a national or global scale.

Appendix 1: A Cost of Injury Calculation

In 2009, the Egyptian Ministry of Health conducted a household survey to measure the burden of road injuries (Egyptian Ministry of Health 2009). The respondents who had injuries were asked to describe the health care utilization and lost work and productivity that they endured. These parameters are listed below.

The Egyptian data revealed an annual age-adjusted rate of injuries of 1,271 per 100,000 population. Taking a hypothetical population of 1 million, this translates to a total of 12,170 injuries of all kinds. Road traffic injuries accounted for 35% of this burden, thus amounting to a total of 4,449 road traffic injuries. The data also reveal some of the details involved in estimating cost components. For example, estimated inpatient costs are the product of the probability of inpatient care \times cost for those hospitalized, with an analogous expression for outpatient costs.

Parameter for patients who have experienced a road injury	Value	Distribution (interquartile range)
Probability of inpatient care	$A = 19.64\%$	
Cost per inpatient stay (US\$) ^a	$B = \$1,547$	168–1,350
Probability of outpatient care	$C = 65.87\%$	
Cost per outpatient visit (US\$)	$D = \$41$	19–60
Probability of lost work	$E = 48\%$	
Duration of lost work for those who lose work	$F = 39$ days	7–45
Median wage (US\$)	$G = 30$ /day	
Probability of property loss	$H = 24\%$	
Value of lost property for those who lose property (US\$)	$I = \$1,350$	443–5,063

Source: For all parameters except cost per inpatient stay is (Egyptian Ministry of Health 2009). For cost of inpatient stay is (World Health Organization 2011)

^aMean length of Inpatient stay = 18.49 days (median 7 days; IQR: 2–16); Cost per day of in-patient care = \$84.35/day (\$31.24 \$International). All costs in table are in US Dollars = \$0.37 International in 2009

One can use the lettered parameters and the equations in the text to calculate the various costs of injury as follows:

	Median	Total	25%ile	75%ile
Property loss	Total RTI $\times H \times I = 4,449 \times 0.24 \times 1,350$	\$1,441,476	\$473,000	\$5,406,000
Lost productivity	Total RTI $\times E \times F \times G = 4,449 \times 0.48 \times 39 \times 30$	\$2,473,573	\$444,000	\$2,854,000
Medical costs	Total RTI $[(A \times B) + (C \times D)] = 4,449$ $[(0.2 \times 1,547) + (0.66 \times 41)]$	\$1,495,532	\$206,000	\$1,377,000
Total		\$5,411,581	\$1,122,000	\$9,637,000

From this exercise, one finds that the total cost of road injuries in Egypt amounts to approximately \$5.4 million per year for a population of 1 million individuals, translating to a cost of \$5.41 per person per year. If one were pursuing an advocacy strategy one would multiply \$5.41 times the 79 million population count of Egypt to announce that road injuries cost Egypt \$427 million and compare this to some other financial statistic like health spending (\$8 billion) or overseas development assistance (\$1.3 billion).⁴

From a planning perspective, one could compare how much the \$5.41 injury cost per citizen might be reduced against the costs of better road safety enforcement to reduce it. If better enforcement could be achieved by deploying one more police officer per 10,000 citizens and if this could reduce crashes by 21%, then one could justify spending \$30 (median daily wage) times 365 days = \$11,000 to balance against the roughly \$1.10 per citizen saved (21% \times \$5.41).

Appendix 2: Sample of a Stated Preference Questionnaire Might Begin

Part 1: Description of blindness: Imagine only being able to see dark or light but not being able to make out faces, read books, or watch television. Walking safely would require you to learn to use a cane or guide dog. You would not be in any pain.

Part 2: Suppose your son were weighing two job offers that both included rewarding work, good benefits, and opportunities for promotion. Both jobs are the same in every way except how much they pay and the risk that an on-the-job blinding injury would occur. Which option would you advise your son to take?

Part 3: Choose the better option

Attribute	Option 1	Option 2
Chance of blindness after 1 year on the job	1%	1.1%
Annual salary	\$50,000	\$51,000

This choice would iterate about 20 or so times per respondent using various values

References

Blomquist, G., Miller, T. R., et al. (1996). Values of risk reduction implied by motorist use of protection equipment: new evidence from different populations. *Journal of Transport Economics and Policy*, 30(1), 55–66.

Champ, P., Boyle, K. J., et al. (2003). *A primer on nonmarket valuation*. Boston, MA: Kluwer.

Cohen, M., & Miller, T. R. (2003). Willingness to award non-monetary damages and the implied value of life from jury awards. *International Review of Law and Economics*, 23, 165–181.

⁴Both health spending and development assistance figures are in current dollars for 2008 and are taken from <http://data.worldbank.org/indicator>.

- Drummond, M., O'Brien, B., et al. (1997). *Methods for the economic evaluation of health care programmes* (Vol. 2). London: Oxford University Press.
- Egyptian Ministry of Health. (2009). National Community Based Injury Survey in Egypt. Cairo: MOH.
- Hensher, D., Rose, J. M., et al. (2005). *Applied choice analysis: a primer*. New York: Cambridge University Press.
- Kniesner, T., Viscusi, J., Kip, W., et al. (2010). Policy relevant heterogeneity in the value of a statistical life: evidence from panel quantile regressions. *Journal of Risk and Uncertainty*, 40(1), 15–31.
- Miller, T. R. (1990). The plausible range for the value of life: Red herrings among the Mackerel. *Journal of Forensic Economics*, 3(3), 75–89.
- Miller, T. R. (2000). Assessing the burden of injury: Progress and pitfalls. In D. Mohan & G. Tiwari (Eds.), *Injury prevention and control* (pp. 51–70). New York: Taylor and Francis.
- Rice, D. (1967). Estimating the cost of illness. *American Journal of Public Health*, 57(3), 424–440.
- Rice, D., & MacKenzie, E. (1989). *Cost of injury in the United States: A report to congress*. San Francisco, CA: University of California.
- Smith, R. S. (1974). The feasibility of an “Injury Tax” approach to occupational safety. *Law and Contemporary Problems*, 38(4), 730–744.
- Thaler, R., & Rosen, S. (1976). The value of saving a life: Evidence from the labor market. In N. E. Terleckyj (Ed.), *Household production and consumption* (pp. 265–298). New York: National Bureau of Economic Research.
- Viscusi, W. K. (1993). The value of risks to life and health. *Journal of Economic Literature*, 31, 1912–1946.
- Waters, H. (2000). *The costing of community maternal and child health interventions*. Washington, DC: US Agency for International Development.
- World Health Organization. (2009). *Guide to identifying the economic consequences of disease and injury*. Geneva: WHO.
- World Health organization. (2011). Choosing Interventions that are Cost Effective (WHO-CHOICE). <http://www.who.int/choice/en/>. Accessed on 27 Oct 2011.

Part IV
Analytical Approaches

Chapter 20

Statistical Considerations

Shrikant I. Bangdiwala and Baishakhi Banerjee Taylor

Introduction

An understanding of how the discipline of statistics, and specifically, some of its principles and concepts play an integral part in our scientific endeavors of acquiring knowledge, is an essential requirement for researchers in all fields. This chapter does not pretend to be exhaustive on all the possible statistical considerations that may impact injury prevention and control research, but mainly to highlight a few key issues. To provide some thread of continuity to the presentation, a common injury problem is considered: motor vehicle-related “accidents.” Motor vehicle-related injuries are the most common morbidity and mortality reason in young adults worldwide (WHO 2011). Most individuals, whether researchers or not, have “an understanding” of the problem, and many may even venture to state the causes and potential solutions. This chapter uses the “motor vehicle-related accidents” problem to illustrate how the application of the principles and concepts of statistics can help one to have a better or an “educated” understanding of the problem. Statistics focuses on quantifications or measurements. Thus, four statistical quantification considerations are specifically addressed here – quantifying uncertainty, quantifying probability, quantifying risk and exposure, and quantifying the strength of relationships – all of which are essential elements of injury prevention and control research.

The Role of Statistics in Quantifying Uncertainty

The field of injury prevention and control is not different from other research fields that attempt to understand the complexities of reality. It is also not different from other fields that often fail to consider the role of uncertainty in their scientific pursuits. Uncertainty arises from the inability to observe

S.I. Bangdiwala, PhD (✉)
Department of Biostatistics and Injury Prevention Research Center,
University of North Carolina, 137 E. Franklin Street, Suite 203, Chapel Hill, NC 27514, USA
e-mail: kant@unc.edu

B.B. Taylor, PhD
Trinity College of Arts and Sciences, Duke University, 114 AAC, Box 90697, Durham, NC 27708, USA
e-mail: bbane2@duke.edu

the truth. When trying to understand the “motor vehicle-related accident” problem by observation and measurement, what is observed and what is measured may differ from the truth because of process variability, sampling “error,” or simply “by chance.”

Uncertainty from Process Variability

Defining “motor vehicle-related accident” first needs to consider its definition – does it involve dealing with incidents involving only motorized vehicles or will it also refer to a pedestrian or a cyclist being hit by a motorized vehicle?; even if only motorized vehicles, is it dealing with frontal collisions, rollovers, single-vehicle against an object? These definitional issues are not the purview of statistics, but do often lead to measuring very disparate incidents as if the same (see section below on “multiple multiplicities”).

Assuming that the definition under consideration is “collisions of only two motorized vehicles,” such incidents are subject to considerable variability. Each “collision” may vary based on type of vehicles, speed and forces at impact, angle of impact, to give a few examples. These other factors (variables) affect the values of the measurements or observations of a collision. The variability observed in the measurements of the outcomes of interest from the collision is a source of uncertainty. Statistical methods can help researchers understand the various sources of variability in their observations, but they require that many observations be made of the same event so as to understand the distribution of the measurements. Even though when it is known that every incident is unique, it can be assumed that they come from a “population” of similar incidents. By studying the distribution of the outcome in the population of similar incidents, one can gain an understanding of the variability in the outcome.

Uncertainty from Sampling and Incompleteness of Information

In addition to process variability as a source of uncertainty, research is subject to the uncertainty from not being able to observe the entire population of incidents, i.e., the so-called “sampling error.” One is rarely in the position to observe the complete population of events, unless one conducts a census or has a registry surveillance system that records all incidents. In such rare situations when one can observe the entire population, statistics as a science of dealing with uncertainty does not enter the picture. The aspect of statistics of summarizing and presenting quantitative information does, so that one may choose to present means and standard deviations of the variables measured, but not inferential summarizations of the distribution such as standard errors or make probability statements by providing confidence intervals or p values. In some countries, one can assume that complete observation of the population may be possible for certain types of outcomes, such as deaths from certain specific events (e.g., deaths in the USA due to collisions involving more than one vehicle), but in general, one normally observes a sample from the population of interest. Thus uncertainty creeps in once again. This chapter now focuses on how *representative* that observed sample may be of the entire population, or more specifically, how closely does the observed empirical distribution of the outcome of interest in the sample match the probability distribution of the outcome in the population.

To quantify the uncertainty from sampling, statisticians rely on methods of inference based on assumptions about the population’s probability distributions. Since these are not observed or known, statisticians’ conclusions are stated in terms of probability statements. Thus, when providing an estimate of a population parameter such as the mean of the outcome, we construct an interval estimate called a “confidence interval,” which provides a range of values from a *lower value* to an *upper value*. We then make statements about having a certain amount of confidence (typically 95%) it contains the true unknown population parameter value.

Table 20.1 Excerpt from Table 20.1 of Weiss et al. (2010) – patient characteristics for youth motorcycle-related hospitalizations in USA in 2006

Characteristics	<i>n</i>	Proportion, %	95% CI
Total estimated number of cases	5,662		
Gender			
• Male	5,016	89.7	(88.7–90.8)
• Female	575	10.3	(9.2–11.3)
Age group			
• 12–14 years	1,023	18.1	(16.3–19.8)
• 15–17 years	1,797	31.7	(30.1–33.4)
• 18–20 years	2,840	50.2	(47.9–52.4)

Weiss et al. (2010) studied youth motorcycle-related hospitalizations in the USA in 2006, but since their estimates were based on a sample of discharges, they reported estimates of proportions of cases by gender and age group along with corresponding 95% confidence intervals (see Table 20.1).

When testing a specific hypothesis related to the population mean of the outcome as the method of inference, we account for the uncertainty of sampling by making statements such as: “we believe that the true mean of the outcome in the population is greater than the hypothesized value because when we assume that the true mean in the population is the hypothesized value, the estimated probability of observing a sample mean equal to or greater than the one we observed is too small.” This probability is called the “*p* value.” How “small” is small is an arbitrary choice, but commonly a probability of less than 0.05 is by tradition considered “small.” The term “statistically significant” is used to denote a result that has a small (of <0.05) probability to be due to chance. Often one is not as careful in one’s statements, and shortens them to “the mean is significantly ($p < 0.05$) greater than the hypothesized value.” “The term ‘significance’ is a technical term in statistics used to determine the crossing of an arbitrary probability threshold in the process of making an inference from a sample characteristic to a population characteristic through a formal test of a hypothesis” (Bangdiwala 2009a).

Weiss et al. (2010) examined the hypothesis of whether the gender distribution was different or not by crash location. They found that in traffic, males were involved in 88.6% of crashes, and in non-traffic, males were involved in 92.1% of crashes in their sample. When they tested whether the true difference in proportions in the population is likely to be zero or not, based on the sample observed difference and on assumptions of the statistical properties of proportions, they found a *p* value that was quite small, less than 0.001, and thus felt comfortable stating that they believe that the true difference in proportions in the populations is likely to not be zero.

Uncertainty from Chance and Randomness

Chance as a source of uncertainty has baffled mankind as a concept. Two exactly same vehicles manufactured exactly the same way in the same place and at the same time, and driven in a test laboratory under exactly the same conditions, will not necessarily have the exact same mechanical failure. What is chance? Statisticians attempt to quantify chance, by posing probability models for the occurrences. When one says that the “vehicle has a 25% chance of failing to stop in 50 m from a speed of 80 kph,” it means that out of 100 attempts under the exact same conditions, the same vehicle would stop 75 times in 50 m, and that 25 times it will not stop in 50 m. However, often models are postulated and not derived from scientific theory, or experimental or observational data. Quantifying probability is elaborated in the next section.

Quantifying Probability Theoretically

While quantifying risk from observations is problematic due to the feature of rare events and the difficulties of quantifying exposure, one can try to use theoretical derivations to quantify these probabilities. Statistics, like all sciences that try to understand nature, relies on theoretical assumptions and models to simplify the complexities observed in nature. For example, the observed or empirical distribution of outcomes from collisions of only two motorized vehicles may be obtained from a sample, but that is subject to uncertainties. However, using arguments and theories derived from physics, mathematics, engineering, and other sciences, one can postulate a theoretical distribution for the outcomes. The possible values and frequency of occurrence of the outcome is thus described by its theoretical probability distribution. Several well-known theoretical probability distributions have been postulated for studying injuries or incidents. They are based on the fact that injuries or incidents are counted and can take on only the nonnegative integer values of 0, 1, 2, 3, ... – a discrete probability distribution. Such probability distributions are specified by providing the possible values and the probability that they occur, with the condition that the sum of all probabilities is 1.

Discrete Distributions for Count Variables

A common theoretical probability distribution that can be applied to count data is the one proposed by Poisson to describe the distribution of the number of occurrences of an event in a given time interval (Bangdiwala 2010). It was proposed as a distribution for rare events, and thus it is considered as a good theoretical option to model injury count data. The Poisson probability distribution is characterized by the following equation:

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, 3, \dots,$$

where λ is the mean value of the distribution and e is the constant base of the natural logarithm (2.71828...). The properties of this theoretical distribution are that its mean λ equals its variance λ and it is right-skewed. Figure 20.1 illustrates the theoretical Poisson distribution function for a specific value of the mean $\lambda=0.6$.

The usefulness of using a theoretical probability distribution such as the Poisson is that one can calculate the predicted probabilities of the number of occurrences of the event based on the theoretical probabilities. Another use is in studying the effect of some other variables on the number of occurrences of the event. For example, there may be other factors such as average age of drivers, road conditions, and so forth, which may explain the number of collisions of only two motorized vehicles in a given community in a given year. Thus, a regression modeling approach is needed. In this case, a standard multiple regression model that assumes a Normal (bell-shaped)-dependent variable is not appropriate, and a Poisson regression model (Kleinbaum et al. 2007) can be used.

Other Discrete Distributions for Count Variables

The Poisson distribution may not always be a good fit for a count variable. This may occur for a variety of reasons. For example, the variance may be larger than the mean, which is called *over-dispersion*. The distribution function that is often better suited for over-dispersed data is the right-skewed *negative binomial* distribution. Another common situation is when an event is so rare, that most of the times a zero count is observed. In this case, an adjustment to the Poisson distribution may

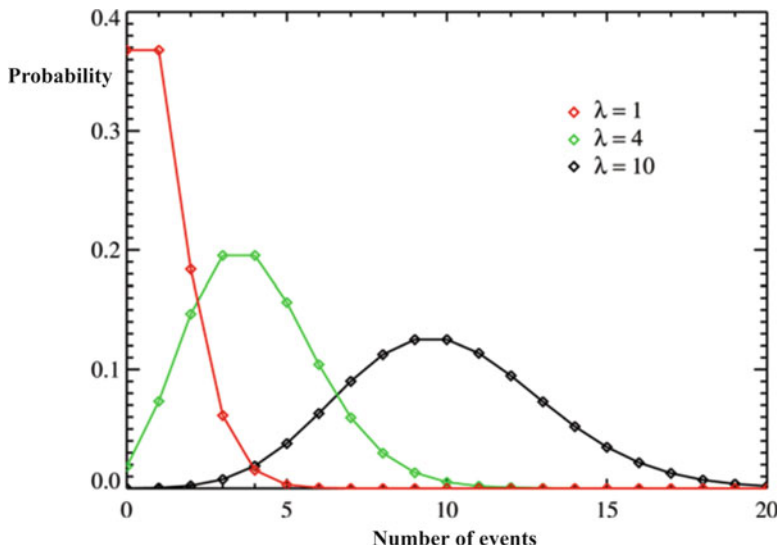


Fig. 20.1 Illustrative plots of Poisson distributions with means $\lambda=1, 4,$ and 10

work better, the so-called *zero-inflated Poisson*, where a combination of modeling the zeroes and the non-zeroes is used (Johnson et al. 2005).

Quantifying Risk and Exposure

As in most fields of research, there are unique characteristics that impact the quantification of variables of interest. In the injury field, there are three main “features – multiple multiplicities, rare events, and working on the extremes – impacting the two main quantification interests of risk and exposure.”

The Feature of Multiple Multiplicities

In the injury field, one commonly gathers information and studies aggregations of types and severities of injuries (Bangdiwala 2009b). Even for a limited and defined specific event as “collisions of only two motorized vehicles,” the types of injuries from such events can include all traumatic/acute transfers of energy, from fractures, penetrating or blunt object injuries, to burns and multiple trauma injuries. The common practice of aggregating all or several types of injuries has potentially serious statistical implications. Relationships between potential causal factors and the outcome may not be readily interpretable (see Sect. “Quantifying Relationships: Evaluating Effectiveness”).

The feature of multiple multiplicities arises because from a single incident such as a collision of only two motorized vehicles, there may be multiple individuals involved. Furthermore, each involved individual may have multiple injuries of multiple types (e.g., fracture, burn) at multiple sites of the body (e.g., upper extremities, torso), and with multiple severities (e.g., mild, moderate, severe). Having such multiplicities is potentially exciting from a statistical quantification standpoint. One may have k collisions, but j individuals with m total injuries, where $k \leq j$ and m is likely to be much bigger than j . So if one is studying “collision level variables,” information on individuals and on

injuries must be aggregated or summarized to the collision level; or if one is studying “individuals,” information on injuries must be aggregated or summarized to the individual level. In addition, if studying injuries, one must account for the fact that the multiple injuries in a given individual are related statistically (correlated) because of individual and collision characteristics; and when studying an outcome on individuals, one must account for the fact that an individual’s overall outcome is correlated with the outcome of other individuals involved in the same collision! These correlations, from the hierarchical nature of the data, can be quantified using traditional variance component analyses if the data are “balanced and complete,” and also by using more advanced regression models such as mixed effects multiple regression models and generalized estimating equation (GEE) models. Failure to account for these correlations in the design may lead to underpowered studies, and when not considered in the analysis, one can obtain false-positive results.

The Feature of Rare Events

Injuries and injury-producing incidents are rare events, relatively speaking, especially in comparison to other health incidents such as influenza or myocardial infarctions in older adults. The number of “collisions of only two motorized vehicles” in any given locale is few relative to the number of vehicles that are in circulation. This is a good thing from the society’s viewpoint, but it makes the use of certain statistical methods problematic. For example, if in a given community there were “ n ” number of collisions in a given year, a meaningful change in number of collision in another year would depend on the size of “ n .” For example, for $n=2,000$, a meaningful change may be say 100 (5% relative change), but for $n=50$, is a 5% (2.5) change meaningful? Thus, to demonstrate a meaningful effect, an intervention may have to be dramatic in absolute change as well as in relative (percent) change.

The small numbers affect also the ability to rule out chance in determining whether the observed changes are real or not. Relatively small communities can have a difficult time interpreting the results of their preventive actions. For example, if the number of injuries from crashes in a small community of ~5,000 persons was on average around 50/year, and some preventive actions are taken, judging effectiveness of the actions is not easy. If the next year they get 53 injuries, was the action ineffective or the difference could be explained by chance? If the number of injuries they get the next year is 45, is the action effective or is the difference explained by chance? This problematic choice arises because the events are “rare.”

The Feature of Working on the Extremes

From a purely common sense standpoint, it makes sense when dealing with a problematic situation to deal with the worse cases. Thus, for example, it makes sense that interventions be considered at those intersections in a community with the highest numbers of crashes since resources are limited. This, however, has statistical implications. Consider the following example, where we let C be the number of crashes per year at intersections as our variable of interest, and let N be the number of intersections that are “similar” in a given community. One is interested in the distribution of the variable C – it is a nonnegative count variable, possibly skewed as hopefully most intersections have a value of “0” but some may have a large value. Assume the true population mean is 3.4 accidents per year. In a given year, suppose we observe a random sample of $n=10$ intersections and get the following values of C : 8, 6, 4, 3, 3, 3, 2, 1, 0, 0. There is funding to intervene on the “most critical” black

Table 20.2 Driver age as a factor in collisions in 1999, Idaho, USA

Age	Drivers		Drivers in all collisions	
	Number	%	Number	%
15	8,334	0.9	488	1.1
16	15,366	1.7	1,521	3.6
17	17,550	2.0	2,084	4.9
18	18,481	2.1	2,085	4.9
19	18,212	2.1	1,930	4.5
20	17,537	2.0	1,554	3.7
21	17,450	2.0	1,420	3.3
22	17,323	2.0	1,263	3.0
23	16,397	1.9	1,063	2.5
24	15,238	1.7	1,041	2.4
25–34	153,815	17.5	7,918	18.6
35–44	179,778	20.4	7,229	17.0
45–54	161,779	18.4	5,488	12.9
55–64	102,960	11.7	3,093	7.3
65–74	70,950	8.1	1,866	4.4
75+	49,989	5.7	1,553	3.7
Not stated or other			923	2.2

Source: Idaho State Government (1999)

spots, so that we intervene in the 3 with 8, 6, and 4 accidents. Their mean is $(8+6+4)/3=6$, so the question is whether in the future their expected future mean with no intervention is it likely to be 6 or 3.4? The statistical correct answer is 6, because observations in a sample will tend to *regress to the mean* of their distribution. In our example, suppose we do intervene, and that after the intervention, we observe the following number of accidents in the 3 selected sites: 6, 3, and 0, so that the observed mean after the intervention in the 3 sites is $(6+3+0)/3=3$. The apparent relative efficiency is $(6-3)/6=0.50$ or 50%, while the real relative efficiency is $(3.4-3)/3.4=0.12$ or only 12%.

Quantifying Risk

Given that events are relatively rare, the process of quantifying the risk or probability of an event is not so easy. As has been mentioned, statisticians estimate the probability of an event by observing multiple occurrences of the same event in multiple opportunities. Thus, to calculate the probability that a standard six-sided die is fair, one must toss the die a large number of times and estimate the probability of each side by the proportion of number of times it occurred divided by the number of tosses. Thus, to calculate the risk of an 18-year-old driver being involved in a collision, he/she would have to be exposed to the exact conditions over a large number of times and count how many different vehicles from different places and under different driving conditions, and count the number of collisions they have had in a given time period, and then divide the count by some measure of exposure. From Table 20.2, we can estimate the risk for 18-year olds in 1999 in Idaho was $2,085/18,481=0.113$, while for 25–34-year olds it was $7,918/153,815=0.051$, less than half the risk.

Another question is whether “being a licensed driver” is the best measure of “being exposed to a collision.” Quantifying “exposure” is extremely difficult in the injury field. In other research fields, such as cardiovascular diseases or cancer, one commonly assumes that by “being alive” one is exposed to getting a heart attack or to getting cancer, but this is not reasonable in the injury field.

Quantifying “Exposure”: When Is One at Risk for an Injury

One cannot be at risk for an injury from a “collision of only two motorized vehicles” if one is not a driver or a passenger in a motorized vehicle. People that do not own or ever ride a motorized vehicle are not at risk. People that do are also not at risk when not engaged in the activity of riding or driving in a motorized vehicle. This characteristic makes “injury” as a disease, different from other diseases such as cancer or cardiovascular disease that one can assume one is constantly at risk or “exposed” to acquire the disease. Thus, the issue of quantifying the denominator of exposure is a difficult one. Bangdiwala et al. (1985) reviewed the use of common motor vehicle exposure quantifications and found that using population or registered vehicles as a measure of exposure is inadequate. They concluded that although often unavailable, better indicators are vehicle-kms or passenger-kms.

Quantifying Relationships: Evaluating Effectiveness

In establishing the causality of a relationship, there are seven considerations (Bangdiwala 2001):

1. *Strength of the association*, as quantified by some statistical measure of association. If the association measure is strong, chance can be ruled out.
2. *Dose–response effect*, i.e., the more of the causal factor, the larger the effect.
3. *No temporal ambiguity*, i.e., the disease follows exposure to the risk factor.
4. *Consistency of the findings*, as shown by external validity or confirmation in other studies.
5. *Biological plausibility*, i.e., the hypothesis is reasonable given what is known in science.
6. *Coherence of evidence*, i.e., consistency and plausibility of findings internally and externally.
7. *Specificity*, i.e., the causal factor causes the disease, and disease is due to the causal factor.”

The main consideration is to examine the strength of the association. When quantifying the association to establish effectiveness of an intervention to prevent or reduce incidents or injuries, one must consider the scale of measurement of the variables involved. Intervention (yes/no) is a binary discrete variable, while the outcome of injury or incidents is a count discrete variable. Thus it is necessary to understand how to quantify the strength of the association between two discrete variables, specifically when one is binary and the other is a count variable.

Comparing Count Variables Across Groups

The observed distribution of counts of the two groups can be presented in a 2-by- k contingency table, where k is the number of categories of counts observed. Measures such as odds ratios (OR) and relative risks (RR) are used to quantify the association. In Table 20.3a, we see how the OR and RR are calculated for a retrospective and a prospective study, respectively. The OR is a ratio of odds, and each odds is understood to be the ratio of the probability of an event occurring divided by the probability of the same event not occurring. Thus, each odds can be any nonnegative number. Thus, the ratio of odds (OR) is also nonnegative, and a value of 1 implies no difference in the odds of the event under one condition and another. The relative risk is also a ratio, but of the probabilities (risks) of an event occurring versus not occurring, and thus it also is nonnegative. The difference between an OR and an RR is actually in its definition – one is a ratio of odds and the other a ratio of probabilities. In Table 20.3a, we see how different they can be, especially when the probability of the event is large. What confuses many folks is the mathematical property due to the feature of “rare” events, as seen in Table 20.3b. With low incidence or probability of an event, the OR and RR for that event can be very similar.

Table 20.3 Hypothetical data and calculation of odds ratios (ORs) and relative risks (RRs)

	Sample of injured (cases)	Sample of non-injured (controls)	
<i>(a1) Retrospective case-control study #1</i>			
Number that did engage in risky behavior	38	23	61
Number that did not engage in risky behavior	9	30	39
	47	53	100
	Number that got injured	Number that did not get injured	
			OR = $(38/47)/(9/47) \div (23/53) / (30/53) = 5.5$
<i>(a2) Prospective cohort study #1</i>			
Sample of those that engage in risky behavior	38	23	61
Sample of those that do not engage in risky behavior	9	30	39
	47	53	100
	Number that got injured	Number that did not get injured	
			RR = $(38/61) \div (9/39) = 2.7$
<i>(b1) Retrospective case-control study #2</i>			
Number that did engage in risky behavior	6	55	61
Number that did not engage in risky behavior	3	36	39
	9	91	100
	Number that got injured	Number that did not get injured	
			OR = $(6/9)/(3/9) \div (55/91) / (36/91) = 1.31$
<i>(b2) Prospective cohort study #2</i>			
Sample of those that engage in risky behavior	6	55	61
Sample of those that do not engage in risky behavior	3	36	39
	9	91	100
	Number that got injured	Number that did not get injured	
			RR = $(6/61) \div (3/39) = 1.28$

Table 20.4 Association between airbag deployment and sustaining a cervical spine injury in motor vehicle collisions (abstracted from Stein et al. 2011)

	Cervical spine injury sustained	No cervical spine injury sustained	Total
Airbag not deployed	163	986	1,149
Airbag deployed	244	2,130	2,374
Total	407	3,116	3,523
OR = 1.44		p value exact (two-sided Fisher's test) = 0.0009	
RR = 1.38		p value approximate (chi-squared test) = 0.0007	

The significance of the comparison is done by Fisher's exact test or the chi-squared test that allows for the approximate calculation of the p value associated with the comparison. Stein et al. (2011) looked at factors associated with cervical spine injuries sustained in motor vehicle collisions, and the example in Table 20.4 illustrates the two concepts mentioned above, namely (1) that when the event is "rare" the OR and the RR are quite similar (1.44 and 1.38, respectively, in this case), and (2) that for large numbers, the exact p value from Fisher's exact test is well approximated by the chi-squared test (0.0009 and 0.0007, respectively, in this case).

Comparing Count Variables over Time

Temporality to establish effectiveness usually involves observing the counts over time. Comparisons of counts over time in a given setting can be done graphically or modeled with time series. For example, Peck et al. (2007) studied the trends in deaths due to motor vehicle collisions in Uzbekistan from 1981 to 1998 – see Fig. 20.2.

However, if multiple settings are available, one can use generalized estimating equations (GEE) or mixed effects linear regression models, to study the relative effects of multiple factors. Kim et al. (2007) use multilevel or mixed effects regression models to study the effects of intersection-level and crash-level factors on the outcomes of crashes. The hierarchical structure of the data, with crashes that occur at the same intersection "nested" within an intersections and thus sharing the effects of the same intersection-level factors, induces a correlation among the outcomes of different crashes that occur at the same intersection. Another fine example of such modeling techniques is a 3-level multilevel model found by Jones and Jørgensen (2003) for predicting outcome of crashes at the individual person level, but with hierarchical data of the individuals nested within vehicles and these vehicles in turn nested within locations.

Practical and Ethical Limits on Possible Study Designs

There is a well-known hierarchy of methodologies to establish the effectiveness of an intervention, when considering point #7 above – specificity, i.e., a cause and effect relationship. Figure 20.3 is a typical display of the hierarchy in terms of a pyramid, while Table 20.5 provides more detail of the different types of study designs.

Observational study designs are subject to various forms of biases, and thus experimental designs are preferred. Within experiments, some designs are considered more robust than others. For example, individual randomized experiments (trials) are considered as stronger than group or

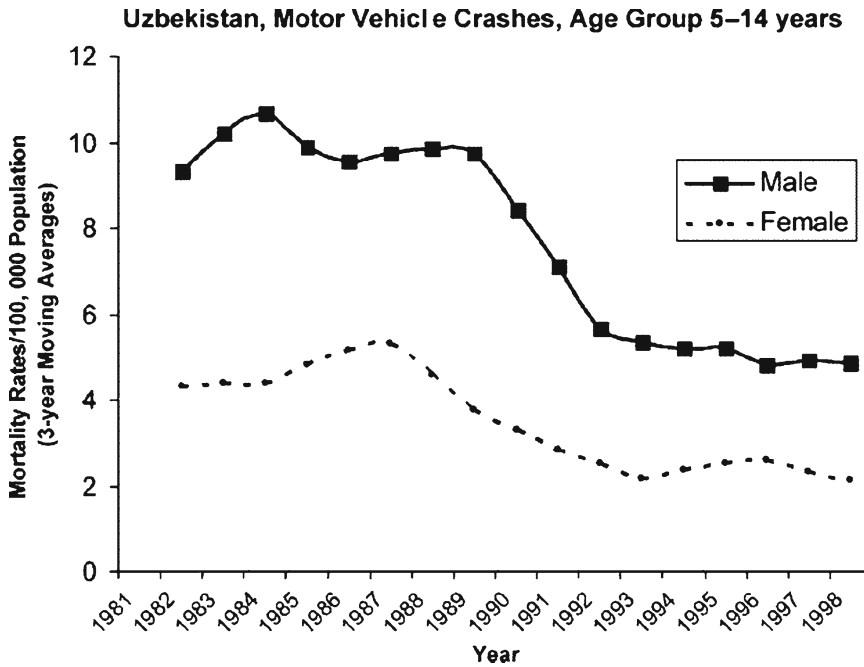
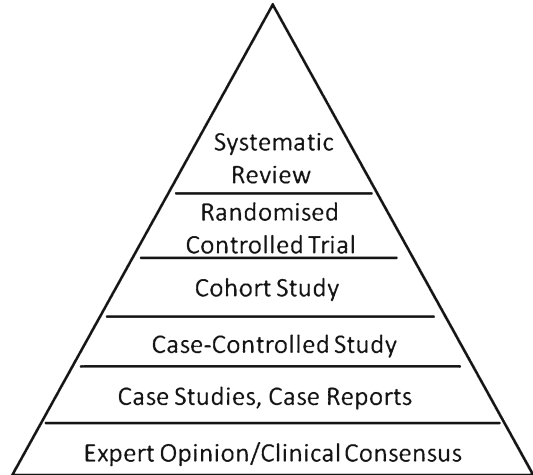


Fig. 20.2 Time trend of mortality due to motor vehicle collisions in Uzbekistan. *Source:* Peck et al. 2007

Fig. 20.3 Pyramid of evidence from different study designs



community randomized experiments. However, in the injury research field, often it is unethical or impractical to do individual randomized experiments. Many interventions are policies or infrastructure changes that affect entire countries or communities, while others such as the use of seatbelts or airbags cannot be ethically denied to individuals. Thus, often specific interventions are carried out in groups of individuals, while the outcomes of the interventions are examined in individuals. These types of experimental study designs are called group- or cluster-randomized experimental designs, and similar to the hierarchical or nested multilevel data form studying individuals from vehicles and

Table 20.5 Hierarchy of evidence from different study designs

-
- Meta-analyses and meta-regressions of RCTs
 - Systematic reviews of RCTs
 - Experimental
 - Randomized-controlled trials (RCTs)
 - Cluster-randomized trials
 - Community intervention trials
 - Natural experiments, field trials
 - Observational
 - Prospective
 - Of individuals: Cohort studies
 - Of groups: surveillance registries
 - Retrospective: case-control studies
 - Cross-sectional: prevalence studies, surveys
 - Case series
-

from locations, one must account for the intra-cluster correlations in these study designs. They are often called “quasi-experimental designs since individuals are not assigned to the interventions but rather groups are assigned.” The evidence from these studies is not considered as strong as that from individualized randomized studies, but often they are the only ones possible in the injury field, and a well-done group randomized design should be considered as providing strong evidence of the effectiveness of an intervention.

How to Produce “Practice-Based Evidence” of Intervention Effects: Mixed Effects Meta-Regression of Observational and Experimental Studies

What we as a field have not done as well as other fields, is to draw strength from numbers. We need to develop the collective evidence that comes from combining the results from multiple studies. This can be done with systematic reviews, a protocol-driven comprehensive review and synthesis of data focusing on a topic or on related key questions. These begin by formulating specific key questions, developing a protocol, refining the questions of interest, then conducting a literature search for evidence, selecting studies that meet the inclusion criteria, appraising the studies critically, and finally synthesizing and interpreting the results. Systematic reviews may or may not include a statistical synthesis called meta-analysis, depending on whether the studies are similar enough so that combining their results is meaningful.

Meta-analysis is a method of combining the results of studies quantitatively, to obtain a summary estimate of the effect of an intervention. They are often restricted to randomized controlled trials, but recently, the Cochrane Collaboration is “branching out” to include both experimental and observational studies in meta-analyses. The combining of results should take into account: the “quality” of the studies, assessed by the reciprocal of the variance of the estimated effect from the study, as well as the “heterogeneity” among the studies, assessed by the variance between studies. The summary estimate of the effect is a weighted average, where the weight of a study is the reciprocal of its variance. To calculate the variance of a study, one can use either a “fixed” effects model or a “mixed” or “random” effects model.

The fixed effects model utilizes no information from other studies, and the variance of each study is given by $\text{var}(Y_i) = \text{var}(e_i) = V_{Y_i} \dots$, where e_i denotes an error term. On the other hand, the random effects model considers both the variance among and within studies, and decomposes the total variance of each study into components, one between studies ζ_i and the error within studies:

$$Y_i = \mu + \zeta_i + e_i,$$

where the variance between studies is

$$\text{var}(\zeta) = \tau^2,$$

and thus the study's variance is

$$\text{var}(Y_i) = \tau^2 + V_{Y_i}^*.$$

The consideration of the variance among studies τ^2 is a way to handle the “heterogeneity” among the studies, and allows one to get a more precise estimate of the overall intervention effect. If it does not handle entirely the heterogeneity and there is still residual heterogeneity, one can expand the mixed effects model to include study-level covariates that may explain some of the residual variability among studies, and this methodology is what is called meta-regression. Now a study's effect Y_i

$$Y_i = \mu + \beta_1 X_{1i} + \beta_2 X_{2i} + \zeta_i + e_i,$$

is decomposed into an overall mean effect μ , its deviation ζ_i from the mean, the error term e_i or variation within the study, plus the effects β of covariates which are study-level potential explanatory variables of the heterogeneity among studies. This heterogeneity is bigger among observational studies than experimental studies, since in experimental studies, one can control such heterogeneity by design, but to a lesser degree in observational studies. Since in the injury field we must rely on more quasi-experimental study designs and observational studies, these techniques are more relevant.

We thus see that methodology does exist for developing stronger collective evidence, evaluating the effectiveness of community-based interventions, using different types of study designs and interventions, and thus of developing the “practice-based evidence” the injury field needs.

Concluding Remarks

This chapter has highlighted a few key statistical issues that may impact injury prevention and control research. The focus has not been on presenting specifics of various statistical methodologies, but on providing some understanding of concepts and principles. We used the “motor vehicle-related accidents” problem to illustrate how the application of the principles and concepts of statistics can help one to have an “educated” understanding of the problem. Researchers that have an understanding of how statistical methods quantify uncertainty, quantify probability, quantify risk and exposure, and quantify the strength of relationships will be better able to understand the utility of statistical methodology in injury prevention and control research.

References

- Bangdiwala, S. I. (2001). Statistical considerations for the design, conduct and analysis of the efficacy of safe community interventions. *Injury Control and Safety Promotion*, 8, 91–97.
- Bangdiwala, S. I. (2009a). Significance or importance: What is the question? *International Journal of Injury Control and Safety Promotion*, 16, 113–114.
- Bangdiwala, S. I. (2009b). Multiple multiplicities. *International Journal of Injury Control and Safety Promotion*, 16, 183–185.
- Bangdiwala, S. I. (2010). The fishy count Poisson. *International Journal of Injury Control and Safety Promotion*, 17, 135–137.
- Bangdiwala, S. I., Anzola-Perez, E., & Glizer, M. (1985). Statistical considerations for the interpretation of commonly utilized road traffic accident indicators: Implications for developing countries. *Journal of Accident Analysis and Prevention*, 17, 419–427.
- Idaho State Government (1999) itd.idaho.gov/ohs/99data/99driver.pdf. <http://itd.idaho.gov/ohs/stats.htm>. Accessed on 27 May 2011.
- Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*. New York: Wiley.
- Jones, A. P., & Jørgensen, S. H. (2003). The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis and Prevention*, 35, 59–69.
- Kim, D. G., Lee, Y., Washington, S., & Choi, K. (2007). Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. *Accident Analysis and Prevention*, 39, 125–134.
- Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Muller, K. E. (2007). *Applied regression analysis and multivariable methods* (4th ed.). Pacific Grove, CA: Duxbury Press.
- Peck, M. D., Shankar, V., & Bangdiwala, S. I. (2007). Trends in injury-related deaths before and after dissolution of the Union of Soviet Socialist Republics. *International Journal of Injury Control and Safety Promotion*, 14, 139–151.
- Rao, C. R. (1989). *Statistics and truth: Putting chance to work*. Fairland, MD: International Co-operative Publishing House.
- Stein, D. M., Kufera, J. A., Ho, S. M., Ryb, G. E., Dischinger, P. C., O'Connor, J. V., & Scalea, T. M. (2011). Occupant and crash characteristics for case occupants with cervical spine injuries sustained in motor vehicle collisions. *Journal of Trauma*, 70, 299–309.
- Weiss, H., Agimi, Y., & Steiner, C. (2010). Youth motorcycle-related hospitalizations and traumatic brain injuries in the United States in 2006. *Pediatrics*, 126, 1141–1148.
- World Health Organization. (2011). *World Health Statistics 2011*, Geneva.

Chapter 21

Video Data Analysis

Andrew E. Lincoln and Shane V. Caswell

Introduction

Coinciding with the growth and increased sophistication of injury-control research activity are two distinct sociological phenomena – the widespread use and availability of video in modern society and the rise in sport participation. Early applications of injury-control research recognized the utility of video capture and analysis of recreational activity, such as Coppens and Gentry's observation of play-ground injury-risk situations (1991). Video was recognized as a tool for the identification of common injury vignettes, or hazard scenarios, and has been primarily employed across the sports and recreational realm in efforts ranging from characterizing events leading to fatal distal limb fractures in horse racing (Parkin et al. 2006) to clarifying when to halt a boxing match (Miele and Bailes 2007).

The primary advantage afforded by video analysis is the accuracy of data collection associated with events that often occur in a fraction of a second and may not be accurately reported by the injured party, parent, coach, or other observer. In the same manner that narrative text has been used to recreate the physical and social environmental factors that may contribute to the injury event or predispose an injury to occur (Lincoln et al. 2004; McKenzie et al. 2010), video offers the ability to capture mechanisms of injury (both contact and noncontact), biomechanics associated with injury events, physiological and behavioral responses to impact, and a host of other outcomes. These investigations contribute to the deeper understanding of injury causation, such as with rupture of the anterior cruciate ligament (ACL) (Quatman et al. 2010) and development of prevention programs. Over the past decade, video analysis has gained recognition as a credible tool for investigators in the injury-control domain.

Application of Video Analysis for Sports Injury Research

Worldwide participation in recreational activities, including organized sport, is increasing. In the USA, a report from the National Federation of State High School Associations (NFHS) and the National Collegiate Athletic Association (NCAA) reported that participation in organized sport

A.E. Lincoln, ScD, MS (✉)

MedStar Sports Medicine Research Center, MedStar Health Research Institute, Union Memorial Hospital,
Room 764, Bauernschmidt, 201 E. University Parkway, Baltimore, MD 21218, USA
e-mail: andrew.e.lincoln@medstar.net

S.V. Caswell, PhD

George Mason University, Bull Run Hall 208B, 10900 University Blvd, MS 4E5, Manassas, VA 20110, USA
e-mail: scaswell@gmu.edu

reached an all-time high with over 7.5 million participants (National Federation of State High School Associations 2010). With greater participation (and corresponding increase in injury exposure) in sport and recreation activities worldwide, the demand for effective, evidence-based injury prevention programs is on the rise. Video analysis plays a key role in the development of evidence that serves as the basis for prevention programs.

To date, a considerable number of epidemiological studies using traditional methodologies and data sources (e.g., medical records, patient interviews) have elucidated the incidence rates and injury patterns in a wide variety of sports and recreational activities (Hinton et al. 2005; Hootman et al. 2007; Lincoln et al. 2011). However, much is unknown regarding the risk factors and mechanisms of injuries specific to various sports and recreational activities. Traditional methodologies that have determined the mechanism of injury via patient self-report (e.g., medical records and patient interviews) or eyewitness accounts (e.g., athletic trainer, coach, physician, teammate, or parent) suffer significant limitations due to recall bias or observation and recording errors when attempting to describe specific hazard scenarios and injury mechanisms in sport. The use of video analysis can overcome these limitations by providing researchers an opportunity to capture, analyze, and describe not only injury mechanisms but also the general and specific characteristics of a given recreational activity or sporting event (Quarrie and Hopkins 2008).

Video has commonly been used in sports to perform match analysis in which coaches evaluate patterns of play and team and player performance (Andersen et al. 2003). This approach has been used in research activities to explore precipitating events and contributing player or environmental factors that lead to an injury, such as the effect of ice rink size on hockey collision rates (Wennberg 2005) and the effect of infraction type on head impact severity (Mihalik et al. 2010). Andersen et al. (2003) further described a newer technique of incident analysis as a refinement of match analysis that is video-based and combines sport-specific and medical information to describe how injuries and high-risk situations of injury occur. This technique has been used to describe the biomechanical characteristics that occur at the time of injury (Andersen et al. 2004a, b; Krosshaug et al. 2007a, b). Other applications of incident analysis describe the immediate behavioral and physiological responses to injury, such as McCrory and Berkovic's (2000) portrayal of the physical manifestations of acute sports-related head injury and Koh and Watkinson's (2002) description of the responses to blows to the head received during taekwondo matches. Selected examples of literature using video analysis can be seen in Table 21.1.

Methodological Considerations

The application of video analysis methodologies to injury research can provide a powerful tool to better understand hazard scenarios and injury mechanisms. Prior to beginning any investigation, whether using video as the basis of data collection or not, some important methodological issues must be considered. For example, examining the kinematics (e.g., player velocity) of a specific event vs. the characteristics of game play (e.g., frequency of fouls) requires different technical approaches. The video analysis process is presented in the following five phases: (1) conceptualization, (2) resources, (3) instrument development, (4) data collection, and (5) data analysis. For a more detailed overview of the process, please refer to Table 21.2. Regardless of the specific methodology chosen, the process of conducting an investigation using video analysis requires that numerous theoretical and practical judgments be made during each phase. In the following section of this chapter, we identify and discuss the primary considerations and decisions that should be addressed when conducting a video analysis study.

Table 21.1 Examples of sports injury epidemiology papers using video incident analysis

Author (year)	Injury focus	Population/level of play	Video and medical sources	Outcomes identified from video	Strengths of video analysis	Limitations of video analysis
Schneiders et al. (2009)	Injury incidence and common mechanisms	Premier grade (nonprofessional club) rugby union	Injury data surveyed from team physiotherapists	Exposures to injury (player challenges)		
Krosshaug (2007b)	Biomechanical analysis of anterior cruciate ligament injury mechanisms	Basketball, downhill skiing, and European team handball	Video of three ACL injury situations recorded with uncalibrated cameras	Kinematics of injury incident	Video yields much better estimates of kinematics compared with the simple visual inspection	Difficulty in obtaining video of (rare) injury incidents
Mihalik et al. (2010)	Effect of infraction type (legal/illegal) on head impact severity	Youth ice hockey	Game video and helmets instrumented with accelerometers	Legal/illegal hits synced to helmet instrumentation (impact forces)	Video enables identification of potentially injurious collisions that may occur beyond referees' field of vision and elucidates the effects of player infractions on measures of head impact severity	
Fuller et al. (2005)	Incidence and causes of head and neck injuries	International football (soccer)	Game video and a standardized injury report form completed by team physicians	Legality of challenges resulting in injury and players' actions most likely to cause a head/neck injury	In the majority of cases, player challenges were deemed to be fair and within the laws of the game	Of the 248 incidents, 163 were identified on videotape and also met the criteria for inclusion in the analysis (66% – potential selection bias)
Quarrie and Hopkins (2008)	Injury risk associated with various characteristics of tackle techniques	Professional rugby union	Video data for tackles resulting in injury was cross-linked to medical data	Inciting events leading to injury and the burden of injuries	Information from video records typically yields more accurate information about the circumstances associated with the injury than is available from player recall of the event or direct observation	

Table 21.2 Video analysis process

1	2	3	4	5
Conceptualization	Resources	Instrument development	Data collection	Data analysis
Devise research question	Data source	Instrument available	Hardware selection	Raw video
Review literature	Video available	Instrument adaptable	Video camera	Compiling video
Target population	Retrieval costs	Instrument development	Specifications	Software
Target event	Video quality and format	Expert focus group	Quantity	Identifying events
Target attributes	Video unavailable	Identify target attributes	Cost	Clipping events
	Access to population	Item development	Data storage and sharing	Storage
	Equipment	Instrument medium	Type	Sharing video
	Personnel	Paper	Cost	Coding video
	Time	Electronic	Filming venue	Work flow
		Cost	Location	Trained raters
		Access	Number	Analysis
		Group review	Weather conditions	Software
		Train raters	Access to power	Reliability
		Pilot study #1	Videography	Statistical testing
		Revise instrument	Camera placement(s)	
		Pilot study #2	Lighting	
			Videographer training	
			Piloting filming	
			Identifying target event	
			Time stamp	
			Linkage of medical records	
			Confirmation of diagnosis	
			Injury severity	
			Case disposition	

Conceptualization

As with other research initiatives, clearly defining the purpose of the study, specific research questions, and sources of available data guide the methodology. During the conceptualization phase, investigators should carefully review the existing literature to inform their research question. Once formulated, the research question must be refined to reflect the target population (e.g., girls’ high school soccer athletes), the target event (e.g., all corner kicks), and the specific attributes associated with that event (e.g., characterize head-to-head collisions). Determination of these three factors will aid in decision-making throughout the video analysis process.

Resources

Once a research question is formulated and the conceptual framework conceived, the next important issue concerns available resources.

Data Sources and Collection

The simplest and most cost-effective approach to obtaining video is to make use of what has already been collected. Televised sporting events at the professional and collegiate level typically feature high-quality video and multiple camera angles. Many studies have employed this approach and produced significant advances in the understanding of the mechanisms of specific injuries, such as ACL rupture (Koga et al. 2010; Krosshaug et al. 2007a) and head injuries (Andersen et al. 2004b; Fuller et al. 2005; Viano et al. 2007). While such findings are critical to the development of prevention efforts at elite levels of play, they may not be applicable to the vast majority of sports participants at lower levels of play, such as those in high schools and club/recreational leagues where participants are typically younger and less skilled. The challenge to perform video analysis among this larger group of athletes (for which we have less understanding of injury incidence and mechanisms) becomes collection of video that is of high quality and (if possible) contains multiple camera angles.

Instrument Development

Researchers may use video analysis to develop detailed observational knowledge relating to a specific population or phenomenon that is not available through other sources of data. The video data represent observational variables comprising a snapshot of the human experience while participating in a specific activity. These data may then be systematically analyzed and used to better understand specific characteristics of hazard scenarios or injury mechanisms. Development of a coding instrument that produces valid and reliable data is perhaps the most important component of any video analysis project. This section describes the general steps and considerations when developing a coding instrument for video analysis. An overview of the instrument development process can be seen in Table 21.2.

Identifying the Research Question

Similar to other coding methodologies, the first step in instrument development is to determine what the tool will measure. This step is driven by the research question and is accomplished by identifying and defining the specific attributes or components of the target event to be examined. For example, McIntosh et al. (2010) performed a study to measure associations between tackle characteristics, level of play, and injury. To accomplish this aim, the researchers developed a qualitative protocol to measure specific attributes of unsafe tackling (McIntosh et al. 2010). The framework of this example can be helpful by enabling the application of Haddon's approach to explore injury etiology, context, contributing factors, and the sequence of events (Bahr and Krosshaug 2005; Haddon 1968).

Review the Literature

Prior to developing any coding instrument, it is vital to extensively review the existing literature. This activity refines the attributes of the target event and may identify a preexisting instrument. If an existing instrument is found, the researcher must determine whether it can be used in the current form or adapted to meet the needs of the present study. Even instruments that were used for other objectives can serve as a model and be altered to meet the specific needs of the research objective.

Expert Focus Group

In this phase, the investigators should obtain feedback from experts in the field. For example, in a recent study of head injuries in women's lacrosse, Caswell et al. (2010) sought the expertise of coaches, officials, and players to refine the research question and develop the coding instrument. Including a panel of subject matter experts helps to establish the utility of the research questions and instrument content validity.

Identifying Specific Attributes of the Target Event

Identifying and analyzing specific injury events amid a complex sporting event are difficult to accomplish through direct observation or interviews. Video analysis of target events can aid investigators in examining the specific attributes (variables) of the target event. Incorporating a combination of published research and expert feedback is effective for determining the specific attributes to include in the coding form (Fuller et al. 2005; Andersen et al. 2003; McIntosh et al. 2010). For example, Fuller et al. (2005) used video analysis to determine the incidence and causal factors associated with head and neck injuries in international football (soccer). They examined specific attributes (e.g., field location, ball possession, tackle characteristics, and body part involved) of game instances deemed to have the greatest impact on head and neck injuries in international football. As an example, Fig. 21.1 provides a brief excerpt from the Lacrosse Incident Analysis Instrument developed for a study examining head injuries in boys' and girls' high school lacrosse (Caswell et al. 2010).

Item Development

Once a list of target attributes is established, individual items are constructed for each attribute. In addition, any items describing demographic or descriptive elements (e.g., location, level of play, weather conditions) should also be developed. An important consideration in the construction of a coding instrument is item format and scaling. Format and scaling should be reflective of the research question and intended purpose of the coding instrument. Various types of item scales can be devised. Examples include selection criteria that force a choice between two dichotomous categories (e.g., yes or no, present or absent) or multiple descriptors (e.g., first, second, third, or fourth quarter). Other item scales can be employed that require a judgment about some element of the target event (e.g., to quantify the height and direction of tackle on the ball carrier, speed of tackler, and speed of ball carrier) (Quarrie and Hopkins 2008). At this stage of development, it is recommended that the initial instrument has more items than is anticipated for the final instrument length. The creation of more items than necessary will permit refinement of the instrument.

Think Aloud and Pilot Testing

When the initial instrument construction is complete, the next step in the process is to conduct a "think aloud." The "think aloud" enables researchers, members of the expert focus group, and raters to use the coding instrument in a group setting with pilot video. Ideally, the "think aloud" session encourages open dialogue between all parties regarding the functioning of each item and the instrument as a whole. Through this process, any ambiguous or otherwise problematic items can be clarified or removed to improve the functioning of the instrument.

Sample General Target Event Attributes

1. Environmental Information
 - a. Describe the weather conditions at the time of injury
 - i. Clear
 - ii. Overcast
 - iii. Precipitating (Rain, Sleet, Snow)
2. Field Information
 - a. Describe the field conditions at the time of injury
 - i. Grass
 - ii. Turf
 - b. Describe the location on the playing field at the time of injury
 - i. Attacking area
 - ii. Midfield
 - iii. Goal

Sample Specific Target Event Attributes

3. Specific information about the head injury
 - a. Describe the initial location of impact to the injured athletes head?
 - i. Face or Facemask
 - ii. Right Parietal
 - iii. Left Parietal
 - iv. Crown
 - v. Occiput
 - vi. Body other than head
4. Impact Source
 - a. What impacted the player ?
 - i. Opponent
 - ii. Playing field
 - iii. Object
 1. Stick
 1. Shaft
 2. Head
 2. Ball
 3. Goal Post
 - b. What part of opponent first impacted the injured player?
 - i. Head / Helmet
 - ii. Shoulder
 - iii. Arm
 - iv. Elbow
 - v. Hand
 - vi. Leg
 - vii. Foot

Fig. 21.1 Sample items from the Lacrosse Incident Analysis coding instrument (reference: Caswell et al. (2010))

Data Collection

The systematic collection of video is critical to the quality of the data generated by the coding instrument and ultimately the outcome of the study. The continual and rapid evolution of technology makes equipment selection challenging for researchers who may not be up to date with the latest available video technology. The reader is directed to the many buyers' guides for selection of video equipment that are available on the Internet and magazines. The purpose of this section is to provide a brief primer regarding the major considerations when planning any video analysis study, including video camera selection, the shooting of video, data storage and sharing methods, and the linking of any companion data (e.g., medical records) to the video data.

Video Equipment Selection

Selection of the "best" video equipment can be a daunting task. Some important specifications include image resolution, light sensitivity, frame rate, lenses, interface/connectivity, output format, memory storage, microphone placement, and battery life. We contend that the "best" video equipment is not the most expensive, but rather the equipment that is "most fitting" for the specific needs of the study.

Shooting of Video

Regardless of the video equipment selected, the shooting of video is crucial to assuring the quality of the video data. Prior to shooting any video is the development of a standard video capture process. For example, standardizing the filming location, field of view, and stopping and starting points among other site-specific considerations should be taken into account as they can impact the quality of video. Moreover, we recommend the assistance of experts in videography when selecting equipment and sport-specific training of videographers. We further recommend shooting pilot video in a realistic competitive setting. These efforts will serve to improve video quality and facilitate data coding.

Linkage of Video to Companion Data Records

For many sporting activities, team physicians and athletic trainers document their evaluation and treatment of player injuries. The linkage of such records with video footage enhances the understanding and details of an incident. For example, an injury record provides confirmation of the actual diagnosis, which often cannot be determined from video alone. The record may also provide information on the severity of injury and return to participation.

Data Analysis

Researchers should anticipate that the time required to properly code and analyze the data will approximate the time required to collect the video. This time frame can be significantly shortened if video is simultaneously recorded and uploaded directly to a computer.

Coding Process and Quality Assurance

A standardized method of coding video will improve the quality of the data. As mentioned in the “Instrument Development” section of this chapter, the training of raters and proper pilot testing of the coding instrument are important components of this process. Many methodologies and computer software tools exist to assist with the coding of data. We recommend that researchers develop a standardized electronic form for the inputting of data generated during the coding of video. This method will normalize the input process and limit typographical and other human errors during data entry. If there are multiple coders, each person should undergo the same training, and the interrater reliability should be calculated during both the pilot testing and the actual study.

Statistical Analyses

Many of the match analysis-oriented papers use descriptive statistics such as frequency, proportion, rates, and relative rates to identify which player positions are at greatest risk of injury, what the injury rates are, what the player activity was at the time of injury, where the injury occurred on the field, and what the burden of a specific injury is. (Andersen et al. 2004a, b; Krosshaug et al. 2007b; Quarrie and Hopkins 2008). Incident analyses often examine differences in player kinematics at the time of injury, such as differences in knee and hip flexion upon jump landing between females and males (Krosshaug et al. 2007a) or the use of upper extremities during heading in football (soccer) and the resulting head injury events (Fuller et al. 2005). The interrater reliability for identifying consistency of coding was typically presented using the Kappa statistic to assess level of agreement among coders. Disagreement among coders was often resolved by a panel of experts. An overview of some of the important methodological considerations pertaining to each phase of the video analysis process can be seen in Table 21.3.

Strengths and Limitations of Video Analysis

According to Quarrie and Hopkins (2008), a strength of video data is that they typically yield more accurate information about the circumstances associated with the injury than is available from player recall of the event or direct observation of data collectors, for example, physicians on the sidelines of matches. When combined with medical information about the injuries, systematic video analysis provides a powerful approach to identifying risk factors for injuries in sport. Nonetheless, it has been suggested that without simultaneous access to medical information from team medical staff, video data alone may result in a biased description of how injuries occur. For example, video footage alone can be expected to reveal injuries resulting from the more spectacular incidents (e.g., body contact) but may overlook a significant number of more subtle but significant injuries (e.g., muscle strains to the thigh or ankle and knee sprains) (Andersen et al. 2004b; Quarrie and Hopkins 2008). An example of the undercounting of events may be found in Fuller et al.’s (2005) investigation of head and neck injuries during international football games in which only 163 incidents were identified on videotape and met the criteria for inclusion in the analysis among the 248 incidents captured on standardized report forms by team physicians. Those incidents identified on videotape (66%) typically represent the more severe injuries resulting from flagrant collisions. Difficulty in identifying the actual injury event on videotape also contributes to an undercount of incidents, which is particularly true of injuries such as concussions that are often the result of seemingly mild impacts or may not be reported by players to the medical staff until the day following the contest.

Table 21.3 Video analysis methodological considerations**Key considerations**

Conceptualization

Research question

- What is the research question?
- Can a population of interest be identified?
- Is there an available source of video that could be used to answer the research question?

Resources

Video

- Does this video already exist? If so, is the video accessible?
- If no video exists, is the target population available for filming?

Technology

- Is there access to necessary video recording resources, computer processing, data storage, and analytical software?

Personnel

- Are qualified personnel who are familiar with the sport/activity available to film, code, and analyze video?

Instrumentation

Coding instrument

- Is an existing coding instrument available or adaptable?
- Is there time to develop an instrument?
- What are the specific attributes of the target event you wish to examine?
- What other characteristics should be measured (e.g., style of play, intensity, aggression, speed of game)?
- How can these attributes be measured/quantified?

Data collection

Target venue characteristics

- Where is the activity located (indoors or outdoors)?
- What is the size of the playing surface?
- Are environmental factors an issue (e.g., rain, wind, low or artificial lighting)?
- What are the best available locations to film?
- Will there be access to a reliable power supply?
- What are the potential obstructions (e.g., structural supports or fans)?

Target activity characteristics

- What is the nature of the target activity?
- Is the game play spread out across a large field with multiple participants (e.g., soccer) or focused within a confined area with fewer participants (e.g., wrestling)?
- Does the activity involve an object that must be followed (e.g., a ball)?
- What is the object's size and speed of movement?
- Is the general speed of the activity fast-paced (e.g., ice hockey)?

Target event characteristics

- What is the estimated frequency of the target event (e.g., two injuries per game)?
- What are the primary injury exposures (e.g., player challenge, tackle, body check)?
- What is the anticipated event location on the playing surface (e.g., near the goal)?
- What is the timing of the target event (e.g., game segment)?

Specific attributes of target event

- How will events be identified in the case of a specific injury (e.g., concussion or knee ligament strain), more general injury (e.g., head injury or lower extremity), or game characteristic (e.g., penalty or surrogate measure for dangerous play)?

Videography

- Can the question be adequately answered with a wide angle approach to shooting video (e.g., match analysis) or does it require a close-up or zoomed view of a specific event (e.g., incident analysis)?
- What are the best locations to shoot video? Can these locations be standardized between venues?
- Can participants obstruct the view of the target event? If so, is more than one camera needed per venue?
- How will the location of the target event be marked for later analysis?
- How will video be stored and shared with others?

(continued)

Table 21.3 (continued)

<i>Supporting records</i>	
	<ul style="list-style-type: none"> • What other data sources (e.g., medical records) are available to corroborate video data? • How will these records be obtained and integrated with video data?
Data analysis	
<i>Capturing, coding, and analyzing video</i>	
	<ul style="list-style-type: none"> • How much time will be required to code and analyze the video data? • How many raters will be required to code the data in an efficient manner? • How will interrater and intrarater reliability be assessed? • What are the appropriate statistical analyses?

Specific injury events, such as ACL injury, are rare events relative to the course of a season. Krosshaug et al. (2007a) succeeded in compiling 39 cases of videotaped ACL injury by requesting these cases from high school, college, and professional basketball coaches across the USA. This approach benefited from the large amount of video that is commonly collected at sporting events across all levels of play. However, this approach was also limited by poor video quality among 17 of the 39 cases, which were not useable for many of the coded variables.

Summary

As video cameras become increasingly accessible to a wider population, the technology to enhance video quality similarly improves. The application of video capture and analysis among recreational and sport activities lends itself well to other areas of research, especially those in which video surveillance may already exist for other purposes (e.g., traffic monitoring and crime prevention). Video analysis is a tool that allows data collection for validation of proposed theories (Laureshyn et al. 2010) and development of interventions. Combined with increased access to share and store video files, the potential for video analysis will largely depend on the creativity and resourcefulness of investigators to apply the techniques described in this chapter for advancements in injury-control research.

References

- Andersen, T. E., Larsen, O., Tenga, A., Engebretsen, L., & Bahr, R. (2003). Football incident analysis: a new video based method to describe injury mechanisms in professional football. *British Journal of Sports Medicine*, 37(3), 226–232.
- Andersen, T. E., Árnason, Á., Engebretsen, L., & Bahr, R. (2004a). Mechanisms of head injuries in elite football. *British Journal of Sports Medicine*, 38(6), 690–696.
- Andersen, T. E., Tenga, A., Engebretsen, L., & Bahr, R. (2004b). Video analysis of injuries and incidents in Norwegian professional football. *British Journal of Sports Medicine*, 38(5), 626–631.
- Bahr, R., & Krosshaug, T. (2005). Understanding injury mechanisms: a key component of preventing injuries in sport. *British Journal of Sports Medicine*, 39, 324–329.
- Caswell, S. V., Lincoln, A. E., Hepburn, L. M., Dunn, R. E., & Almquist, J. L. (2010). *Identifying mechanisms of sports-related injuries through video analysis*. Paper presented at the American College of Sports Medicine, Baltimore, MD.
- Fuller, C. W., Junge, A., & Dvorak, J. (2005). A six year prospective study of the incidence and causes of head and neck injuries in international football. *British Journal of Sports Medicine*, 39, i3–i9.
- Haddon, W. (1968). The changing approach to epidemiology, prevention, and amelioration of trauma: the transition to approaches etiologically rather than descriptively based. *American Journal of Public Health*, 58(8), 37–41.

- Hinton, R. Y., Lincoln, A. E., Almquist, J. L., Douoguih, W. A., & Sharma, K. M. (2005). Epidemiology of lacrosse injuries in high school-aged girls and boys: a 3-year prospective study. *American Journal of Sports Medicine*, 33(9), 1305–1314.
- Hootman, J. M., Dick, R., & Agel, J. (2007). Epidemiology of collegiate injuries for 15 sports: summary and recommendations for injury prevention initiatives. *Journal of Athletic Training*, 42(2), 311.
- Koga, H., Nakamae, A., Shima, Y., Iwasa, J., Myklebust, G., Engebretsen, L., et al. (2010). Mechanisms for noncontact anterior cruciate ligament injuries: knee joint kinematics in 10 injury situations from female team handball and basketball. *American Journal of Sports Medicine*, 38(11), 2218–2225.
- Koh, J. O., & Watkinson, E. J. (2002). Video analysis of blows to the head and face at the 1999 World Taekwondo Championships. *Journal of Sports Medicine & Physical Fitness*, 42(3), 348–353.
- Krosshaug, T., Nakamae, A., Boden, B. P., Engebretsen, L., Smith, G., Slauterbeck, J. R., et al. (2007a). Mechanisms of anterior cruciate ligament injury in basketball: video analysis of 39 cases. *American Journal of Sports Medicine*, 35(3), 359–367.
- Krosshaug, T., Slauterbeck, J. R., Engebretsen, L., & Bahr, R. (2007b). Biomechanical analysis of anterior cruciate ligament injury mechanisms: three-dimensional motion reconstruction from video sequences. *Scandinavian Journal of Medicine & Science in Sports*, 17(5), 508–519.
- Laureshyn, A., Svensson, A., & Hyden, C. (2010). Evaluation of traffic safety, based on micro-level behavioural data: theoretical framework and first implementation. *Accident; Analysis and Prevention*, 42(6), 1637–1646.
- Lincoln, A. E., Sorock, G. S., Courtney, T. K., Wellman, H. M., Smith, G. S., & Amoroso, P. J. (2004). Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. *Injury Prevention*, 10(4), 249–254.
- Lincoln, A. E., Caswell, S. V., Almquist, J. L., Dunn, R. E., Norris, J. B., & Hinton, R. Y. (2011). Trends in concussion incidence in high school sports: a prospective 11-year study. *American Journal of Sports Medicine*, 39(5), 958–963.
- McCrary, P. R., & Berkovic, S. F. (2000). Video analysis of acute motor and convulsive manifestations in sport-related concussion. *Neurology*, 54(7), 1488–1491.
- McIntosh, A. S., Savage, T. N., McCrary, P., & Frechede, B. O. (2010). Tackle characteristics and injury in a cross section of rugby union football. *Medicine & Science in Sports & Exercise*, 42(5), 977–984.
- McKenzie, K., Scott, D. A., Campbell, M. A., & McClure, R. J. (2010). The use of narrative text for injury surveillance research: a systematic review. *Accident; Analysis and Prevention*, 42(2), 354–363.
- Miele, V. J., & Bailes, J. E. (2007). Objectifying when to halt a boxing match: a video analysis of fatalities. *Neurosurgery*, 60(2), 307–315. discussion 315–306.
- Mihalik, J. P., Greenwald, R. M., Blackburn, J. T., Cantu, R. C., Marshall, S. W., & Guskiewicz, K. M. (2010). Effect of infraction type on head impact severity in youth ice hockey. *Medicine & Science in Sports & Exercise*, 42(8), 1431–1438.
- National Federation of State High School Associations. (2010). NFHS Participation Figures. http://www.nfhs.org/custom/participation_figures/default.aspx. Accessed on 10 Dec 2008.
- Parkin, T. D., Clegg, P. D., French, N. P., Proudman, C. J., Riggs, C. M., Singer, E. R., et al. (2006). Analysis of horse race videos to identify intra-race risk factors for fatal distal limb fracture. *Preventive Veterinary Medicine*, 74(1), 44–55.
- Quarrie, K. L., & Hopkins, W. G. (2008). Tackle injuries in professional rugby union. *American Journal of Sports Medicine*, 36(9), 1705–1716.
- Quatman, C. E., Quatman-Yates, C. C., & Hewett, T. E. (2010). A ‘plane’ explanation of anterior cruciate ligament injury mechanisms: a systematic review. *Sports Medicine*, 40(9), 729–746.
- Schneiders, A. G., Takemura, M., & Wassinger, C. A. (2009). A prospective epidemiological study of injuries to New Zealand premier club rugby union players. *Physical Therapy in Sport*, 10(3), 85–90.
- Viano, D. C., Casson, I. R., & EJ, P. (2007). Concussion in professional football: biomechanics of the struck player: Part 14. *Neurosurgery*, 61(2), 313–327.
- Wennberg, R. (2005). Effect of ice surface size on collision rates and head impacts at the World Junior Hockey Championships, 2002 to 2004. *Clinical Journal of Sports Medicine*, 15(2), 67–72.

Chapter 22

Age–Period–Cohort Modeling

Katherine M. Keyes and Guohua Li

The prevalence and incidence of fatal and non-fatal injuries have exhibited substantial trends over time (Martinez-Schnell and Zaidi 1989). By examining these trends we can gain insight into the causes of injury at the population level, for example: the effectiveness of public health prevention and intervention efforts for gun control, or the magnitude of change in social norms regarding driving practices, and can forecast the future burden of injury outcomes in the population. Quantitative evaluation of trends over time in injury is aided by a comprehensive approach to age-period-cohort analysis, an analytic tool to partition trends into components that are associated with changes over time within a given age structure of the population, time period, and birth cohort. In this chapter we will review essential concepts and definitions in age-period-cohort analysis, provide examples of historical uses of age-period-cohort analysis, illustrate the statistical problem in simultaneously estimating age, period, and cohort effects, offer an example of a multi-phase method for quantifying cohort effects using data on suicide in the United States from 1910–2004, and summarize and describe new directions and innovations in age-period-cohort analysis.

Essential Concepts and Definitions in Age–Period–Cohort Effect Estimation

Age Effects

Age effects describe the common developmental processes that are associated with particular ages or stages in the life course, regardless of the time period or birth cohort to which an individual belongs. For example, motor vehicle mortality (among those born after 1910) is highest among those aged 20–24 among men and 15–19 among women, and homicide mortality in the latter half of the twentieth century among men is highest among those in the late teens and early-1920s (Shahpar and Li 1999). Figure 22.1 shows a hypothetical mortality rate over time in which only age effects are operative. As shown, within each birth cohort, the mortality rate increases linearly across age

K.M. Keyes, PhD (✉)

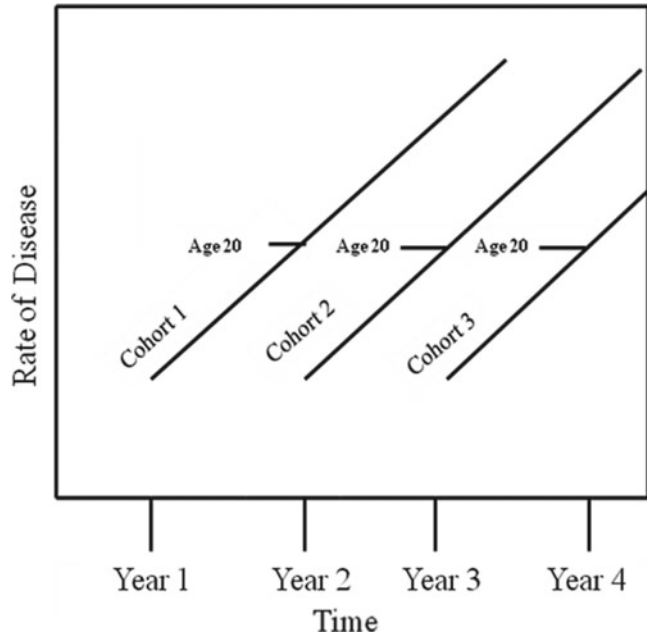
Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA
e-mail: kmk2104@columbia.edu

G. Li, MD, DrPH

Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, USA

Department of Anesthesiology, Columbia University College of Physicians and Surgeons, 622 West 168th St.,
PH5-505, New York, NY 10032, USA
e-mail: gl2240@mail.cumc.columbia.edu

Fig. 22.1 Hypothetical rate of disease in three birth cohorts over time with only age effects operative



(indicating the presence of age effects). Between each birth cohort, the mortality rate is constant across the lifespan (indicating little evidence of birth cohort effects). In the absence of period and cohort effects, age effects can explain *trends* in health outcomes only if the age distribution of the population shifts over time.

Period Effects

Period effects describe changes in the prevalence of health outcomes associated with certain calendar years across all age groups. For example, increases in death from accidental poisoning (predominately illegal and prescription drug overdose) since the year 2000 have been shown to be due in large part to period effect (Miech et al. 2011); that is, the death rate from accidental poisoning increased across all age groups simultaneously, posited to be caused by the increase in availability, use, and abuse of prescription opioid and stimulant medication across all age groups. Period effects can also arise in data due to methodological changes in outcome definitions, classifications, or method of data collection. Figure 22.2 shows a hypothetical mortality rate over time in which only period effects are operative. As shown, everyone in the population evidences an increase in the rate of disease at Time 2, regardless of age.

Cohort Effects

Cohort effects (sometimes referred to as “generation effects”) (Last 2001) are generally conceptualized as variation in the risk of a health outcome according to the year of birth, often coinciding with shifts in the population exposure to risk factors over time. For example, the risk of hip fracture in several countries is increasing in more recently born cohorts. Data from the Framingham study in the USA indicate that cohorts born after 1911 had a higher risk than those born in the late nineteenth century (Samelson et al. 2002), data from the UK suggest a combination of period and cohort effects for hip fracture influencing increased rates in the 1980s (Evans et al. 1997), and data from Finland indicate substantial increases in hip fractures in the 1990s attributable to increases in more recently

Fig. 22.2 Hypothetical rate of disease in three birth cohorts over time with only period effects operative

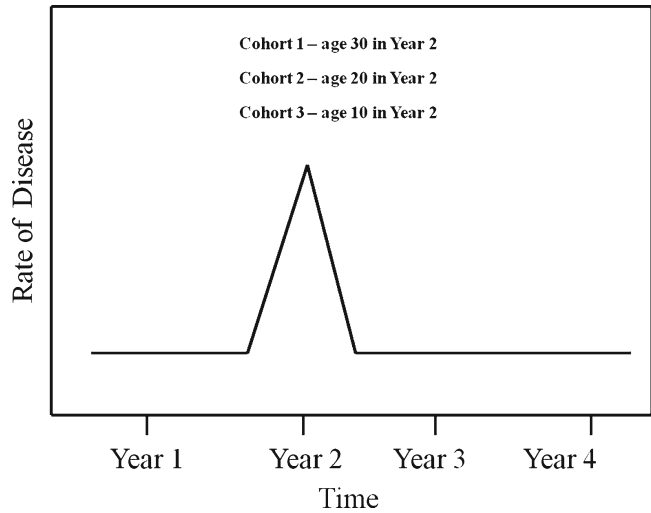
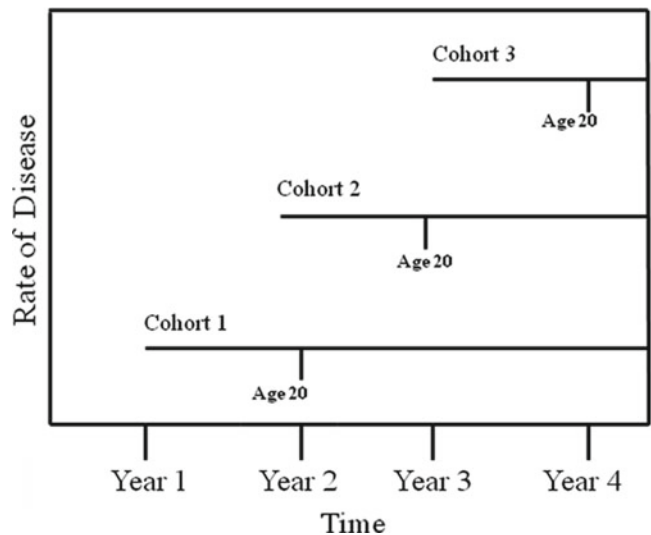


Fig. 22.3 Hypothetical rate of disease in three birth cohorts over time with only cohort effects operative



born cohorts (Kannus et al. 1999). The underlying cause of this cohort effect is posited to be due at least in part to improved life expectancy with advancements in medication and treatment of chronic disease; those in more recently born cohorts survive longer, and since hip fracture increases with age, more recently born cohorts have a higher incidence of hip fracture. Thus, population-level improvements in medication and treatment have a greater effect among one age group compared with another, resulting in the emergence of a cohort effect.

Figure 22.3 shows a hypothetical mortality rate over time in which only cohort effects are operative. As shown, Cohort 1 has an unvarying rate of disease over calendar time and age, which is a lower rate than that of Cohort 2 and that of Cohort 3.

Figures 22.1–22.3 provide simple examples of mortality rates over time when only one of three effects: age, period, or cohort is operative. In Fig. 22.4, all three cohorts exhibit a similar trajectory of mortality with respect to age (the rate is increasing linearly across time). The rate of disease in Cohort 2, however, is higher than Cohort 1 at every time point. Similarly, the rate of disease in Cohort 3 is higher than Cohorts 1 and 2 at every time point. Thus, the rate of disease exhibits age and cohort effects. In Fig. 22.5, each cohort has a different rate of disease at every time point (indicative of a cohort effect), but all three cohorts exhibit a similar increase in the disease rate at Year 2 (indicative of a period effect).

Fig. 22.4 Hypothetical rate of disease in three birth cohorts over time with age and cohort effects operative

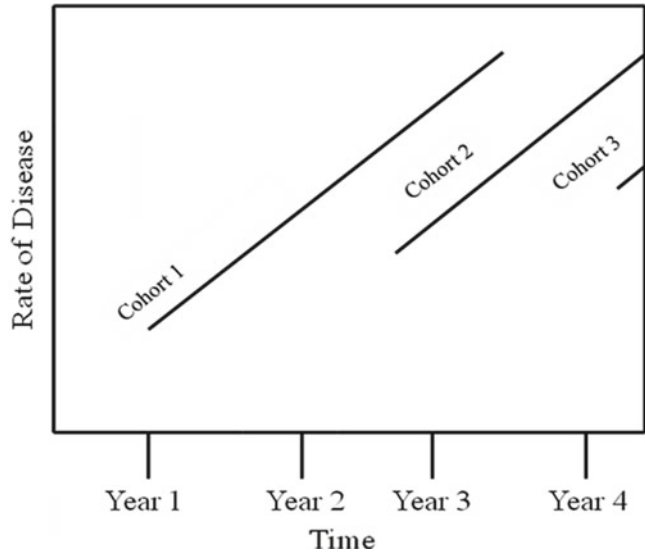
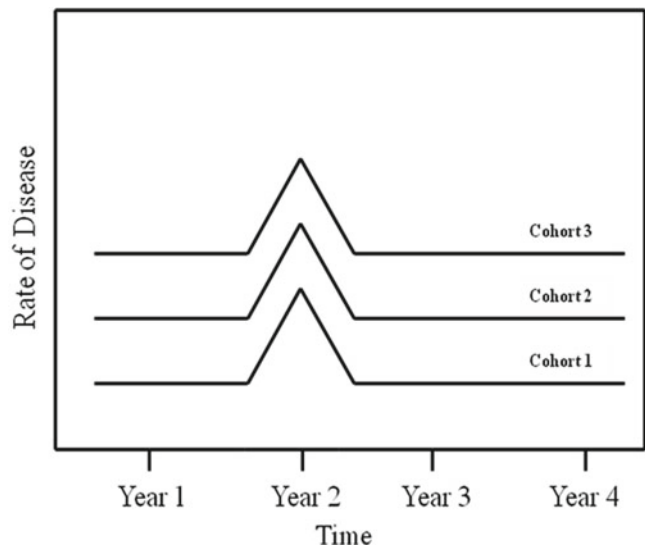


Fig. 22.5 Hypothetical rate of disease in three birth cohorts over time with period and cohort effects operative



Graphically Analyzing Age, Period, and Cohort Changes in Disease Rates over Time

By graphing disease rates by age, period, and/or cohort, patterns can be elucidated which give clues to the underlying magnitude of age, period, and cohort effects. In practice, however, rates of morbidity and mortality are rarely as simple as shown in Figs. 22.1–22.5. Instead, rates often have complex patterns over time creating difficulties in the estimation of age, period, and cohort effects. Further, methodological quandaries preclude a simple estimation of the three effects. While conceptually distinct, age, period, and cohort effects cannot be formally disentangled. Within a fixed time period, for example, an individual's age determines his/her birth cohort because $\text{Cohort} = \text{Period} - \text{Age}$. Within a given cohort moving through time, period of observation and age

will be perfectly correlated. Thus, any analysis seeking to formally estimate age, period, as well as cohort effects must contend with this intractable problem.

Despite the linear link among the three variables, age–period–cohort analyses should always begin with comprehensive graphical examination of the data across multiple time-related dimensions. Age by period, age by cohort, and period by cohort graphs should be constructed and visually examined for the presence of age, period, and cohort effects. Note that only two of the three effects can be examined at any one time in a graphical analysis; the third variable will be uncontrolled. Strong effects will visually emerge when graphs are examined. In more complicated cases, multiple regression techniques may aid in disentangling age, period, and cohort effects. We describe these techniques in more detail below (Sects. “Statistical Approaches to Analyzing Age–Period–Cohort Effects” and “The Multi-phase Method for Analyzing Cohort Effects in Age–Period Contingency Table Data: Suicide in the USA from 1910 to 2004”). Lexus diagrams have been developed to attempt a three-axis examination of age, period, and cohort effects (Carstensen 2007), although methodological issues remain (Rosenbauer and Strassburger 2007).

Historical Uses of Age–Period–Cohort Analysis

While the term “cohort analysis” often refers to longitudinal methods for individual-level follow-up data, in the early part of the twentieth century it was used to describe the graphical method for analyzing age–period contingency table data (Susser 2001). This method is applicable when individual-level data is unavailable; typically researchers have rates specific to age and time period of observation. From these rates, the health experience of cohorts can be described, but the same individuals in each cohort are not followed over time. Instead, the cohort itself becomes the unit of analysis. Several early uses of graphical approaches to age–period–cohort analysis demonstrated the power and utility of graphically assessing rates of health outcomes by age, period, and especially birth cohort (Derrick 1927; Kermack et al. 1934; Frost 1939). These analyses stand as the forebears to the development of life course epidemiology, in recognition that conditions and environmental exposures experienced early in infancy and childhood shape health throughout the life span (Ben-Shlomo and Kuh 2002; Hall et al. 2002; Lynch and Smith 2005).

Age–period–cohort analysis became popularized in epidemiology after the posthumous publication of cohort effects in tuberculosis mortality rates in Massachusetts from 1880 to 1930 by Wade Hampton Frost (1939). While overall mortality rates were on the decline over the time period in which Frost was analyzing his data, he noted that the age distribution of tuberculosis was simultaneously changing. In 1880, those in the youngest and oldest age groups evidenced the highest death rate from tuberculosis. By 1930, the death rate peaked among those in middle-age (around 40–60). The change in the age distribution was perplexing, as it was generally thought that individuals who survive past infancy during peak periods of tuberculosis exposure possess immunological protection against future illness. By analyzing his data by birth cohort rather than time period, however, Frost realized that those in middle age in 1930 were of the same birth cohort as those who were infants in 1880 and that each successively younger birth cohort had a lower risk of tuberculosis death throughout the life course, regardless of age. Frost concluded “... to have passed through a period of high mortality risk confers not protection, but added hazard later in life.” Thus, the high mortality rate among infants in 1880 was carried through the life course to those who were middle aged 1930.

Influential publications of the early- to mid-20th using graphical approaches documented cohort effects in many health outcomes including all-cause mortality (Case 1956), cancer epidemiology (Korteweg 1951; Macmahon and Terry 1958; Doll 1971), and peptic ulcer mortality (Susser 1961). The common thread through these analyses is a focus on birth cohorts as analytic units through which patterns in health outcomes can be better understood. These publications changed the way in

which health was conceived in epidemiologic practice, yet the analytic tools used across these analyses remained basic and graphical in nature.

There are limitations to the graphical method publicized by Frost and others, however. While it is a useful tool for visually assessing the cohort component in the age–period contingency table data, it is qualitative and nonparametric, and lacks an explicit definition for cohort effects. Specifically, it does not differentiate cohort effects (health outcome rates that are common to all individuals observed in a specific cohort and different across cohorts of observation) from period effects (health outcome rates that are common to all individuals observed in a specific period and different across time periods of observation). This inability to differentiate period from cohort effects stems from the structural link among age, period, and cohort. Age is defined by chronological age at the time the health outcome is experienced; period is defined by the chronological time at which the health outcome is experienced. As described above (Sect. “Essential Concepts and Definitions in Age–Period–Cohort Effect Estimation”), the cohort of any individual can be determined by the subtraction of age from period. When examining rates over time in a graphical approach, only two of the three variables can be examined simultaneously (i.e., one can graph age as a function of period or cohort, period as a function of age or cohort, and cohort as a function of age or period). But this process will necessarily leave one of the three effects uncontrolled. Thus, researchers must use outside data, theories, and assumptions to draw conclusions from the graphs of a health outcome over time.

Statistical Approaches to Analyzing Age–Period–Cohort Effects

The inability to quantitatively estimate the contribution of age, period, and cohort to health outcomes over time motivated a strong interest in sociology and biostatistics to develop models for age, period, and cohort. This was especially the case after Norman Ryder published “The Cohort as a Concept in the Study of Social Change” in 1965 (Ryder 1965). Cohorts, according to Ryder, have emergent group properties that may impact health (e.g., cohort size) and should be considered a structural category influential in health similar to race or social class. Publication of the Ryder paper sparked a surge of interest in sociology regarding the estimation of the *unique* effects of belonging to a certain cohort. Cohort membership was conceived of as an exposure with effects that exists outside of the concurrent socio-historical circumstances in which the cohorts come of age (period effects) and outside of the observed variation in disease rates across age that often occur independent of the socio-historical period (age effects). It is in the post-Ryder biostatistical and methodological developments that the term “age–period–cohort effects” becomes prominently featured in sociology, demography, and epidemiology (Mason et al. 1973; Fienberg and Mason 1979).

The sociological conceptualization of cohort effects as unique from age and period effects motivated statisticians to develop models which estimate cohort effects controlling for age and period effects. Under this conceptualization, researchers were confronted with an insurmountable challenge – the non-identification issue resulting from the collinearity among age, period, and cohort. Due to this link, statistical models cannot simultaneously estimate age, period, and cohort effects because the design matrix is less than one full rank and thus yields a non-invertible estimator. The non-identification problem makes simultaneously modeling the linear functions of age, period, and cohort effects impossible without imposing additional restrictions in the model.

Research aimed at solving the “non-identifiability” problem in the past three decades has generated a considerable body of literature and fostered the development of a variety of methodological approaches. Mason and colleagues recognized that model identifiability can be achieved if one additional constraint is placed on the parameters (Mason et al. 1973). For instance, the effect of two adjacent age groups can be constrained to be equal, or the effect of one age group and one period group. These methods ensure identifiability of the three-factor model, but the parameters are considerably sensitive to the constraint chosen, thus the method has been criticized as rendering estimates invalid and uninterpretable (Glenn 1976; Kupper et al. 1985; Holford 1991). Throughout the 1980s, a

flurry of proposals for APC modeling were made by epidemiologists and biostatisticians. A popular approach in epidemiology is to characterize the trends in two components: linear trend (i.e. “drift”) and deviations from linearity (i.e. “curvature”) (Clayton and Schifflers 1987; Holford 1992). These methods have been shown to produce reliable results compared with other methods (McNally et al. 1997; Robertson et al. 1999), being widely used in areas such as cancer research (e.g., Zheng et al. 1995, 1996; Cleries et al. 2006). The method has been criticized for limited interpretive value; however, given a definition of a cohort effect that allows existence outside of period and age effects, it is difficult, if not impossible, to assess the cohort effect based on curvature alone (Kupper et al. 1985; McNally et al. 1997). While parameter constraints and second-order functions are among the most widely used methods in age–period–cohort analysis, myriad approaches have also been developed (Nakamura 1986; Robertson and Boyle 1986; Wickramaratne et al. 1989; Tarone and Chu 1992; Berzuini and Clayton 1994; Lee and Lin 1996; O’Brien 2000; Yang et al. 2004; Yang and Land 2006).

Taken together, the decades of statistical development of age–period–cohort models have generated substantial technical innovation and creativity. However, the above methods share an underlying conceptualization of a cohort as a meaningful category which indexes barriers and resources that exist independently of the ubiquitous environmental conditions coinciding with the cohort’s collective experience through the life course. That is, that cohort effects are confounded by period and age effects. An alternative interpretation is that, rather than obscuring the effects of cohort in data collected over time, period and age effects often interact to produce the unique experiences of each cohort through the life course (Keyes et al. 2010).

The consideration of cohort effects as a partial interaction between age and period is a more rational approach to age–period–cohort analysis, and an approach that does not suffer statistically from an identification problem. This approach was first described by Greenberg et al. in their analysis of syphilis rates in the 1940s (Greenberg et al. 1950). Greenberg states “A rate may not be completely described by combining only the effect of age with the effect of time. That is, an interaction factor between age and time may be necessary in the model to describe other influences. For example, an epidemic at time T may affect one particular age group, whereas another epidemic at time T' may attack a different age group.” Thus, instead of considering a cohort effect as a proxy for past environmental influences and period effects as a proxy for contemporaneous environmental influences, this conceptualization offers an explicit, statistically identifiable definition of a cohort effect as the nonadditivity of age–period influences.

Statistical methods that capitalize on this definition have been developed in the epidemiological literature (Selvin 1996; Keyes and Li 2010); for example, the median polish approach essentially quantifies the cohort effect in the residuals of the nonadditive influences of age and period on rates over time. For example, the median polish method, developed by (Tukey 1977) and first applied to age–period–cohort analysis by (Selvin 1996), explicitly estimates the nonadditivity of age and period influences, and correlates that residual nonadditivity with a birth cohort index. Thus, this method considers cohort effects to be nonadditive influences of age and period. In the next section, we describe a multi-phase method for analyzing cohort effects using the median polish approach. We apply this approach to the analysis of suicide mortality data in the USA from 1910 to 2004.

The Multi-phase Method for Analyzing Cohort Effects in Age–Period Contingency Table Data: Suicide in the USA from 1910 to 2004

Background on Age–Period–Cohort Effects on Suicide

We illustrate the multi-phase method drawing upon data on rates of suicide completion in the USA from 1910 to 2004. In the USA, approximately 30,000 individuals die of suicide yearly; the scope of the public health burden of suicide has led to calls for increased information on determinants of and

trends in suicide risk (US Public Health Service 1999). While older age has been repeatedly identified as a risk factor for suicide, there has been a shift in the age distribution of suicide over the last century in Western countries (Stockard and O'Brien 2002a, b). Throughout the nineteenth century, suicide was rare in early- to middle life, and individuals in the oldest age groups were most likely to commit suicide. The last century, however, has seen increasing rates of suicide at younger ages. By 1995, the age distribution in suicide has become bimodal; individuals aged 20–24 and 75+ have similar rates of suicide at approximately 18 per 100,000 (higher than any other age group) (Stockard and O'Brien 2002a, b; Paulozzi et al. 2007).

The variation in the age distribution of suicide over time has prompted questions as to the possibility of cohort effects in suicide. In the USA and in other developed countries, several analyses have examined evidence of cohort effects in suicide during the twentieth century. Results in the USA have indicated relative stability in the risk of suicide among men in their 40s from 1951 to 1988 (whereas rates for women and young people fluctuate), suggesting that men in this age group may experience particular stressors that consistently affect risk for suicide (Riggs et al. 1996). There is also evidence for variation in risk that is unique to birth cohorts; sociological analyses from 1930 to 1995 have estimated that cohorts with characteristics such as large percentages of members born from nonmarital unions (i.e., born to unwed mothers) have a higher risk of suicide (Pampel 1996; Stockard and O'Brien 2002a, b). Further, data following cohorts through the early-1970s suggested a continuously increasing risk of suicide by cohort (Murphy and Wetzel 1980). Some recent age–period–cohort analyses from other developed countries suggest cohort-specific trends over time in suicide mortality. Studies of suicide among men based in Switzerland and in the general UK population found increases in the suicide rate among cohorts observed following World War II (Gunnell et al. 2003; Ajdacic-Gross et al. 2006). Several studies in Spain (Granizo et al. 1996) and the USA (Joe 2006) have noted that individuals in more recently born cohorts are more likely to complete suicide, although data from Australia suggest that increases in the suicide rate among younger cohorts is most likely due to a period effect rather than a cohort effect (Lynskey et al. 2000), and age–period–cohort analyses in Brazil suggest a decreasing rate of suicide among younger-born cohorts (Rodrigues and Werneck 2005). In Japan, cohort effects have been documented for men born after 1926 and women born after 1956; in addition, an increase in suicide across age (a period effect), especially among men, was documented in 1998 (Odagiri et al. 2011).

No investigation of age–period–cohort effects in suicide in the USA has included information from the last decade, yet recent evidence suggests that a cohort effect may be emerging among middle-aged men. Specifically, recent data from the Centers for Disease Control and Prevention (CDC) revealed unexpected increases in suicide among individuals aged 45–54 between 1999 and 2004 (Paulozzi et al. 2007). Between 1999 and 2004, individuals aged 45–54 had a 19.4% increase in suicide rates (13.1% increase among those 55–64). These data indicate that the age distribution of suicide, a critical piece of information for prevention and intervention planning, may again be shifting. Additionally, there may be an emergence of a cohort effect; specifically, the early baby boom cohorts may be evidencing a higher risk for suicide at older ages than previous cohorts.

Because of these recent changes to the epidemiology of suicide in the USA, a comprehensive examination of cohort effects including data from the last decade can provide necessary information to better characterize at-risk cohorts in the context of a changing society.

Median Polish Procedure and Methods

The median polish method was developed by Tukey (1977) and applied to age–period–cohort analysis by Selvin (1996). The purpose of median polish analysis is to remove the additive effect of age (row) and period (column) by iteratively subtracting the median value of each row and column. After

several iterations, the row and column medians approximate zero and the residual values in the cells contain the nonadditive data. In Table 22.1, we display the age-specific suicide mortality rates for men in the USA from 1910 to 2004. From this table, the median polish procedure is performed.

Two methodological issues are worth noting. First, in contingency table data, mutually exclusive cohort risks cannot be estimated because of overlapping cohorts. We have followed the convention in an age–period–cohort analysis of aggregated data to label the cohort intervals by subtracting the youngest age from the earliest year and the latest year in the interval. For instance, for the individuals aged 30–34 in 1955–1959, we subtract 30 from 1955 and 1959 to label the cohort interval, 1925–1929. The convention introduces misclassification as some of the individuals in this category will be born from 1921 to 1924 (e.g., those aged 34 in 1955–1959). This issue is not unique to the median polish method but common to all methods using aggregated data. Since the primary purpose of an age–period–cohort analysis is to estimate general trends in cohort-specific risk rather than a precise quantification of a “true” causal risk, the overlap in cohort serves as a caution against over-interpretation of generated estimates. Second, age–period–cohort analyses of contingency table data will always be limited by missing data. For example, we have only one data point available for the 1840–1844 cohort (those aged 70–74 in 1910–1914), and similarly, we only have one data point available for the 1990–1994 cohort (those aged 10–14 in 2000–2004). The impact of this missing data on cohort effect estimates depends on the association between age and the outcome of interest. Since cohort effects are an average estimate of the outcome experience of each birth cohort across age, estimates limited to the youngest or oldest age categories may be influenced by age effects.

In Table 22.2, we show the median polish residuals after removal of the log-additive age and period effects from the log-transformed suicide rates shown in Table 22.1. The median of each row and each column is simultaneously zero, indicating that the nonadditive influences of age and period have been removed, and the remaining value reflects all nonadditive influences.

Plotting the residuals against cohort category is an efficient descriptive procedure to assess the presence and size of cohort effects. If no cohort effects exist, the residuals tend to evenly distribute around zero (more specifically, in the absence of cohort effects, the expectation of the residuals should be approximately zero); a marked deviation from zero may indicate the presence of a cohort effect (nonlinearity of period and age effects). The residuals can also be subtracted from the cells of the original table, leaving cells reflecting only the additive effects of age and period for qualitative comparison with original contingency table. For a full description and examples of median polish analysis, we refer readers to the both Tukey (1977) and Selvin (1996).

We plot the residuals of the suicide data (shown in Table 22.2) by birth cohort category in Fig. 22.6. We see visual evidence of a positive cohort effect in later born cohorts, indicating that the rate is higher than what we would expect if age and period were acting additively. Note again that the data in this graph are based on the actual residual cell values from the median polish, after the removal of the additive age and period influence.

Regression of Median Polish Residuals on Cohort Category

After residuals are identified from the median polish analysis, the final step to statistically assess the relative magnitude of cohort effects is to regress residuals (e_k) on cohort category (entered as a collection of indicator variables for the $m+n-2$ cohort, $k=1, 2, \dots, m+n-2$) using linear regression, where e_k is a function of intercept μ_k , a vector of cohort effects γ_k , and a vector of error terms e_{ijk} (the errors term representing the random error unaccounted for by the cohort effect across i age, j period, and k cohort categories):

$$e_k = \mu_k + \gamma_k + e_{ijk}, \quad (22.1)$$

Table 22.1 Suicide mortality among men in the USA, 1910–2004

Age	Time period																			
	1910–1914	1915–1919	1920–1924	1925–1929	1930–1934	1935–1939	1940–1944	1945–1949	1950–1954	1955–1959	1960–1964	1965–1969	1970–1974	1975–1979	1980–1984	1985–1989	1990–1994	1995–1999	2000–2004	
10–14	0.61	0.63	0.53	0.55	0.54	0.67	0.64	0.79	0.59	0.69	0.86	0.92	1.03	1.28	1.53	2.22	2.34	2.25	1.89	
15–19	6.33	4.94	3.57	4.07	4.92	4.83	3.90	3.81	3.77	4.19	5.77	7.22	10.29	12.87	13.94	16.89	17.94	14.98	12.46	
20–24	21.79	15.30	10.98	11.79	13.33	13.02	10.72	10.09	9.55	9.97	12.06	14.94	21.62	27.49	25.54	26.21	26.39	23.57	20.70	
25–29	27.91	21.10	14.50	16.02	18.49	17.77	14.35	12.26	12.22	12.16	14.39	16.86	21.56	26.88	26.25	25.36	25.09	22.79	20.29	
30–34	32.38	26.98	19.03	18.93	22.78	22.04	17.69	15.19	13.61	14.28	17.03	18.09	20.55	23.39	23.93	24.24	24.65	22.61	20.52	
35–39	36.72	28.78	24.75	25.86	29.44	26.91	21.44	19.80	17.27	16.79	19.71	20.69	21.39	22.69	22.71	23.41	24.14	23.56	22.14	
40–44	42.96	33.50	27.44	30.97	36.66	32.62	24.22	24.08	21.98	22.49	23.26	23.53	23.06	22.81	22.31	21.97	23.28	23.88	24.16	
45–49	50.28	37.32	33.38	38.29	48.22	40.27	28.78	27.52	27.49	27.05	28.00	25.77	26.64	24.63	22.69	22.55	22.61	23.01	24.44	
50–54	59.49	41.79	39.22	46.18	58.44	48.27	35.06	33.27	32.32	33.30	33.90	30.18	27.84	25.72	24.13	23.80	22.88	21.52	23.29	
55–59	67.68	53.98	42.47	49.09	66.72	53.24	42.78	39.15	35.75	37.57	37.91	34.31	31.61	27.40	25.97	25.65	23.68	21.87	22.40	
60–64	67.93	52.21	49.89	57.34	74.63	57.42	45.46	43.80	42.13	40.38	36.67	35.58	31.31	29.50	25.48	26.26	24.77	20.84	20.37	
65–69	64.48	59.30	51.52	64.44	76.92	59.21	47.83	43.32	44.21	41.17	36.41	34.23	33.56	31.26	27.87	29.97	26.74	23.86	21.15	
70–74	64.30	57.64	55.65	65.68	82.67	58.22	52.77	48.20	48.41	44.92	40.72	36.84	36.92	36.44	35.06	39.18	34.03	29.77	26.40	
75–79							57.54	56.13	50.15	51.12	48.09	42.56	41.16	44.53	42.92	50.78	46.05	37.83	33.36	
80–84							61.17	58.68	59.30	58.64	55.67	47.69	48.12	45.77	49.91	63.00	59.18	49.22	42.18	

Table 22.2 Residual values from median polish of suicide rates shown in Table 22.1

Age	Time period																			
	1910-1914	1915-1919	1920-1924	1925-1929	1930-1934	1935-1939	1940-1944	1945-1949	1950-1954	1955-1959	1960-1964	1965-1969	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	
10-14	-0.78	-0.50	-0.51	-0.61	-0.81	-0.47	-0.22	0.00	-0.23	-0.11	0.12	0.25	0.33	0.54	0.75	0.98	1.11	1.20	1.02	
15-19	-0.13	-0.12	-0.30	-0.30	-0.28	-0.18	-0.11	-0.11	-0.07	0.00	0.33	0.62	0.94	1.16	1.27	1.31	1.45	1.41	1.22	
20-24	0.03	-0.06	-0.24	-0.31	-0.35	-0.26	-0.16	-0.21	-0.21	-0.20	0.00	0.28	0.62	0.85	0.81	0.68	0.77	0.79	0.66	
25-29	0.02	0.00	-0.22	-0.26	-0.29	-0.21	-0.13	-0.27	-0.22	-0.26	-0.08	0.14	0.35	0.56	0.58	0.39	0.46	0.50	0.38	
30-34	0.00	0.08	-0.12	-0.26	-0.25	-0.16	-0.09	-0.23	-0.28	-0.27	-0.08	0.04	0.14	0.26	0.32	0.18	0.27	0.32	0.22	
35-39	-0.02	0.00	0.00	-0.09	-0.13	-0.10	-0.04	-0.10	-0.19	-0.25	-0.08	0.03	0.03	0.08	0.12	0.00	0.11	0.22	0.15	
40-44	0.05	0.06	0.02	0.00	0.00	0.00	-0.01	0.00	-0.03	-0.05	0.00	0.07	0.02	0.00	0.01	-0.15	-0.01	0.15	0.15	
45-49	0.04	0.01	0.05	0.05	0.11	0.04	0.00	-0.03	0.02	-0.03	0.02	0.00	0.00	-0.09	-0.13	-0.29	-0.21	0.06	0.00	
50-54	0.05	-0.04	0.04	0.07	0.14	0.06	0.03	0.00	0.02	0.02	0.05	0.00	-0.12	-0.21	-0.23	-0.40	-0.36	-0.28	-0.21	
55-59	0.06	0.09	0.00	0.01	0.15	0.04	0.11	0.04	0.00	0.01	0.03	0.00	-0.12	-0.27	-0.29	-0.45	-0.45	-0.39	-0.37	
60-64	0.00	0.00	0.11	0.11	0.20	0.06	0.11	0.09	0.11	0.03	-0.06	-0.02	-0.18	-0.25	-0.36	-0.48	-0.46	-0.50	-0.52	
65-69	0.00	0.18	0.19	0.27	0.28	0.14	0.21	0.13	0.21	0.10	-0.01	-0.01	-0.06	-0.14	-0.22	-0.30	-0.33	-0.31	-0.44	
70-74	-0.10	0.05	0.16	0.19	0.26	0.02	0.21	0.14	0.20	0.09	0.00	-0.03	-0.07	-0.09	-0.09	-0.13	-0.19	-0.19	-0.32	
75-79							0.19	0.18	0.12	0.11	0.05	0.00	-0.07	0.00	0.00	0.02	0.00	-0.06	-0.19	
80-84							0.05	0.02	0.09	0.04	0.00	-0.09	-0.11	-0.17	-0.05	0.03	0.05	0.00	-0.16	

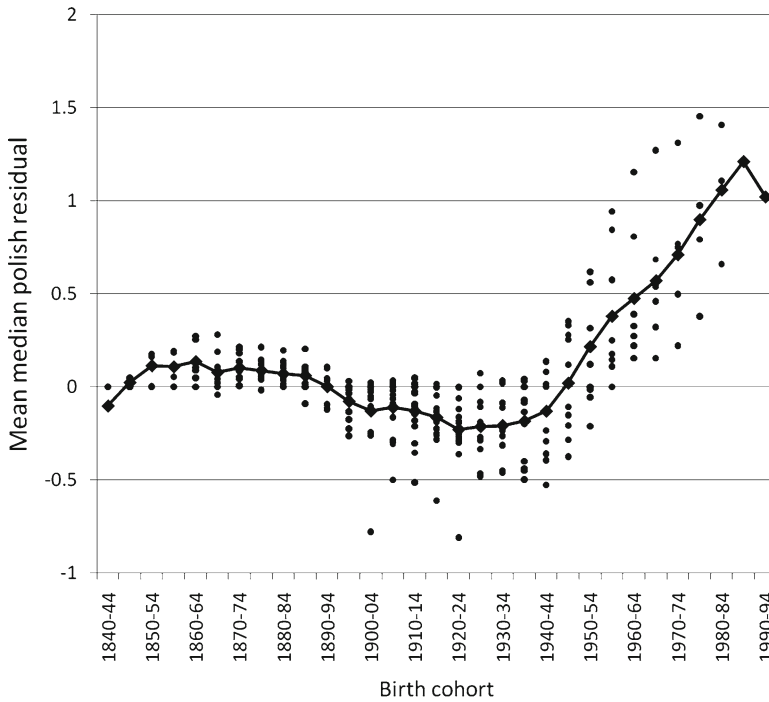


Fig. 22.6 Nonadditive influences of age and period by birth cohort on suicide mortality, 1910–2004

with the expectation of intercept μk approximately equal to zero. This step produces k beta estimates (one for each cohort category) reflecting the log rate that indicates a ratio of cohort effects (i.e. the ratio of the nonadditive effect for one cohort to that of the nonadditive effect for a reference cohort). The exponentiation of each beta estimate derived from equation three indicates the excess rate attributable to each cohort category. Each cohort category can then be compared to the referent cohort to obtain a relative estimate of the size of the cohort effect. The residuals from this model can be examined for violations of parametric assumptions.

In Table 22.3, we show the results of a regression of the median polish residuals on cohort category in the suicide data. The 1910–1914 cohort was used as the referent as it contained the most complete data. Among men, prior to the 1910–1914 cohort there is evidence of small but significant increases in cohort effect in those born throughout the second half of the nineteenth century (from approximately 1845 to 1894). Those born approximately 100 years later had a substantially stronger cohort effect compared to the reference cohort; men born between 1945 and 1949 had 1.2 times the risk of suicide (95% C.I. 1.08–1.25) and by the youngest (i.e., most recent) cohort of observation, the risk of suicide is over threefold that of the reference cohort.

Taken together, we have shown in this example an efficient method for estimating cohort effects in age–period contingency table data. First, we used graphical procedures to visually assess nonadditivity of age and period by cohort categories. In the suicide data, there was clear evidence of a cohort-specific suicide rate. Second, we use the median polish to remove the additive effects of age and period, and graphed the residuals by cohort category. We visually saw positive cohort effect in later born cohorts. Finally, we used a linear regression to quantify the cohort effects, documenting statistically significant cohort effects.

Table 22.3 Estimated risk ratio and 95% confidence interval for the effect of birth cohort on suicide mortality, 1910–2004

Year of birth	Risk ratio	95% Confidence interval	
1840–1844	1.03	0.86	1.22
1845–1849	1.17	1.03	1.32
1850–1854	1.28	1.15	1.42
1855–1859	1.27	1.15	1.40
1860–1864	1.30	1.20	1.42
1865–1869	1.23	1.14	1.32
1870–1874	1.26	1.18	1.36
1875–1879	1.24	1.16	1.33
1880–1884	1.22	1.14	1.31
1885–1889	1.21	1.13	1.29
1890–1894	1.14	1.07	1.22
1895–1899	1.05	0.99	1.12
1900–1904	1.00	0.94	1.06
1905–1909	1.02	0.96	1.08
1910–1914	1	Reference	
1915–1919	0.97	0.91	1.03
1920–1924	0.91	0.85	0.96
1925–1929	0.92	0.86	0.98
1930–1934	0.92	0.87	0.99
1935–1939	0.95	0.89	1.01
1940–1944	1.00	0.93	1.07
1945–1949	1.16	1.08	1.25
1950–1954	1.41	1.32	1.52
1955–1959	1.67	1.55	1.79
1960–1964	1.83	1.69	1.98
1965–1969	2.02	1.86	2.19
1970–1974	2.32	2.12	2.53
1975–1979	2.80	2.54	3.08
1980–1984	3.28	2.95	3.65
1985–1989	3.82	3.37	4.34
1990–1994	3.17	2.66	3.77

Implications of the Age–Period–Cohort Analysis for Our Understanding of Suicide in the USA

This age–period–cohort analysis of suicide in the USA documents that more recently born cohorts have a higher risk of suicide than previous born cohort. Future presentation and analysis of suicide surveillance data should focus on presenting information on trends by year of birth instead of year of death, as traditionally presented for surveillance summary. Presentation of data by year of birth rather than year of death provides an informative description of suicide data, summarizing complex trends across time and simplifying seemingly complicated patterns across age and year. Additionally, we did not find compelling evidence of period effects in suicide, suggesting that there is little variation in suicide outcome that does not vary across age. Therefore, secular trends in suicide can be best described according to birth cohorts.

The demonstration that risk varies across cohort as well as sex provides the basis for tests of competing hypotheses; for instance, note that the cohort effect ratio for males is significantly increased

in the two decades after World War II, whereas for females, the increase in cohort effect ratios begins after approximately 1960. Causal explanations should therefore focus on those social-level factors that influence men rather than women. There is evidence that traumatic war-related experiences have been associated with suicide risk especially among young men returning from the Vietnam war (Bullman and Kang 1996). Those born in the decades after World War II would come of age during the Korean and Vietnam wars; thus, these data are consistent with a hypothesis that Vietnam War exposure is a possible contributory factor to the elevated cohort-specific risk.

This analysis was motivated in part by the increased rate of suicide in middle-aged men and women from 1999 to 2004 (Paulozzi et al. 2007). These individuals would be in the cohort born 1945–1950; we did not find evidence of a systematic increase in the cohort effect for these individuals compared with the cohort effect for those born 1910–1914, indicating that the increased suicide rate in these groups is not a unique component of the birth cohort risk. An earlier analyses of these data also indicated an increase in the risk for suicide among men aged 40–44 that was constant across cohorts born 1951–1988 (Riggs et al. 1996). These independent analyses suggest that individuals in middle-age may possess unique psychosocial risk factors for suicide that thus far have not shown evidence of variation across cohorts.

We have also demonstrated a significantly increased cohort effect for recently born male and female cohorts. While there may be myriad plausible factors as explanations for the emergence of the cohort effect, the principle advantage of an age–period–cohort analysis is the ability to use the data to discriminate among competing hypotheses. For instance, two factors that have been cited as possible causes of suicide include firearm exposure and illegal drug use; the age–period–cohort method can help us differentiate the plausibility of these hypotheses. Firearm access and exposure is an important determinant of individual-level suicide risk, with a firearm present in approximately 61% of suicide deaths in the USA (Ajdacic-Gross et al. 2008); additionally, it has been documented that States with more firearms have higher rates of suicide (Miller and Hemenway 2008). Thus, the increased access among young people (Cook and Laub 1998) is consistent with the observed cohort effect in US suicide. Alternatively, national probability samples of young adults indicate increases in drug use in the population beginning in the 1960s (Johnston et al. 2007), yet rates of use were highest in the 1960s for marijuana and 1970s for cocaine. Thus, drug use does not explain the consistently increasing risk in cohorts born after the peak of drug use in the population. Firearm exposure and illegal drug use are only two of a number of factors that may be implicated in these patterns; income inequality, violence on television and in video games, and changing social norms may also be a part of a more comprehensive explanation of these rates, and the plausibility of these factors should be rigorously investigated.

We also note that the cohort effects identified for suicide parallel those observed for homicide. Sociological scholars have for centuries studied the relationship between homicide and suicide as comparable phenomena (Lester 1996). Recent research indicates that in the USA, there is an inverse relationship between suicide and homicide rates (Bills and Li 2005), yet this analysis indicates that cohorts born chronologically later in time are at an elevated risk of both suicide and homicide. Thus, theories regarding the increased risk of suicide should focus on those factors which may influence homicide as well; increased access to firearms, for instance, is consistent as an important determinant in both domains. Taken together, the present evidence offers substantially more information to rule out and in competing hypotheses regarding population-level causes of suicide in the US population.

Inference from the present data is limited by potential changes in the quality of national suicide statistics throughout the course of the twentieth century. Definitions, recording, and reporting of suicide have changed over time and across regions of the USA, though changes in ICD codes are not likely to have affected these results (Stockard and O'Brien 2002a, b); further, suicide is a stigmatized cause of death and is likely to be underreported. While there is no direct evidence to confirm the quality of the information provided by national statistics, we cannot rule out changes in reporting and recording practices as potentially influential in these results.

Future suicide surveillance using age–period–cohort analysis will increase in importance as the current cohorts of adolescents progress through the primary risk period for suicide attempt during a time period in which mental health policy is undergoing shifts. In 2004, the US Food and Drug Administration mandated a package warning on all prescription antidepressant labels due to possible increases in pediatric suicidality. Several studies have now documented an increase in the rate of youth suicide following the FDA action (Gibbons et al. 2007) and a decrease in pediatric depression treatment (Libby et al. 2007; Valuck et al. 2007). Further, the suicide rate among former active duty veterans involved in the ongoing war in Iraq is higher than the national average for similar age, gender, and race groups (Kang and Bullman 2008). Ongoing follow-up of young men engaging in and returning from combat will be important for national suicide surveillance. This underscores the necessity of presenting suicide information by year of birth, as examination of cohort effects can reveal important patterns in the data and the manifestation of cohort-specific changes in suicide risk can only be rigorously evaluated in an age–period–cohort framework.

Summary and New Directions and Innovations in Age–Period–Cohort Analysis

To summarize the chapter thus far, age–period–cohort analysis is a conceptually useful method to uncover hidden patterns in rates over time. Despite the conceptual utility, statistically the quantification of age–period–cohort effects is problematic. This is due to the extreme colinearity among the three variables, and has been termed the “identification” problem. Despite the fundamental lack of a solution to this identification problem, many statistical techniques have been offered to estimate the contribution of age, period, and cohort to disease rates over time. Each statistical method requires varying degrees of assumptions about the underlying distribution of the data and the relation among age, period, and cohort effects. Comparisons across method can sometimes result in conflicting findings. For example, we compared three methods for age–period–cohort analysis on trends over time in the prevalence of obesity in the USA and found that commonly used constraint-based models suggested the presence of a cohort effect, whereas a model based on the Holford approach as well as the median polish approach suggested no cohort effect (Keyes et al. 2010). The interpretation of the constraint-based model does not necessarily conflict with the median polish or Holford results, however; the interpretation depends on the definition of a cohort effect. The epidemiologic definition of a cohort effect suggests that a cohort effect occurs when different distributions of disease arise from a changing or new environmental cause affecting age groups differently. A more sociologically oriented definition, however, is that birth cohorts index the conditions, barriers, and resources that each cohort is born into and in which they live their collective lives, which may uniquely shape the patterns and experiences of health and mortality for that cohort. Adopting the epidemiological definition of a cohort effect, the results of our method comparison suggest that the environmental causes of obesity have not varied across age groups to cause the emergence of the obesity epidemic in America. If viewing the results from a sociological definition, however, we may conclude that there are structural factors unique to the experience of each cohort as they progress through the life course that would produce higher obesity rates independently of the concurrent environmental conditions that ubiquitously impact the population at large.

Taken as a whole, this literature suffers from a serious limitation: ages, periods, and cohorts are distant proxies for constructs that mediate the explanation of these trends. Statistical problems in the identification of age–period–cohort effects arise due to a direct mathematical relationship of the three variables that are almost always used to estimate age–period–cohort effects (Cohort = Period – Age). If more proximal constructs can be tested rather than the three variables of age, period, and cohort, the resulting hypothesis tests are more conceptually meaningful and meth-

odologically sound (Preston and Wang 2006; Winship and Harding 2008). Hobcraft et al. (1982) and others (Winship and Harding 2008) have noted that the constructs which we use age, period, and cohort to represent are often distinct proxies for the true constructs of interest. Ambiguity in interpretation of age–period–cohort models can arise when we enter the analysis unsure of the specific constructs we aim to represent with year of birth. Ultimately, age–period–cohort models do not test hypotheses about the effects of environmental or historical influences; instead, they organize data and provide useful mathematical formulae for summarizing disease rates over time. Researchers with specific hypotheses about the causes through which age, period, and cohort effects arise will be better armed to achieve salient public health conclusions if the constructs can be directly measured and tested. For example, a recent age–period–cohort analysis demonstrated that sex differences in mortality rates across time can be fully accounted for by changing sex differences in cohort-specific smoking patterns (Preston and Wang 2006). Statistical identification problems were not an issue, because specific variables measuring the smoking patterns of cohorts were used to measure cohort effects (rather than measuring cohort effects through year of birth alone).

Innovative directions in age–period–cohort analysis include models to integrate mechanistic variables that more directly test the causes through which disease patterns aggregate by age, period, and cohort. Unfortunately, often these variables are unmeasured, leaving the analyst with only age, period, and year of birth as markers for these underlying constructs. However, careful thinking about hypotheses will greatly inform model selection and model building. Use of graphical approaches and straightforward analytic tools such as the multi-phase method will aid researchers in the attempt to tease apart rates over time and inform public health intervention and prevention efforts by identifying the population-level structure of rates over time by age, period, and birth cohort effects.

References

- Ajdacic-Gross, V., Bopp, M., Gostynski, M., Lauber, C., Gutzwiller, F., & Rössler, W. (2006). Age-period-cohort analysis of Swiss suicide data, 1881–2000. *European Archives of Psychiatry and Clinical Neuroscience*, 256(4), 207–14.
- Ajdacic-Gross, V., Weiss, M. G., et al. (2008). Methods of suicide: International suicide patterns derived from the WHO mortality database. *Bulletin of the World Health Organization*, 86(9), 726–32.
- Berzuini, C., & Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, 13(8), 823–38.
- Ben-Shlomo, Y., D. Kuh (2002). A life course approach to chronic disease epidemiology: Conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31(2), 285–293.
- Bills, C. B., & Li, G. (2005). Correlating homicide and suicide. *International Journal of Epidemiology*, 34(4), 837–45.
- Bullman, T. A., & Kang, H. K. (1996). The risk of suicide among wounded Vietnam veterans. *American Journal of Public Health*, 86(5), 662–7.
- Carstensen, B. (2007). Age-period-cohort models for the Lexis diagram. *Statistics in Medicine*, 26(15), 3018–45.
- Case, R. A. M. (1956). Cohort analysis of mortality rates as an historical or narrative technique. *British Journal of Preventative and Social Medicine*, 10, 159–171.
- Clayton, D., & Schifflers, E. (1987). Models for temporal variation in cancer rates. I: Age-period and age-cohort models. *Statistics in Medicine*, 6(4), 449–67.
- Cleries, R., Ribes, J., et al. (2006). Time trends of breast cancer mortality in Spain during the period 1977–2001 and Bayesian approach for projections during 2002–2016. *Annals of Oncology*, 17(12), 1783–91.
- Cook, P., & Laub, J. (1998). *The unprecedented epidemic in youth violence*. Chicago: University of Chicago Press.
- Derrick, V. P. A. (1927). Observations on (1) errors of age in the population statistics of England and Wales, and (2) the changes in mortality indicated by the national records. *Journal of the Institute of Actuaries*, LVIII, 117–159.
- Doll, R. (1971). The age distribution of cancer: Implications for models of carcinogenesis. *Journal of the Royal Statistical Society*, 134, 133–155.
- Evans, J. G., Seagroatt, V., et al. (1997). Secular trends in proximal femoral fracture, Oxford record linkage study area and England 1968–1986. *Journal of Epidemiology and Community Health*, 51(4), 424–9.

- Fienberg, S. E., & Mason, W. M. (1979). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, 10(1), 1–67.
- Frost, W. H. (1939). The age selection of morality from tuberculosis in successive decades. *American Journal of Hygiene*, 30, 91–96.
- Gibbons, R. D., Brown, C. H., et al. (2007). Early evidence on the effects of regulators' suicidality warnings on SSRI prescriptions and suicide in children and adolescents. *The American Journal of Psychiatry*, 164(9), 1356–63.
- Glenn, N. D. (1976). Cohort analysts' futile quest: Statistical attempts to separate age, period, and cohort effects. *American Sociological Review*, 41, 900–905.
- Granizo, J. J., Guallar, E., & Rodriguez-Artalejo, F. (1996). Age-period-cohort analysis of suicide mortality rates in Spain, 1959–1991. *International journal of epidemiology*, 25(4), 814–20.
- Greenberg, B. G., Wright, J. J., et al. (1950). A technique for analyzing some factors affecting the incidence of syphilis. *American Statistical Association Journal*, 45(251), 373–399.
- Gunnell, D., Middleton, N., et al. (2003). Influence of cohort effects on patterns of suicide in England and Wales, 1950–1999. *The British Journal of Psychiatry*, 182, 164–70.
- Hall, A. J., Yee, L. J., et al. (2002). Life course epidemiology and infectious diseases. *International Journal of Epidemiology*, 31(2), 300–301.
- Hobcraft, J., Menken, J., et al. (1982). Age, period, and cohort effects in demography: A review. *Population Index*, 48(1), 4–43.
- Holford, T. R. (1991). Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review of Public Health*, 12, 425–57.
- Holford, T. R. (1992). Analysing the temporal effects of age, period and cohort. *Statistical Methods in Medical Research*, 1(3), 317–37.
- Joe, S. (2006). Explaining changes in the patterns of black suicide in the United States from 1981 to 2002: An age, cohort, and period analysis. *Journal of Black Psychology*, 32(3), 262–284.
- Johnston, L. D., O'Malley, P. M., et al. (2007). *Monitoring the Future national survey results on drug use, 1975–2006: Volume I, Secondary school students*. Bethesda, MD: National Institute on Drug Abuse.
- Kang, H. K., & Bullman, T. A. (2008). Risk of suicide among US veterans after returning from the Iraq or Afghanistan war zones. *The Journal of the American Medical Association*, 300(6), 652–3.
- Kannus, P., Niemi, S., et al. (1999). Hip fractures in Finland between 1970 and 1997 and predictions for the future. *Lancet*, 353(9155), 802–5.
- Kermack, W. O., McKendrick, A. G., et al. (1934). Death-rates in great Britain and Sweden. Some general regularities and their significance. *Lancet*, 31, 698–703.
- Keyes, K. M., & Li, G. (2010). A multiphase method for estimating cohort effects in age-period contingency table data. *Annals of Epidemiology*, 20(10), 779–85.
- Keyes, K. M., Utz, R. L., et al. (2010). What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971–2006. *Social Science and Medicine*, 70(7), 1100–8.
- Korteweg, R. (1951). The age curve in lung cancer. *British Journal of Cancer*, 5, 21–27.
- Kupper, L. L., Janis, J. M., et al. (1985). Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Disease*, 38(10), 811–30.
- Last, J. M. (2001). *A dictionary of epidemiology* (4th ed.). New York, NY: Oxford University Press.
- Lee, W. C., & Lin, R. S. (1996). Autoregressive age-period-cohort models. *Statistics in Medicine*, 15(3), 273–81.
- Lester, D. (1996). *Patterns of homicide and suicide in the World*. Commack, NY: Nova Science Publishers, Inc.
- Libby, A. M., Brent, D. A., et al. (2007). Decline in treatment of pediatric depression after FDA advisory on risk of suicidality with SSRIs. *The American Journal of Psychiatry*, 164(6), 884–91.
- Lynch, J. & Smith, G. D., (2005). A life course approach to chronic disease epidemiology. *Annual Review of Public Health*, 26, 1–35.
- Lynskey, M., Degenhardt, L., & Hall, W. (2000). Cohort trends in youth suicide in Australia 1964–1997. *The Australian and New Zealand journal of psychiatry*, 34(3), 408–12.
- Macmahon, B. & Terry, W. D. (1958). Application of cohort analysis to the study of time trends in neoplastic disease. *Journal of Chronic Diseases*, 7(1), 24–35.
- Martinez-Schnell, B., & Zaidi, A. (1989). Time series analysis of injuries. *Statistics in Medicine*, 8(12), 1497–508.
- Mason, K. O., Mason, W. M., et al. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review*, 38, 242–258.
- McNally, R. J., Alexander, F. E., et al. (1997). A comparison of three methods of analysis for age-period-cohort models with application to incidence data on non-Hodgkin's lymphoma. *International Journal of Epidemiology*, 26(1), 32–46.
- Miech, R., Koester, S., et al. (2011). Increasing U.S. mortality due to accidental poisoning: The role of the baby boom cohort. *Addiction*. 106(4), 806–815.

- Miller, M., & Hemenway, D. (2008). Guns and suicide in the United States. *The New England Journal of Medicine*, 359(10), 989–91.
- Murphy, G. E., & Wetzel, R. D. (1980). Suicide risk by birth cohort in the United States, 1949 to 1974. *Archives of General Psychiatry*, 37(5), 519–23.
- Nakamura, T. (1986). Bayesian cohort models for general cohort table analysis. *Annals of the Institute of Statistical Mathematics*, 38(Part B), 353–70.
- O'Brien, R. M. (2000). Age period cohort characteristic models. *Social Science Research*, 29, 123–139.
- Odagiri, Y., Uchida, H., et al. (2011). Gender differences in age, period, and birth-cohort effect on suicide mortality rate in Japan 1985–2006. *Asia-pacific Journal of public Health*, 23(4), 581–7.
- Pampel, F. C. (1996). Cohort size and age-specific suicide rates: A contingent relationship. *Demography*, 33(3), 341–355.
- Paulozzi, L., Crosby, A., et al. (2007). Increases in age-group – specific injury mortality – United States, 1999–2004. *Morbidity and Mortality Weekly Report*, 56(49), 1281–1284.
- Preston, S. H., & Wang, H. (2006). Sex mortality differences in the United States: The role of cohort smoking patterns. *Demography*, 43(4), 631–46.
- Riggs, J. E., McGraw, R. L., et al. (1996). Suicide in the United States, 1951–1988: Constant age-period-cohort rates in 40- to 44-year-old men. *Comprehensive Psychiatry*, 37(3), 222–5.
- Robertson, C., & Boyle, P. (1986). Age, period and cohort models: The use of individual records'. *Statistics in Medicine*, 5, 527–538.
- Robertson, C., Gandini, S., et al. (1999). Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology*, 52, 569–583.
- Rodrigues, N. C., & Werneck, G. L. (2005). Age-period-cohort analysis of suicide rates in Rio de Janeiro, Brazil, 1979–1998. *Social Psychiatry and Psychiatric Epidemiology*, 40(3), 192–6.
- Rosenbauer, J., & Strassburger, K. (2007). Letter to the Editor: Comments on "Age-period-cohort models for the Lexis diagram". *Statistics in Medicine*, 26, 3018–3045.
- Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review*, 30(6), 843–61.
- Samelson, E. J., Zhang, Y., et al. (2002). Effect of birth cohort on risk of hip fracture: Age-specific incidence rates in the Framingham Study. *American Journal of Public Health*, 92(5), 858–62.
- Selvin, S. (1996). *Statistical analysis of epidemiologic data*. New York: Oxford University Press.
- Shahpar, C., & Li, G. (1999). Homicide mortality in the United States, 1935–1994: Age, period, and cohort effects. *American Journal of Epidemiology*, 150(11), 1213–22.
- Stockard, J., & O'Brien, R. M. (2002a). Cohort effects on suicide rates: International variations. *American Sociological Review*, 67, 854–872.
- Stockard, J., & O'Brien, R. M. (2002b). Cohort variations and changes in age-specific suicide rates over time: Explaining variations in youth suicide. *Social Forces*, 81(2), 605–642.
- Susser, M. (1961). Environmental factors and peptic ulcer. *Practitioner*, 186(302–311).
- Susser, M. (2001). Commentary: The longitudinal perspective and cohort analysis. *International Journal of Epidemiology*, 30(4), 684–7.
- Tarone, R. E., & Chu, K. C. (1992). Implications of birth cohort patterns in interpreting trends in breast cancer rates. *Journal of the National Cancer Institute*, 84(18), 1402–10.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MS: Addison-Wesley Publishing Company.
- US Public Health Service. (1999). *The surgeon general's call to action to prevent suicide*. Washington, DC: U. S. P. H. Service.
- Valuck, R. J., Libby, A. M., et al. (2007). Spillover effects on treatment of adult depression in primary care after FDA advisory on risk of pediatric suicidality with SSRIs. *The American Journal of Psychiatry*, 164(8), 1198–205.
- Wickramaratne, P. J., Weissman, M. M., et al. (1989). Age, period and cohort effects on the risk of major depression: Results from five United States communities. *Journal of Clinical Epidemiology*, 42(4), 333–43.
- Winship, C., & Harding, D. J. (2008). A general strategy for the identification of age, period, cohort models: A mechanism based approach. *Sociological Methods and Research*, 36(3), 362–401.
- Yang, Y., Fu, W. J., et al. (2004). A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, 34, 75–110.
- Yang, Y., & Land, K. C. (2006). *A mixed models approach to age-period-cohort analysis of repeated cross-section surveys: Trends in verbal test scores*. *Sociological Methodology*. R. M. Stolzenberg. Boston: Blackwell Publishing. 36.
- Zheng, T., Holford, T. R., et al. (1995). Time trend in pancreatic cancer incidence in Connecticut, 1935–1990. *International Journal of Cancer*, 61(5), 622–7.
- Zheng, T., Holford, T. R., et al. (1996). Time trend and age-period-cohort effect on incidence of bladder cancer in Connecticut, 1935–1992. *International Journal of Cancer*, 68(2), 172–6.

Chapter 23

Multilevel Modeling

David E. Clark and Lynne Moore

Introduction

Mortality is the most frequently modeled outcome in injury research. It is easy to recognize, relatively free from measurement error, and fundamentally interesting. Injury researchers in public health or clinical medicine have become familiar with logistic regression as a standard way to model a binary outcome like mortality (or alternatively survival). Many other outcomes encountered in injury research can also be considered binary, such as the occurrence of a serious complication or an extended length of stay in hospital.

Traditionally, mortality modeling has been based on single-level logistic regression models, which assume that individual observations are independent and have the same error variance. However, individual observations in epidemiologic or health services data often occur naturally within groups that have certain properties in common. If an observation is more likely to be correlated with observations in the same group than it is to be correlated with observations in other groups, then it may be more appropriate to devise a statistical model that does not assume that all observations share a single error variance.

In the field of injury prevention and control, “clustering” of this sort might arise, for example, when patient outcomes from different hospitals are compared, when occupational injuries are recorded for different industries, or when repeated episodes of violence are suffered by the same person. One categorization may even be nested within another, such as when traffic fatalities are grouped by county and counties are grouped by state. Models incorporating such a structure could be considered “hierarchical.” However, we might also be interested in multiple categorizations that

Disclaimers

Content reproduced from the National Trauma Data Bank remains the full and exclusive copyrighted property of the American College of Surgeons. The American College of Surgeons is not responsible for any claims arising from works based on the original data, text, tables, or figures.

D.E. Clark, MD (✉)

Maine Medical Center, Suite 210, 887 Congress Street, Portland, ME 04102, USA

Harvard Injury Control Research Center, Harvard School of Public Health, Boston, MA, USA

e-mail: clarkd@mmc.org

L. Moore, PhD

Département de Médecine Sociale et Préventive, Université Laval, Québec City, QC, Canada,

Centre Hospitalier Affilié Universitaire de Québec, Pavillon Enfant-Jésus, 1401,

18ème rue, Québec City, QC, Canada G1J 1Z4

e-mail: lynne.moore.cha@sss.gouv.qc.ca

are not nested, such as traffic crashes grouped by geographic area and also by vehicle type. Some have therefore preferred the less restrictive term “multilevel models,” although this still preserves the idea that one “level” is higher than another. The fundamental mathematical innovation in “hierarchical models” or “multilevel models” is to consider one or more terms of the model to be random rather than fixed, leading to the even more general term “random effects models”; because there are usually both random and fixed coefficients, these are sometimes called “mixed models,” although the latter term is so general that it becomes almost meaningless.

This chapter will use the general term “multilevel” to mean a data structure that allows individual observations to be categorized into one or more groupings that are of interest to the researcher. However, attention will be mostly limited to a regression model with only one level (or hierarchy) of categorization above the individual observations, constructed only by considering the intercept to be a random variable; this type of multilevel model can therefore be called a “two-level random intercept model.”

For studies where observations are grouped, but the groups are not of direct interest, the different levels of variation can be taken into account using other techniques such as generalized estimating equations or robust variance estimation (Cook and DeMets 2008; Rabe-Hesketh and Skrondal 2008). This might be appropriate, for example, when there are multiple victims from the same crash, in which case the research would not usually compare one crash to another. However, when the variability among the groups is itself of interest, for example, comparing crashes grouped by different intersections or road sections, then it might be useful to employ a more descriptive multilevel model that could enable predictions about the specific locations.

Random intercept models are the simplest form of multilevel models. They have been widely used to compare hospital mortality for cardiac surgery centers (Normand et al. 1997; Shahian et al. 2001; Austin et al. 2003; Krumholz et al. 2006; Normand and Shahian 2007) and have more recently been proposed for trauma centers (Glance et al. 2010; Moore et al. 2010a; Clark et al. 2010a). In addition to their applications for hospital profiling, multilevel logistic regression models have been used to study traffic crash mortality in Norway, grouping persons by vehicle and geographic location, and in France, grouping persons by vehicle and crash event (Jones and Jorgensen 2003; Lenguerrand et al. 2006). Multilevel logistic regression models have also been used to study the probability of certain crash types in the state of Georgia, grouping events by specific roadway intersections, and crash severity in the state of Washington, grouping events by specific roadway segments (Kim et al. 2007; Milton et al. 2008).

The goal of this chapter is to describe how random intercept multilevel logistic regression models are similar to yet differ from standard logistic regression models, explain why they may be preferable, and discuss some practical methods to implement them. The practical implications of using multilevel modeling will be demonstrated by comparing trauma center profiling results generated with a multilevel logistic regression model to those obtained from a single-level logistic regression model. Along the way, some useful mathematical approximations will be derived, and the accuracy of these approximations will be evaluated using a working example. Simplifications of the theory and practice using the methods and approximations suggested in this chapter may reduce the complexity of the calculations and concepts necessary to implement these models when appropriate and increase their acceptance among injury control researchers.

Single-Level Modeling of Injury Mortality

Mortality and other binary outcomes are usually modeled using logistic regression. If the data refer to each patient i in each hospital j , standard single-level logistic regression models estimate the log odds (logit) of the probability of death,

$$y_{ij} = \text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = b_0 + b_1x_{ij1} + \cdots + b_kx_{ijk}, \quad (23.1)$$

where p_{ij} is the estimated probability of death, b_0 is the intercept, $b_1 \dots b_k$ are the regression coefficients associated with covariates $X_1 \dots X_k$, and $x_{ij1} \dots x_{ijk}$ are the values of these covariates for the i th patient in hospital j . The probability of death (p_{ij}) can then be predicted using the antilogit function, that is,

$$p_{ij} = \text{logit}^{-1}(y_{ij}) = \frac{e^{y_{ij}}}{1 + e^{y_{ij}}} = \frac{1}{1 + e^{-y_{ij}}}. \quad (23.2)$$

The logistic model assumes that the errors in estimating p_{ij} follow a binomial distribution (since the outcomes can only be 1 or 0). The standard logistic model also assumes that the outcomes for individual patients are not influenced by their membership in any group.

For the specific example of hospital profiling, a single-level logistic regression model of the sort described in (23.1) is typically used to generate a ratio of the observed to expected (O/E) number of deaths for each hospital under evaluation. The expected number of deaths can be obtained from a logistic regression model fitted to a reference population, which usually but not necessarily contains hospital j . The probability of death for patient i in hospital j (p_{ij}) is estimated by applying the coefficients of the model to the risk factors of the patient. The expected number of deaths for hospital j is then the sum of p_{ij} for all patients in that hospital.

Clinicians can intuitively understand an observed/expected (O/E) ratio, which is sometimes referred to as a standardized mortality ratio (SMR). However, a SMR must be interpreted with caution because it has been risk-adjusted using the case mix of the hospital under evaluation. The SMR is therefore not strictly comparable from one hospital to another, especially in situations where a hospital has a very different case mix from the reference population (Rothman 1986; Chan et al. 1988; Jones et al. 1995; Moore et al. 2010b). Sometimes the O/E ratio is multiplied by the overall crude mortality in a reference population to obtain a risk-adjusted mortality rate. However, if a hospital has an unusual case mix resulting in a very large O/E ratio, this may produce absurd results (e.g., a predicted incidence of mortality greater than 100%).

The calculation of an interval estimate (confidence limits) for a SMR (or any other measurement of hospital effect) is as important as the calculation of the point estimate. Hospitals whose confidence limits do not include the reference value (e.g., SMR=1) may be labeled as “outliers,” and therefore singled out for praise or criticism.

If the probability of death p_{ij} for each patient follows a Bernoulli distribution (outcome equals 1 with probability p_{ij} , and 0 otherwise), the variance will be $p_{ij}(1 - p_{ij})$ for each case. Further assuming that these individual distributions are independent, the expected number of deaths in a given hospital (E_j) will have a variance (v_j) equal to the sum of its individual patient variances (Flora 1978). That is,

$$v_j = \text{var}(E_j) = \sum_{i=1}^{n_j} p_{ij}(1 - p_{ij}), \quad (23.3)$$

where n_j is the number of patients in hospital j . A score test can be constructed as

$$z_j = \frac{O_j - E_j}{\sqrt{v_j}},$$

where O_j is the observed number of deaths in hospital j , and the resulting “Z statistic” (assuming that Z has a standard normal distribution) has been widely used in the past as a measure of trauma center performance (Champion et al. 1990). A lower confidence limit (LCL) and an upper confidence limit (UCL) for O_j can also be constructed as

$$\begin{aligned} \text{LCL}(O_j) &= O_j - 1.96\sqrt{v_j} \\ \text{UCL}(O_j) &= O_j + 1.96\sqrt{v_j}. \end{aligned}$$

However, the assumption of a (symmetrical) normal distribution may be inappropriate if E_j is small and/or n_j is small and may produce absurd results (e.g., $LCL < 0$).

Another formulation of standard logistic regression would create an indicator variable ($x_j = 1$ for patients in hospital j and $x_j = 0$ otherwise) and estimate a separate model for each hospital substituting

$$b_0 = b_{00} + b_j x_j, \quad (23.4)$$

into (23.1). If the reference database is very large compared to any individual hospital, the intercept b_{00} for patients other than those in hospital j will be little different from b_0 , and the fixed coefficients $b_1 \dots b_k$ will also be little different. The coefficient b_j and confidence limits constructed using the Z -score from a Wald test performed by standard regression software can be used as a measure of effect for hospital j (DeLong et al. 1997; Pollock 1999).

The score test leads to a confidence interval for O_j under the hypothesis that $O_j = E_j$, whereas the Wald test leads to a confidence interval around the observed value of O_j . Since both are based on the assumption of normality in large samples, they will give very similar determinations of significance (Z -scores) (Rothman 1986; Cook and DeMets 2008). If the indicator variable regression approach is used, the expected number of deaths for hospital j can be estimated as

$$E_j = \sum \text{logit}^{-1}(b_{00} + b_1 x_{ij1} + \dots + b_k x_{ijk}).$$

Similarly, a confidence interval can be constructed for O_j that has the desired property of symmetry on the logit scale while being limited to positive numbers on the probability scale, namely,

$$LCL(O_j) = \sum \text{logit}^{-1}(b_{00} + LCL(b_j) + b_1 x_{ij1} + \dots + b_k x_{ijk}),$$

$$UCL(O_j) = \sum \text{logit}^{-1}(b_{00} + UCL(b_j) + b_1 x_{ij1} + \dots + b_k x_{ijk}).$$

One practical limitation of the indicator variable regression method is that an effect cannot be estimated if hospital j has zero mortality. If there are many hospitals, it may also take some time to calculate a separate regression for each one, although a computer can be programmed to perform this repetitive task.

A simpler approach to constructing confidence intervals for the SMR assumes that the observed number of deaths in a given hospital follows a Poisson distribution (Ulm 1990). A Poisson distribution with parameter O_j provides a good approximation to a binomial distribution with parameters n_j and p_j , assuming that n_j is large, that p_j is the mean of p_{ij} for hospital j , and that O_j is approximately equal to $n_j \times p_j$. However, p_{ij} for some patient populations (e.g., trauma) may not be distributed binomially around an expected value p_j but instead may be separated into two subpopulations with p_{ij} either near 0 or near 1, in which case the assumption of a Poisson distribution may be inappropriate.

In recent years, the SMR (or some other measure of hospital performance) has often been displayed on a vertical axis with its confidence interval, ranking the hospitals from best to worst on the horizontal axis to produce a "rank plot." Because of its graphical appearance, this is sometimes called a "caterpillar plot," and indeed if there are very many hospitals, it may be as hard to distinguish them as the legs of a caterpillar. Furthermore, graphing or even listing ranks is not very useful for the majority of hospitals in the middle of the group where a large difference in rank may actually represent only a small difference in measured outcomes. If the statistical model is changed, hospital ranks may change, and it can therefore be difficult to evaluate the effect of changing models by comparing graphs of this type.

A more informative display may be obtained using a "funnel plot," in which an outcome and its confidence limits are graphed against some factor associated with greater precision and narrower

confidence limits (Spiegelhalter 2005; Kirkham and Bouamra 2008). Outliers are easily identified as those observations outside of the resulting funnel-shaped confidence limits. If the SMR for each hospital is graphed against its observed mortality, the position of each hospital on the x -axis will be fixed, so it is easier to determine how the graphical results are affected by any change in statistical models.

If confidence limits have been calculated for the observed mortality of hospital j , then division by E_j allows us to obtain confidence limits for the observed SMR (O_j/E_j). The null hypothesis $O_j = E_j$ can be tested graphically as

$$\begin{aligned} \text{LCL}(O_j) < E_j < \text{UCL}(O_j), \\ \frac{\text{LCL}(O_j)}{E_j} < \frac{E_j}{E_j} < \frac{\text{UCL}(O_j)}{E_j}, \end{aligned}$$

and these limits can be plotted on a rank graph with the null hypothesis that the confidence interval will contain 1. For a funnel graph, the null hypothesis of no hospital effect can be restated as

$$\frac{1}{\text{LCL}(O_j)} > \frac{1}{E_j} > \frac{1}{\text{UCL}(O_j)},$$

so confidence limits for the observed SMR can be constructed as

$$\frac{O_j}{\text{UCL}(O_j)} < \frac{O_j}{E_j} < \frac{O_j}{\text{LCL}(O_j)}, \quad (23.5)$$

and these can be plotted with the null hypothesis that the confidence interval will contain the observed SMR (O_j/E_j).

Multilevel Modeling of Injury Mortality

Suppose the overall mortality for patients with gunshot wounds treated in American trauma centers is known, but a particular center reports that last year it had no mortality (taking an extreme case). An instinctive prediction of the mortality for gunshot victims treated in this center in subsequent years will not be zero (Dimick and Welch 2008), although it might be somewhat less than the national average. Empirical Bayes estimation is a theoretically consistent method to make predictions in this setting by combining weighted estimates of the observed data from one center and prior knowledge about similar hospitals. Multilevel modeling offers a method to obtain appropriate weights for empirical Bayes estimation.

In the multilevel approach, a hospital's contribution to mortality predicted by this equation is a "shrunk" estimate that weights the observed mortality by its reliability; that is, the hospital-specific effect (u_{0j}) is shrunk or "pulled" toward zero, with the greatest shrinkage among hospitals producing the least data for estimation. Theoretical considerations and empirical evidence (Raudenbush and Bryk 2002; Clark et al. 2010a) demonstrate that, on average, the shrunk estimates will tend to be more accurate predictors of future performance than those based on single-level regression.

The first level of a random intercept multilevel logistic regression model is similar to the single-level logistic regression model of (23.1), estimating the log odds of hospital death as

$$y_{ij} = \text{logit}(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = b_{0j} + b_1 x_{ij1} + \dots + b_k x_{ijk}. \quad (23.6)$$

However, in this case, the intercept b_{0j} is specific for hospital j , as determined by a second level of the model

$$b_{0j} = b_{00} + u_{0j}, \tag{23.7}$$

where b_{00} is the mean intercept across hospitals and u_{0j} is the random effect of hospital j on the mean, generally assumed to be normally distributed with mean of zero and a variance v_2 estimated from the data. This combination of (23.7) with (23.6) is analogous to the combination of (23.4) with (23.1) using the indicator variable approach to single-level logistic regression.

The theoretically optimal weights to be used for empirical Bayes estimation are related to the “reliability” of the level 1 (patient-level) estimate for hospital j , and a “reliability coefficient” may be symbolized as λ_j (Snijders and Bosker 1999; Raudenbush and Bryk 2002). This quantity may be derived by considering the “precision” of an estimate to be the inverse of its variance, and then calculating λ_j as the proportion of the total precision attributable to the patient-level data from hospital j . That is,

$$\lambda_j = \frac{1/v_{1j}}{1/v_{1j} + 1/v_2} = \frac{v_2}{v_{1j} + v_2},$$

where v_{1j} is the level 1 variance (calculated using (23.3) for the patients in hospital j) and v_2 is the level 2 variance (the variance among hospital means).

If the expected number of deaths for hospital j has been estimated using data from a reference group of hospitals, a shrunken prediction of the number of deaths for hospital j may be defined as

$$S_j = \lambda_j O_j + (1 - \lambda_j) E_j = E_j + \lambda_j (O_j - E_j). \tag{23.8}$$

A shrunken standardized mortality rate (sSMR) may be obtained by dividing the shrunken prediction by the expected incidence of mortality, that is,

$$\frac{S_j}{E_j} = \frac{E_j + \lambda_j (O_j - E_j)}{E_j} = 1 + \lambda_j (\text{SMR} - 1). \tag{23.9}$$

SMR in (23.9) refers to the nonshrunken SMR, that is, O_j/E_j . Since λ_j is used to “shrink” the raw estimate toward the group mean, it can also be called a “shrinkage coefficient” for hospital j .

Expressions (23.6–23.7) can be used to estimate

$$S_j = \sum_j \text{logit}^{-1}(b_{0j} + b_1 x_{ij1} + \dots + b_k x_{ijk}),$$

and confidence limits can be constructed as

$$\text{LCL}(S_j) = \sum \text{logit}^{-1}(b_{00} + \text{LCL}(u_{0j}) + b_1 x_{ij1} + \dots + b_k x_{ijk}),$$

$$\text{UCL}(S_j) = \sum \text{logit}^{-1}(b_{00} + \text{UCL}(u_{0j}) + b_1 x_{ij1} + \dots + b_k x_{ijk}).$$

When the outcome is continuous instead of binary, a linear random intercept multilevel model is simply

$$y_{ij} = b_{0j} + b_1 x_{ij1} + \dots + b_k x_{ijk} + e_{ij},$$

where e_{ij} is an error term assumed to follow a normal distribution with mean 0 and variance σ^2 estimated from the data. As in (23.7), the intercept b_{0j} is determined by a second level

$$b_{0j} = b_{00} + u_{0j},$$

where b_{00} is the mean intercept across hospitals and u_{0j} is the random effect of hospital j on the mean. The random variable u_{0j} is usually assumed to follow a normal distribution with mean 0 and variance v_2 estimated from the data. A major difference in logistic regression is that the errors are not estimated for y_{ij} but assume that p_{ij} (23.2) follows a Bernoulli distribution. This additional conceptual and computational complexity will be explored when considering approximations to the shrinkage coefficient later in this chapter.

Estimates from the random-intercept multilevel model can be adapted relatively easily to provide confidence limits for the SMR that incorporate the conservatism of empirical Bayes theory. Using the null hypothesis $S_j = E_j$ (23.5) can be modified to obtain

$$\frac{S_j}{\text{UCL}(S_j)} < \frac{S_j}{E_j} < \frac{S_j}{\text{LCL}(S_j)},$$

and combining this with the definition of S_j from (23.8) gives

$$\frac{S_j}{\text{UCL}(S_j)} < \frac{\lambda_j O_j + (1 - \lambda_j) E_j}{E_j} < \frac{S_j}{\text{LCL}(S_j)}.$$

Algebraic rearrangements of the above result in

$$1 - \left(\frac{1}{\lambda_j} \right) \left(1 - \frac{S_j}{\text{UCL}(S_j)} \right) < \frac{O_j}{E_j} < 1 + \left(\frac{1}{\lambda_j} \right) \left(\frac{S_j}{\text{LCL}(S_j)} - 1 \right), \quad (23.10)$$

which can be depicted along with the unshrunk SMR on a funnel graph. This avoids having to explain the shrunken SMR while preserving the conservatism of the empirical Bayes approach as a result of expanding the confidence interval by the factor $1/\lambda_j$.

Empirical Bayes estimation grows out of ideas developed using Bayesian theory, but strictly speaking it is not a Bayesian analysis (O'Hagan 1994). While empirical Bayes analysis is based only on observed data, a "fully Bayesian" analysis would begin with subjective specifications of a prior distribution for each random variable and would update them using the observed data and Bayes' theorem to derive posterior distributions. This generally requires simulation methods for computation, which may be very time-consuming. In many applications, the absence of prior information leads to the use of "noninformative" prior distributions (with large variances). In such circumstances, the prior distribution will have essentially no influence on the posterior distribution, so results will be similar to those obtained from empirical Bayes methods (Browne and Draper 2006).

Even without fully Bayesian computation, and even when limited to two-level random intercept models, the multilevel approach does make the statistical methodology more complex. In addition to the textbooks cited in this chapter, several monographs about multilevel modeling at different levels of sophistication are available in the medical and public health literature (Diez-Roux 2000; Goldstein et al. 2002; O'Connell and McCoach 2004; Adewale et al. 2007; Lipsky et al. 2010). Legitimate questions can be raised whether the theoretical benefits of multilevel models, or even their improved predictions, warrant the additional complexity (Hannan et al. 2005; Cohen et al. 2009; Mukamel et al. 2010). However, it is useful for the analyst to be familiar with them, at least in order to consider when they may or may not be appropriate.

Software for Estimating Multilevel Models

Estimating the coefficients for multilevel models, and in particular for nonlinear multilevel models, involves maximum-likelihood theory and sophisticated iterative matrix algorithms, the details of which are far beyond the scope of this chapter. Until recently, most software algorithms used

approximate “quasi-likelihood” or “pseudo-likelihood” methods for nonlinear multilevel models, which were reasonably accurate but occasionally failed to converge to a solution or tended to be biased in some situations (Snijders and Bosker 1999). Simulation methods can be used to derive parameter estimates to any desired accuracy by increasing the number of iterations but are slow and do not explicitly estimate the likelihood. Adaptive quadrature methods (of which the simplest is the Laplace approximation) are also relatively slow but can be made more accurate by increasing the number of quadrature points and do estimate the likelihood, so likelihood-ratio tests can be applied (Rabe-Hesketh and Skrondal 2008).

Fortunately, rapid progress in computing speed and clever programming has now made nonlinear multilevel methods more accessible. This rapid progress also means that the descriptions in this chapter may be outdated by the time they are published. All of the software packages mentioned below can estimate models with random intercepts and binary responses, such as are described in this chapter. Each can also estimate other kinds of linear and nonlinear multilevel models. They use different computational algorithms and generally offer one or more relatively fast approximations to be used during model development, along with more precise methods that require more time to converge to a solution. Results from different algorithms may differ to some degree, although these differences have diminished as subsequent versions of the software have been improved in recent years. At the present stage of development, it may still be reassuring if similar results are obtained using more than one method.

HLM (Scientific Software International, Lincolnwood IL) is a package designed specifically for the implementation of hierarchical linear models (hence the name). It arose out of educational research (e.g., students grouped by schools), and its user community is predominantly American educators. In order to import data into HLM, they must be formatted into separate files for each level; once this step has been accomplished, models can be estimated following simple instructions to implement partial quasi-likelihood (PQL) or adaptive quadrature methods. The academic cost for the most recent version is advertised at \$495, and regular courses are offered in Chicago to learn the mechanics and interpretation. Detailed theoretical background is provided in a textbook by the principal developers of HLM (Raudenbush and Bryk 2002) and numerous other articles by these authors and their colleagues.

MLwiN (Center for Multilevel Modelling, Bristol UK) is another package designed specifically for the implementation of multilevel models, using the Windows operating system (hence the name). It also arose out of educational research, and its user community is predominantly British educators. In order to import data into MLwiN, larger data files may need to be split into several columns at a time; once this step has been accomplished, models can be estimated following simple instructions to implement PQL or Markov Chain Monte Carlo (MCMC) simulation. The academic cost for the most recent version is advertised at \$564, and regular courses are offered in Bristol to learn the mechanics and interpretation. Detailed theoretical background is provided in the text by the principal developer of MLwiN (Goldstein 2003), numerous articles by this author and his colleagues, and an extensive website devoted to the general topic (Centre for Multilevel Modelling 2011).

Stata (StataCorp, College Station TX) is a well-known general purpose statistical package, and obviously, regular users have an incentive to work within a familiar data management environment. Multilevel modeling commands are now standard in Stata, implementing adaptive quadrature methods including the Laplace approximation. In addition to documentation and references available in its software manuals, Stata has published a textbook (Rabe-Hesketh and Skrondal 2008) and has recently begun to offer courses on multilevel modeling in various US cities.

SAS (SAS Institute, Cary NC) is another well-known general purpose statistical package, likewise offering its regular users the incentive to work within a familiar data management environment. The most frequently used multilevel modeling commands are PROC MIXED for continuous response variables and PROC GLIMMIX (pseudo-likelihood) for binary or ordinal response variables. PROC

NLMIXED (adaptive quadrature) can be used for any type of response variable but is not as user friendly. SAS has published guides to all of these procedures, and further information is available in papers from user group meetings.

Other statistical computing tools can also be used to implement multilevel models, including the popular open-source language R (Gelman and Hill 2007). Statisticians approaching hierarchical models from a strictly Bayesian perspective (Normand et al. 1997) generally use simulation methods for computation, especially the open-source “BUGS” language; (The BUGS Project 2011) interfaces are available between MLwiN and BUGS and between SAS and BUGS.

Approximating and Applying the Shrinkage Coefficient

In linear multilevel models, v_{1j} is a simple function of n_j , the number of patients in hospital j (or more generally individuals in any group j) who are assumed to share with all other groups the same patient-level variance, say σ^2 . If the outcome of interest was the total of some continuous-valued measurement for all patients in hospital j (analogous to the total number of deaths in our logistic regression example), then $v_{1j} = n_j \sigma^2$. More commonly, the outcome of interest is the group mean for some measurement, in which case $v_{1j} = \sigma^2/n_j$, and

$$\lambda_j = \frac{v_2}{v_2 + \frac{\sigma^2}{n_j}}. \quad (23.11)$$

For a single individual ($n_j = 1$) in either case, (23.11) reduces to an intraclass correlation coefficient (ICC), which may be interpreted as measuring the proportion of the variance for an average subject that can be attributed to membership in a group.

On the other hand, with multilevel logistic regression models for binary outcomes like mortality, there is no common variance σ^2 , and furthermore, v_2 is not measured on the same scale as v_{1j} , so there can be no exact analogue to the ICC. Some methods to approximate an ICC in this situation have been proposed (Goldstein 2003), and these methods can be extended to calculate approximations for λ_j .

In a random intercept multilevel model of the type specified by (23.6–23.7), all the coefficients are fixed except for $b_{0j} = b_0 + u_{0j}$, where u_0 is a normal random variable having mean 0 and variance v_{2L} . This chapter will henceforth use the subscript 2L as a reminder that this level 2 variance, as reported by standard computer programs, is measured on the logit (log-odds) scale and is therefore not directly comparable to v_{1j} for the purpose of partitioning the total variance. The level 1 variance for hospital j (from 23.3) will be henceforth symbolized as v_{1pj} as a reminder that it is measured on the probability scale.

In order to make the variances comparable, it is possible to use a first-order Taylor series approximation, sometimes referred to as the “delta method” (Cook and DeMets 2008). If the variance for a random variable w is known, then the variance for $f(w)$, some function of w , may be approximated by

$$\text{var}(f(w)) \approx \left(\frac{d(f(w))}{dw} \right)^2 \text{var}(w), \quad (23.12)$$

where the derivative of $f(w)$ is evaluated at its mean.

The estimate of $y_{ij} = \text{logit}(p_{ij})$ may be considered as the random variable w with variance v_{2L} . The function $p_{ij} = \text{logit}^{-1}(y_{ij})$ may be considered as the function $f(w)$. The delta method can then be used

to estimate v_{2P_j} , the variance on the probability scale for the estimated sum of p_{ij} for all patients in hospital j . Referring to (23.2),

$$\frac{d\left(\frac{e^w}{1+e^w}\right)}{dw} = \frac{e^w}{(1+e^w)^2},$$

and this expression evaluated at the mean of $f(w)$ will be the sum of this expression over all values of b_{ij} . Accordingly, the derivative of $f(w)$, evaluated at its mean, becomes

$$\sum_j \frac{e^{b_{ij}}}{(1+e^{b_{ij}})^2} = \sum_j \frac{e^{b_{ij}}}{(1+e^{b_{ij}})(1+e^{b_{ij}})} = \sum_j p_{ij}(1-p_{ij}) = v_{1P_j}.$$

Substituting this convenient result into (23.12), an approximate level 2 variance on the probability scale is then simply

$$v_{2P_j} \approx (v_{1P_j})^2 v_{2L}, \tag{23.13}$$

noting that this will be different for each hospital. An approximate shrinkage coefficient for hospital j can then be calculated as

$$\lambda_j \approx \frac{v_{2P_j}}{v_{2P_j} + v_{1P_j}} \approx \frac{(v_{1P_j})^2 v_{2L}}{(v_{1P_j})^2 v_{2L} + v_{1P_j}} = \frac{v_{1P_j} v_{2L}}{v_{1P_j} v_{2L} + 1}.$$

The delta method can also be used in the reverse direction to transform v_{1P_j} to the logit scale, differentiating

$$\frac{d\left(\sum_j \log\left(\frac{p_{ij}}{1-p_{ij}}\right)\right)}{dp_{ij}} = \frac{1}{\sum_j p_{ij}(1-p_{ij})},$$

and evaluating at the mean of p_{ij} to produce the similarly convenient result that v_{1L_j} on the logit scale can be approximated by $1/v_{1P_j}$. Then,

$$\lambda_j \approx \frac{v_{2L}}{v_{2L} + v_{1L_j}} \approx \frac{v_{2L}}{v_{2L} + 1/v_{1P_j}} = \frac{v_{1P_j} v_{2L}}{v_{1P_j} v_{2L} + 1}, \tag{23.14}$$

so that the same formula for λ_j is obtained on either the probability scale or the logit scale.

An iterative Taylor series procedure has been presented that allows more accurate approximations of λ_j , as well as b_{0j} , on the logit scale (Clark et al. 2010b). If a multilevel model (including v_{2L}) has been estimated on the logit scale using a software package, simulation can also be used to obtain a value for λ_j on the probability scale. From (23.6–23.7) and the following algorithm, compute

- For $j=1$ to J (each hospital)
 - For $z=1$ to Z (some large number of iterations)
 - Generate u_{0jz} from a Normal $(0, v_{2L})$ distribution
 - Let $b_{0jz} = b_{00} + u_{0jz}$
 - For $i=1$ to n_j (each patient in hospital j)
 - Let $b_{ijz} = b_{0jz} + b_1 x_{ij1} + \dots + b_k x_{ijk}$
 - Let $p_{ijz} = 1/(1 + \exp(-b_{ijz}))$

$$\begin{aligned}
&\text{Let } p_{jz} = \text{Sum}(p_{ijz}) \\
&\text{Let } v_{jz} = \text{Sum}(p_{ijz}(1-p_{ijz})) \\
&\text{Let } v_{2pj} = \text{Variance}(p_{jz}) \\
&\text{Let } v_{1pj} = \text{Mean}(v_{jz}) \\
&\text{Calculate } \lambda_j = \frac{v_{2pj}}{v_{2pj} + v_{1pj}}
\end{aligned}$$

The worked example below will show that a similar estimate for λ_j is obtained regardless of which method is used and whether it is estimated on the logit scale or the probability scale.

All the methods described above for obtaining λ_j assume that a multilevel model has been estimated and that data are available from hospital j but do not necessarily assume that hospital j was a part of the reference database used to estimate the model. If hospital j was indeed in the reference database, and a computer program has given the total variance (or its square root, the standard error) for b_{0j} , then some algorithm can be presumed to have utilized an approximation for v_{1Lj} . An expression for the variance of b_{0j} in terms of v_{1Lj} , v_{2j} , and λ_j can be derived from the previously described concept that the variance of an estimate is the inverse of the precision of that estimate, along with the concept that the total precision is the sum of the level 1 precision and the level 2 precision. This leads to

$$\text{var}(b_{0j}) = \frac{1}{\frac{1}{v_{1Lj}} + \frac{1}{v_{2L}}} = (1 - \lambda_j)v_{2L}. \quad (23.15)$$

The formula described by (23.15) has been published (not necessarily with this derivation) in standard textbooks (Snijders and Bosker 1999; Raudenbush and Bryk 2002) and can be rearranged to give

$$\lambda_j = 1 - \frac{\text{var}(b_{0j})}{v_{2L}}, \quad (23.16)$$

so that λ_j may be calculated using $\text{var}(b_{0j})$ and v_{2L} obtained from the computer output. It is reasonable to assume, and demonstrable in practice, that if u_{0j} is set equal to 0, the estimates of y_{ij} for each patient i in hospital j to be used in (23.1) or (23.6) will be almost identical whether single-level or multilevel logistic regression is used. Therefore, an approximate variance for the shrunken estimate of the predicted mortality for hospital j can be obtained as

$$\text{var}(S_j) \approx \frac{1}{\frac{1}{v_{1pj}} + \frac{1}{v_{2pj}}} \approx \lambda_j v_{1pj}, \quad (23.17)$$

assuming that v_{1pj} is essentially the same as v_j calculated in (23.3).

The relationship between the confidence intervals for SMR and sSMR can then be compared using the approximate variance calculated in (23.17). Since the estimates of v_{1pj} will be virtually the same,

$$\begin{aligned}
\text{SE}(\text{unshrunk}) &\approx \sqrt{v_{1pj}}, \\
\text{SE}(\text{shrunk}) &\approx \sqrt{\lambda_j v_{1pj}} = \sqrt{\lambda_j} \sqrt{v_{1pj}}.
\end{aligned} \quad (23.18)$$

When compared to the unshrunk SMR, sRSMR will thus be shrunken toward 1 by a factor approximately equal to the reliability coefficient (23.9), whereas its LCL and UCL will be shrunken by a

factor approximately equal to the square root of the reliability coefficient (23.18). Since $0 < \lambda_j < 1$, it is true that $\lambda_j < \sqrt{\lambda_j} < 1$, so that the confidence limits will be shrunken less than the point estimates.

To illustrate the calculations of the preceding paragraph, suppose for some hospital j that the SMR (O_j/E_j) is 5, the UCL for the SMR is 4, and the shrinkage coefficient λ_j is 0.25. From (23.9), the sSMR will be shrunken approximately to $1 + 0.25(5 - 1) = 2$. Using (23.18), the UCL for the sSMR will be shrunken approximately to $1 + 0.5(4 - 1) = 2.5$. In this case, hospital j was an outlier with single-level logistic regression but is no longer an outlier with multilevel logistic regression; in general, (23.18) means that number of outliers must be fewer with the multilevel method.

Worked Example

National Trauma Data Bank data for admission year 2008 were obtained from the American College of Surgeons (ACS) Committee on Trauma, in compliance with its standard Data Use Agreement. Initial data management and modeling were performed using Stata (Version 11, College Station, TX), and programs to replicate these findings are available from the first author on request.

Data files in text format were converted for use with Stata and merged into a single analytic file. An outcome variable “died” was created if the patient had either an emergency department outcome of “dead” or a hospital outcome of “expired.” Patient characteristics were categorized as required for the Trauma Quality Improvement Program (Hemmila et al. 2010). Hospitals were characterized by their ACS verification status as trauma centers.

Analysis was based on cases with an Injury Severity Score (ISS) of at least nine, admitted to an ACS-verified level I or level II trauma center (note that the “level I” and “level II” terminology here has a different meaning from the “level 1” and “level 2” terminology for the statistical models). Hospitals were excluded if they had fewer than 200 cases with an ISS of at least 9, if more than 1% of their cases were missing vital status (dead/alive) at discharge, or if more than 20% of their remaining cases did not have valid data recorded at the time of hospital admission for vital status or any of the required predictive factors, namely, age, mechanism of injury, abbreviated injury score for the head or abdominal regions (AIS_h, AIS_a), Glasgow Coma Scale motor score (GCS_m), systolic blood pressure (BP), pulse rate (PR), or status as a transfer from another hospital (Transin).

Data were used from the 125 remaining hospitals, but individual cases were excluded if they lacked data for any of the variables described in the preceding paragraph, or if they were declared “dead on admission,” or if they were transferred out to another acute care hospital.

In order to provide an example that could be replicated relatively easily, and to focus on the effect of using a multilevel model, no further exclusions, imputations, or adjustments were undertaken. The many potential misinterpretations of hospital mortality as an outcome and many other possible methods of risk adjustment (Clark et al. 2007; Gorra et al. 2008; Moore et al. 2009a, b) will not be addressed in this chapter.

A standard logistic regression model predicting hospital mortality was first created using the “logit” command in Stata. Hospital effects were estimated using (23.1) above. A corresponding multilevel logistic regression model was created using the “xtmelogit” command in Stata. The number of quadrature points was started at 1 (Laplace approximation) and increased in increments of 2 until the variance of the random intercept did not change at three significant digits. This resulted in a random intercept logistic regression model in the form of (23.6–23.7) and an estimate of the level 2 variance (v_{21}). A posterior prediction of the effect for each hospital (b_{0j}) and its standard error were generated using the “predict, reffects” and “predict, reses” commands in Stata.

Patient-level effects obtained using single-level or multilevel modeling were very similar (Table 23.1). As would be expected, mortality was associated with older age, firearm injury, greater

Table 23.1 Fixed effects estimated from standard LR (logit command) and multilevel random intercept LR (xtmelogit command) using Stata

Variable	Standard LR		Multilevel LR	
	Effect	95% CI	Effect	95% CI
Age >65 years	1.69	1.61, 1.76	1.70	1.63, 1.78
Firearm mechanism of injury	1.24	1.14, 1.34	1.23	1.12, 1.33
Injury Severity Score >24	1.09	1.02, 1.17	1.12	1.04, 1.19
AISh 1–2=1, AISh 3–4=2, AISh 5–6=3	0.43	0.39, 0.47	0.44	0.40, 0.47
AISa 1–2=1, AISa 3–4=2, AISa 5–6=3	0.14	0.09, 0.18	0.14	0.09, 0.19
Initial GCS motor score 1	2.55	2.47, 2.62	2.58	2.50, 2.66
Initial GCS motor score 2–5	1.59	1.50, 1.68	1.57	1.48, 1.66
Initial systolic BP 0	3.10	2.81, 3.39	3.11	2.82, 3.40
Initial systolic BP 1–90	1.15	1.04, 1.25	1.15	1.05, 1.25
Initial pulse rate 0–40	1.50	1.26, 1.75	1.52	1.26, 1.77
Transfer from another hospital	–0.24	–0.31, –0.18	–0.25	–0.32, –0.18

LR logistic regression, CI confidence interval, AISh maximum Abbreviated Injury Scale in the head region, AISa maximum Abbreviated Injury Scale in the abdominal region, GCS Glasgow Coma Scale, BP blood pressure. “Initial” physiologic data were those recorded in the emergency department. Odds ratios can be obtained by exponentiating the effect obtained on the logit scale

Table 23.2 Correlation coefficients between different estimates of the shrinkage coefficient obtained using computer output, first-order Taylor series (delta method), iterative Taylor series, or simulation

Estimation method	A	B	C	D
A. Computer output	1			
B. Delta method	0.980	1		
C. Iterative Taylor series	>0.999	0.980	1	
D. Simulation	0.997	0.979	0.997	1

See text for methods of calculation

injury severity (especially in the head and abdominal regions), depressed mental status (low GCSm), low blood pressure, and low pulse rate. Patients received in transfer from another hospital had lower mortality.

An estimate of the shrinkage coefficient for each hospital (λ_j) was obtained from the computer output and (23.16). In order to validate the approximation methods described above, estimates of λ_j were also obtained using the delta method (23.12–23.14 described above), the iterative Taylor series method described in another publication (Clark et al. 2010b), and the simulation method described above with 1,000 replications for each hospital. Correlation coefficients comparing these estimates were obtained using the “corr” command in Stata and are shown in Table 23.2. The estimates are all very similar, although the delta method was not correlated quite as closely as the others.

A comparison of the hospital effects obtained from standard logistic regression and the “shrunk” effects obtained from the random intercept model is shown in Fig. 23.1. Figure 23.2 shows a funnel plot of SMR obtained using standard logistic regression, while Fig. 23.3 shows a funnel plot of sSMR obtained using multilevel logistic regression for comparison. Figure 23.4 shows a funnel plot of the unshrunk SMR, but with confidence limits expanded using (23.10).

Multilevel models were replicated using identical NTDB data imported into SAS (Version 9.2), HLM (Version 6.04), and MLwiN (Version 2.02). Estimates of b_0 and v_{2L} obtained with different algorithms, as well as the approximate time to reach a result, are compared in Table 23.3. Computing times may of course be affected by the particular machine used. The estimates of the fixed effects

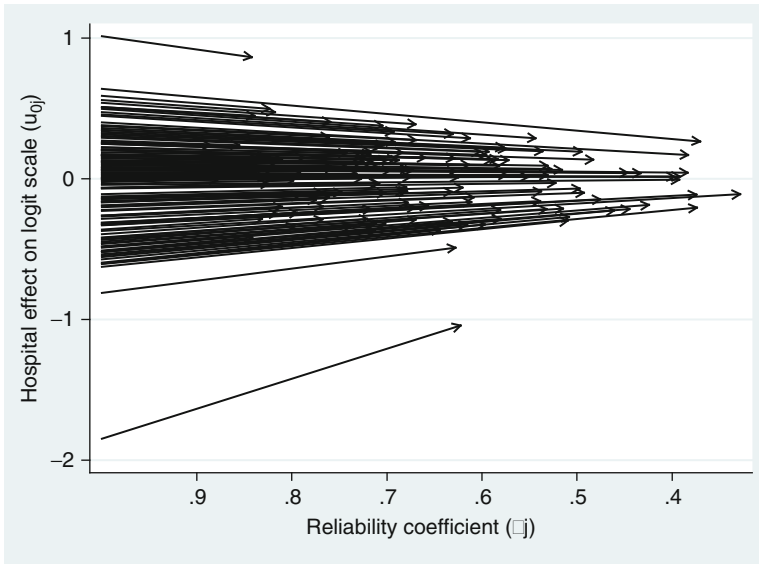


Fig. 23.1 Effect of the reliability coefficient for the level 1 data from hospital j (λ_j) on the empirical Bayes prediction of the hospital effect on mortality. The standard logistic regression estimate is on the left y-axis (corresponding to $\lambda_j=1$), and the arrowhead is at the “shrunk” effect estimated by a random intercept multilevel model

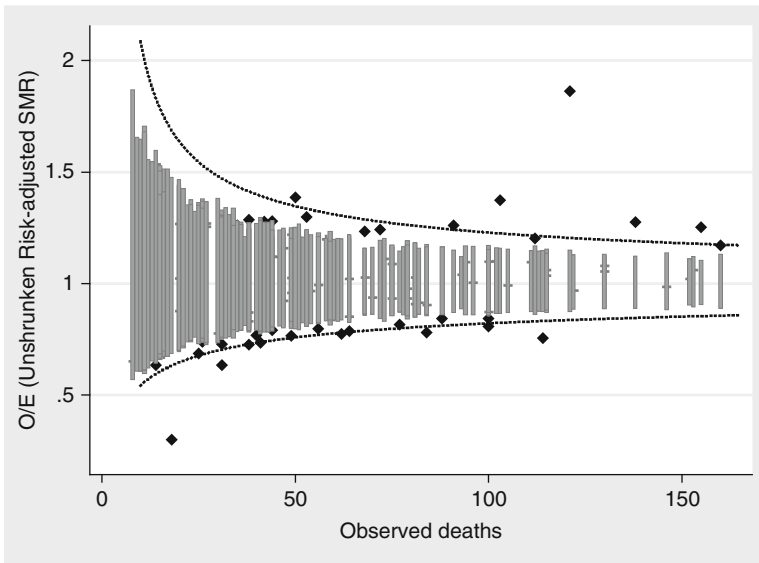


Fig. 23.2 Standardized mortality ratio (SMR) with confidence intervals calculated from single-level logistic regression. Data refer to levels I–II trauma centers in the National Trauma Data Bank, using a risk-adjustment model based on that proposed for the Trauma Quality Improvement Program. Point estimates within the confidence intervals are shown as crosses, while outliers are shown as diamonds. The dotted lines indicate 95% confidence/control limits based upon a Poisson approximation

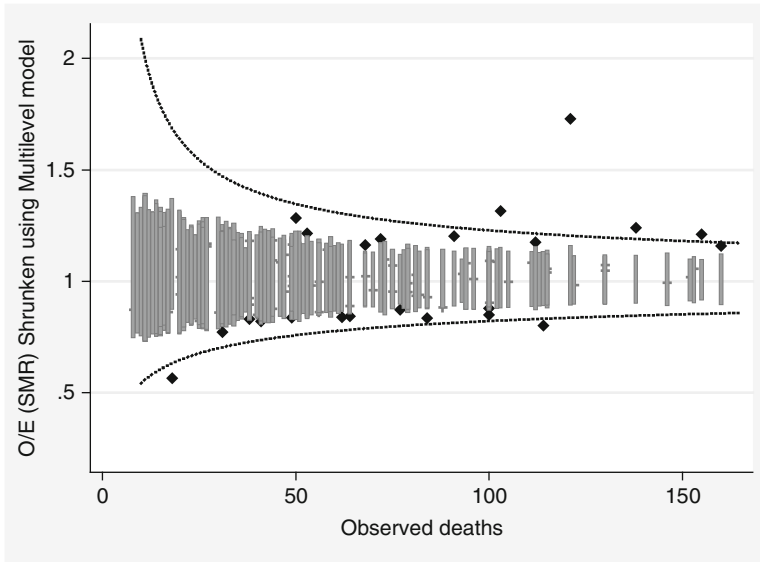


Fig. 23.3 Shrunken standardized mortality ratio (sSMR) with confidence intervals calculated from multilevel logistic regression. Data refer to levels I–II trauma centers in the National Trauma Data Bank, using a risk-adjustment model based on that proposed for the Trauma Quality Improvement Program. Point estimates within the confidence intervals are shown as crosses, while outliers are shown as diamonds. The dotted lines indicate 95% confidence/control limits for the unshrunk SMR based upon a Poisson approximation. Compared to Fig. 23.2, the point estimates have been shrunken toward the mean by a factor of approximately λ_j , while the confidence limits have been shrunken toward the mean by a factor of approximately $\sqrt{\lambda_j}$. Consequently, there are fewer outliers

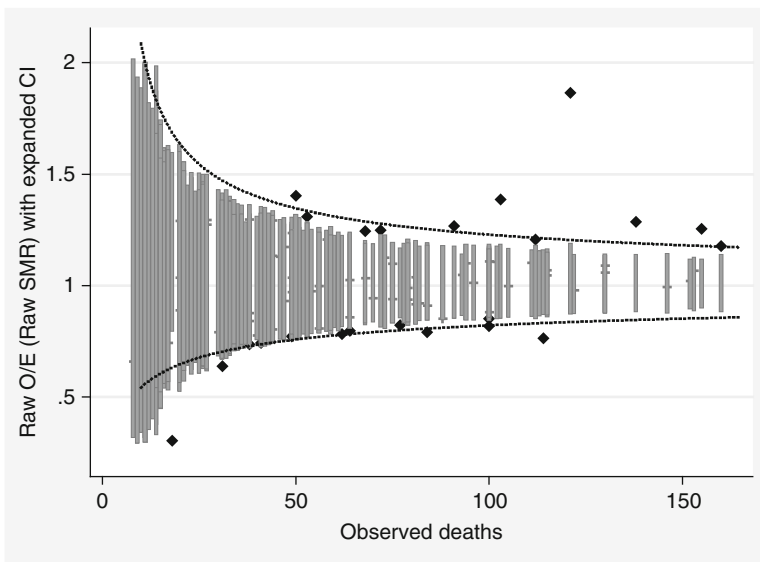


Fig. 23.4 Unshrunk standardized mortality ratio (SMR) with confidence intervals calculated from multilevel logistic regression. Data refer to levels I–II trauma centers in the National Trauma Data Bank, using a risk-adjustment model based on that proposed for the Trauma Quality Improvement Program. Point estimates within the confidence intervals are shown as crosses, while outliers are shown as diamonds. The dotted lines indicate 95% confidence/control limits for the unshrunk SMR based upon a Poisson approximation. Compared to Fig. 23.2, the point estimates are identical, but the confidence intervals have been expanded by a factor of approximately $1/\sqrt{\lambda_j}$, so there are fewer outliers. Compared to Fig. 23.3, point estimates and confidence intervals have both been expanded by the same factor of approximately $1/\lambda_j$, so there are the same outliers

Table 23.3 Results obtained for the random portions of the described multilevel model using different software and estimation methods, including times until convergence

Results by algorithm	Estimates		Time
	b_{00}	v_{2L}	
Stata, Version 11			
“xtmelogit” AQ (1 integration point – Laplace)	–5.20	0.082	5 min
AQ (7 integration points)	–5.20	0.082	4 min
SAS, Version 9			
“GLIMMIX” RPL	–5.18	0.082	<1 min
“NLMIXED” AQ	–5.19	0.085	6 min
MLwiN, Version 2			
PQL (2nd order)	–5.20	0.082	<2 min
MCMC (5,000 iterations)	–5.20	0.084	28 min
HLM, Version 6			
PQL (1st order)	–5.18	0.082	<1 min
Laplace	–5.18	0.081	<2 min

AQ adaptive quadrature, *RPL* residual pseudo-likelihood, *PQL* penalized quasi-likelihood, *MCMC* Markov Chain Monte Carlo simulation

(not shown) and the random effects were very similar regardless of the software. Computing time was very fast for pseudo-likelihood or quasi-likelihood methods, not quite as fast for adaptive quadrature methods, and slowest for the simulation method.

Other Extensions of Multilevel Modeling

Extensions of the random-intercept model briefly described in this section will include models incorporating level 2 variables, models with random slopes as well as random intercepts, models describing more than two levels of hierarchy, and generalizations of the linear model other than the logistic transformation.

One important advantage of multilevel models is that they can incorporate level 2 as well as level 1 covariates while still evaluating the effect of individual level 2 groups. For example, the registry-based trauma center comparison in the previous section could include a hospital-level variable indicating trauma center status (e.g., $t=1$ if verified as a level I trauma center, $t=0$ otherwise). It would not be possible to include such a covariate when combining (23.1) and (23.4) due to collinearity between the hospital and its verification status. However, in the multilevel model, (23.6) and (23.7) can be modified as

$$y_{ij} = b_{0j} + b_1 x_{ij1} + \dots + b_k x_{ijk},$$

$$b_{0j} = b_{00} + b_{10} t_j + u_{0j}.$$

One caution after applying hospital-level variables in a multilevel model is that estimates will then be shrunken toward the mean of any subgroup sharing the same hospital-level predictors, which may or may not be the intention of the analysis. In this case, hospitals in different subgroups can no longer be directly compared, but the overall effect of different types of hospitals might be evaluated. In the worked NTDB example, the addition of an indicator variable for ACS level I verification status did not have a significant effect, and it was therefore eliminated from the model.

The discussion thus far has also been limited to random-intercept multilevel models. However, if the effect of a predictor (e.g., injury severity) is thought to vary among groups (e.g., hospitals), this

phenomenon can be modeled by adding random slopes. For example, a hospital might perform well with moderately injured patients but not as well with those more severely injured. If X_1 were some measure of injury severity, then the model described in (23.6) and (23.7) could be extended by adding a random coefficient, that is

$$y_{ij} = b_{0j} + b_{1j}x_{ij1} + b_2x_{ij2} \dots + b_kx_{ijk}.$$

In this case, there is not only a random intercept as before but also a random slope in the second level of the model

$$b_{0j} = b_{00} + u_{0j},$$

$$b_{1j} = b_{10} + u_{1j}.$$

This might be valuable from the standpoint of helping a hospital focus its efforts to improve quality of care but would make it more difficult to evaluate the overall performance of this hospital in comparison with others.

The discussion so far in this chapter has only considered two levels of hierarchy, with the example of patients grouped within hospitals. However, if hospitals were thought to have characteristics more in common with other hospitals in their geographic region than in other regions, this correlation could be modeled by adding another level of hierarchy. The model described in (23.6) and (23.7) would then become

$$y_{ijr} = b_{0jr} + b_{1jr}x_{ijr1} + b_2x_{ijr2} \dots + b_kx_{ijrk},$$

$$b_{0jr} = b_{00r} + u_{0jr},$$

$$b_{00r} = b_{000} + u_{00r},$$

where r specifies the region containing hospital j . Such a model might be useful for exploring why certain interhospital differences were found but then would not allow direct comparison of hospitals from different regions. Higher-level models of this sort can also be used to incorporate a time dimension, which is useful for “repeated measures” studies, in which the same individual provides data on several occasions.

This chapter has concentrated on multilevel logistic models as an extension of multilevel linear models. However, the linear model can also be extended to involve Poisson or negative binomial regression, generally used to model counts or incidence rates for relatively infrequent events. Examples where such models have been applied to injury data include a study on the incidence of overall injury mortality in populations grouped by neighborhoods of Barcelona and a study on the incidence of hospitalization after traffic crashes in young men grouped by health service areas of British Columbia, both based on multilevel Poisson regression (Borrell et al. 2002; MacNab 2003). A Greek study of the incidence of traffic crashes used negative binomial multilevel regression, grouping events by geographical regions (Yannis et al. 2007). An Austrian study counting repeated suicide attempts used a Poisson multilevel model with the individual person as a grouping variable, thus demonstrating the use of multilevel modeling for repeated measures over time (Antretter et al. 2006).

Summary and Conclusions

Given the frequency with which grouped data occur in observational studies of injury, and in other areas of epidemiology and health care research, models that account for correlated variance structures can be an important part of the analyst’s toolkit. Extending multilevel models beyond the linear

setting involves some additional assumptions and computational issues. However, for random-intercept multilevel logistic models, the theoretical complexity can be reduced by considering approximate formulas related to the reliability coefficient (shrinkage coefficient). Computation is now fairly rapid using any one of several software packages.

Empirical Bayes estimation can theoretically and practically be shown to give better predictions, especially for groups with scant data. In addition to allowing for empirical Bayes estimation, the multilevel approach (for logistic or linear models) offers other potential advantages, including (1) appropriately modeling the clustered nature of the data and correlation of outcomes within hospitals; (2) naturally adjusting for inflation of the type-I error rate caused by multiple comparisons; (3) allowing for estimation in cases where the observed hospital mortality is zero; (4) allowing for incorporation of hospital characteristics as well as patient characteristics in the model; and (5) reducing the occurrence of outlying mortality estimates, especially when these are felt to be an artifact of small sample sizes.

Acknowledgments Supported in part by Grant R21HD061318 (PI Clark) from the National Institutes of Health, Grants R01HS015656 (PI Clark) and R01HS017718 (PI Shafi) from the Agency for Healthcare Research and Quality, a grant from the Maine Medical Center Research Strategic Plan (PI Clark), and a research award from the Canadian Institutes of Health Research (PI Moore).

References

- Adewale, A. J., Hayduk, L., Estabrooks, C. A., et al. (2007). Understanding hierarchical linear models: Applications in nursing research. *Nursing Research, 56*, S40–6.
- Antretter, E., Dunkel, D., Osvath, P., et al. (2006). Multilevel modeling was a convenient alternative to common regression designs in longitudinal suicide research. *Journal of Clinical Epidemiology, 59*, 576–86.
- Austin, P. C., Tu, J. V., & Alter, D. A. (2003). Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: Should we be analyzing cardiovascular outcomes data differently? *American Heart Journal, 145*, 27–35.
- Borrell, C., Rodriguez, M., Ferrando, J., et al. (2002). Role of individual and contextual effects in injury mortality: New evidence from small area analysis. *Injury Prevention, 8*, 297–302.
- Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*, 473–514.
- Centre for multilevel modelling. www.bristol.ac.uk/cmm. Accessed 1 Mar 2011.
- Champion, H. R., Copes, W. S., Sacco, W. J., et al. (1990). The Major Trauma Outcome Study: Establishing national norms for trauma care. *Journal of Trauma, 30*, 1356–65.
- Chan, C. K., Feinstein, A. R., Jekel, J. F., & Wells, C. K. (1988). The value and hazards of standardization in clinical epidemiologic research. *Journal of Clinical Epidemiology, 41*, 1125–34.
- Clark, D. E., Hannan, E. L., & Wu, C. (2010a). Predicting risk-adjusted mortality for trauma patients: Logistic versus multilevel logistic models. *Journal of the American College of Surgeons, 211*, 224–31.
- Clark, D. E., Hannan, E. L., & Raudenbush, S. W. (2010b). Using a hierarchical model to estimate risk-adjusted mortality for hospitals not included in the reference sample. *Health Services Research, 45*, 577–87.
- Clark, D. E., Lucas, F. L., & Ryan, L. M. (2007). Predicting hospital mortality, length of stay, and transfer to long-term care for injured patients. *Journal of Trauma, 62*, 592–600.
- Cohen, M. E., Dimick, J. B., Bilimoria, K. Y., et al. (2009). Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: A comparison of logistic versus hierarchical modeling. *Journal of the American College of Surgeons, 209*, 687–93.
- Cook, T. D., & DeMets, D. L. (2008). *Introduction to statistical methods for clinical trials*. Boca Raton, FL: Chapman and Hall.
- DeLong, E. R., Peterson, E. D., DeLong, D. M., et al. (1997). Comparing risk-adjustment methods for provider profiling. *Statistics in Medicine, 16*, 2645–64.
- Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health, 21*, 171–92.
- Dimick, J. B., & Welch, H. G. (2008). The zero mortality paradox in surgery. *Journal of the American College of Surgeons, 206*, 13–6.

- Flora, J. D., Jr. (1978). A method for comparing survival of burn patients to a standard survival curve. *Journal of Trauma*, 18, 701–5.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Glance, L. G., Osler, T. M., Dick, A. W., et al. (2010). The survival measurement and reporting trial for trauma (SMARTT): Background and study design. *Journal of Trauma*, 68, 1491–7.
- Goldstein, H. (2003). *Multilevel statistical models*. London: Hodder Arnold.
- Goldstein, H., Browne, W., & Rasbash, J. (2002). Multilevel modelling of medical data. *Statistics in Medicine*, 21, 3291–315.
- Gorra, A. S., Clark, D. E., Mullins, R. J., & Delorenzo, M. A. (2008). Regional variation in hospital mortality and 30-day mortality for injured Medicare patients. *World Journal of Surgery*, 32, 954–9.
- Hannan, E. L., Wu, C., DeLong, E. R., & Raudenbush, S. W. (2005). Predicting risk-adjusted mortality for CABG surgery: Logistic versus hierarchical logistic models. *Med Care*, 43, 726–35.
- Hemmila, M. R., Nathens, A. B., Shafi, S., et al. (2010). The Trauma Quality Improvement Program: Pilot study and initial demonstration of feasibility. *Journal of Trauma*, 68, 253–62.
- Jones, A. P., & Jorgensen, S. H. (2003). The use of multilevel models for the prediction of road accident outcomes. *Accident Analysis and Prevention*, 35, 59–69.
- Jones, J. M., Redmond, A. D., & Templeton, J. (1995). Uses and abuses of statistical models for evaluating trauma care. *Journal of Trauma*, 38, 89–93.
- Kim, D. G., Lee, Y., Washington, S., & Choi, K. (2007). Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. *Accident Analysis and Prevention*, 39, 125–34.
- Kirkham, J. J., & Bouamra, O. (2008). The use of statistical process control for monitoring institutional performance in trauma care. *Journal of Trauma*, 65, 1494–501.
- Krumholz, H. M., Brindis, R. G., Brush, J. E., et al. (2006). Standards for statistical models used for public reporting of health outcomes. *Circulation*, 113, 456–62.
- Lenguerrand, E., Martin, J. L., & Laumon, B. (2006). Modelling the hierarchical structure of road crash data – application to severity analysis. *Accident Analysis and Prevention*, 38, 43–53.
- Lipsky, A. M., Gausche-Hill, M., Vienna, M., & Lewis, R. J. (2010). The importance of “shrinkage” in subgroup analyses. *Annals of Emergency Medicine*, 55(544–52), e3.
- MacNab, Y. C. (2003). A Bayesian hierarchical model for accident and injury surveillance. *Accident Analysis and Prevention*, 35, 91–102.
- Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention*, 40, 260–6.
- Moore, L., Hanley, J. A., Turgeon, A. F., et al. (2009a). A multiple imputation model for imputing missing physiologic data in the National Trauma Data Bank. *Journal of the American College of Surgeons*, 209, 572–9.
- Moore, L., Lavoie, A., Turgeon, A. F., et al. (2009b). The trauma risk adjustment model: A new model for evaluating trauma care. *Annals of Surgery*, 249, 1040–6.
- Moore, L., Hanley, J. A., Turgeon, A. F., & Lavoie, A. (2010a). Evaluating the performance of trauma centers: Hierarchical modeling should be used. *Journal of Trauma*, 69, 1132–7.
- Moore, L., Hanley, J. A., Turgeon, A. F., et al. (2010b). A new method for evaluating trauma centre outcome performance: TRAM-adjusted mortality estimates. *Annals of Surgery*, 251, 952–8.
- Mukamel, D. B., Glance, L. G., Dick, A. W., & Osler, T. M. (2010). Measuring quality for public reporting of health provider quality: Making it meaningful to patients. *American Journal of Public Health*, 100, 264–9.
- Normand, S.-L. T., Glickman, M. E., & Gatsonis, C. A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *Journal of the American Statistical Association*, 92, 803–14.
- Normand, S.-L. T., & Shahian, D. M. (2007). Statistical and clinical aspects of hospital outcomes profiling. *Statistical Science*, 22, 206–26.
- O’Connell, A. A., & McCoach, D. B. (2004). Applications of hierarchical linear models for evaluations of health interventions: Demystifying the methods and interpretations of multilevel models. *Evaluation and The Health Professions*, 27, 119–51.
- O’Hagan, A. (1994). *Kendall’s advanced theory of statistics, volume 2B: Bayesian inference*. New York: Halsted Press.
- Pollock, D. A. (1999). Summary of the discussion: Trauma registry data and TRISS evidence. *Journal of Trauma*, 47, S56–S8.
- Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. London: Sage.
- Rothman, K. J. (1986). *Modern epidemiology*. Boston: Little, Brown and Co.
- Shahian, D. M., Normand, S.-L. T., Torchiana, D. F., et al. (2001). Cardiac surgery report cards: Comprehensive review and statistical critique. *The Annals of Thoracic Surgery*, 72, 2155–68.

- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage.
- Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24, 1185–202.
- The BUGS project*. www.mrc-bsu.cam.ac.uk/bugs Accessed 1 Mar 2011.
- Ulm, K. (1990). A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). *American Journal of Epidemiology*, 131, 373–5.
- Yannis, G., Papadimitriou, E., & Antoniou, C. (2007). Multilevel modelling for the regional effect of enforcement on road accidents. *Accident Analysis and Prevention*, 39, 818–25.

Chapter 24

Geographical Information Systems

Becky P.Y. Loo and Shenjun Yao

Introduction

Road traffic crashes rank among the top global burdens of disease in terms of disability-adjusted life year (DALY) (Murray and Lopez 1994). According to the World Health Organization (WHO), approximately 1.3 million people die each year and another 20–50 million sustain nonfatal injuries on roads worldwide (WHO 2009). Road trauma is an important public health problem, especially for low- and middle-income countries, which accounted for over 90% of the world's road fatalities but only about 48% of the world's vehicle fleet (WHO 2009). Unlike contagious diseases such as influenza, injury is not spatially contagious. Yet, their occurrence is clearly influenced by locational factors. Early research on the investigation of geographical distribution of road crashes, for example, simply mapped the location of cases and visually examined crash maps *directly* to identify spatial clusters and potentially hazardous road locations. More advanced spatial analyses have been made possible only with the development of geographic information systems (GIS), which are generally defined as computer-based systems for collecting, storing, integrating, analyzing, and displaying spatially referenced data (Gatrell and Loytonen 1998). Together with other information technologies such as remote sensing (RS) and global positioning system (GPS), GIS has transformed the ways in which we study spatial phenomena such as traffic crashes. With GIS as an enabling technology, researchers can move their focus away from simple mapping (that is, producing maps at multiple scales) to more advanced spatial analyses through integrating large quantities of spatial and nonspatial data and examining relationships through hypothesis testing. In this chapter, we first introduce different types of GIS-based spatial analysis methods for crash analysis and prevention. Then, we introduce a GIS-based network analysis approach for the identification of hazardous road locations. Finally, we conclude by suggesting ways forward for better utilizing the spatial data and spatial modeling capacities to reduce traffic injury.

B.P.Y. Loo, PhD (✉) • S. Yao, PhD
Department of Geography, The University of Hong Kong, Pokfulam Road, Hong Kong, China
e-mail: bpyloo@hku.hk; shenjun@hku.hk

GIS Applications in Crash Analysis and Prevention

From a user's perspective, the main functions of GIS fall into three categories: database management, visualization and mapping, and spatial analysis (Cromley and McLafferty 2002). In this section, we select some examples to illustrate how these three important categories of GIS functions can be used to support crash analysis and prevention.

Database Management

GIS allows researchers to integrate and manage large quantities of spatial and nonspatial data that contain relevant information about events causing injury. For road traffic crashes, these data pertain to vehicles, road users, road environment, and other factors such as weather. Through the creation of a relational database, various types of spatial data can be integrated and analyzed; these include coordinates measured by GPS, land use imagery derived from aerial photographs or satellites, and street centerlines or census tract boundaries digitized from atlases, as well as nonspatial data such as police investigation materials, trauma registry records, other medical and nonmedical interventions, and various demographic and socioeconomic statistics.

The key linkage among different types of datasets is the geographical coordinates that tie the data together by their common spatial locations. For instance, a research team may want to explore the relationship between crashes and (alcohol-selling) pubs at the census-tract level in the USA. Since both crash and pub locations usually only record the street name, it needs to be linked with census tract boundaries. Such linkage relies on the "joining" operation in GIS, by which the tabular data could be "joined" to the spatial layer based upon a common field such as the street name or other spatial identifiers. With the database integration, the densities of crashes and pubs of different census tracts can be easily calculated and additional information, such as the severity of crashes happening near pubs (say, within a half-mile radius), can be obtained for further analysis.

Most GIS packages provide a series of data collection, conversion, transformation, and generalization procedures for integrating diverse types of spatial datasets that differ in scale, geographical extent, or image resolution. A typical data-integration example is the transformation of two spatial datasets with different coordinate systems. For instance, the location of crashes or falls measured by GPS devices is often recorded using *WGS84* as the reference system. *WGS84* belongs to the type of geographic coordinate system that has its coordinates calculated by latitude and longitude in degrees. However, for a particular administrative unit, most spatial data such as road centerlines, buildings, and district boundaries use projected coordinate system that is measured by *X* and *Y* positions based upon a grid lying on a flat surface. For example, Hong Kong Spatial Data are represented by projected coordinate system, known as *Hong Kong 1980 Grid* (Loo 2006). If one wants to integrate the *WGS84* data into the *Hong Kong 1980 Grid* database, one needs to transform the coordinate system of the former in accordance with the latter. Currently, most GIS platforms such as ArcGIS of Environmental Systems Research Institute (ESRI) include a set of tools for conducting such coordinate transformation.

Then, integrating spatial and nonspatial data in a meaningful way is the first step of any scientific GIS application. There are many good examples. For instance, Odera et al. (2007) has linked trauma records to the geographical coordinates of traffic crash sites to establish an electronic injury surveillance system for improving patient care and monitoring injury incidence and distribution patterns. When researchers aim to examine the influence of various environmental factors on the spatial pattern of crashes, huge amounts of additional data associated with environmental risks need to be collected and integrated. With the use of GIS, they could conveniently integrate and manage the

collected data that cover both natural and socioeconomic environmental information such as road type, land use, district boundaries, population age structure, and weather conditions (Graham and Glaister 2003; Loo and Tsui 2005, 2010; Loo and Yao 2010; Wier et al. 2009).

Visualization and Mapping

The visualization functions of GIS enable users to explore data interactively in a form of “visual thinking” and “visual communication” (MacEachren et al. 1992; Hearnshaw and Unwin 1994). In GIS, injury events can either be “viewed by attributes” using Boolean operators such as “OR,” or be “viewed by location” with simple spatial query filters such as “INTERSECT” (Cromley and McLafferty 2002). Moreover, the development of computer graphics has enabled GIS to cope with three-dimensional (3D) representations, scene generations, and other kinds of displays. Recently, researchers have applied 3D GIS to many studies of crash analysis and prevention (Han et al. 2006; Jha et al. 2001; Khattak and Shamayleh 2005). Using Khattak and Shamayleh’s (2005) work as an example, the study introduced a 3D-GIS visualization method to check for stopping and passing sight distance. The geographical data were collected by flying a plane equipped with a light detection and ranging (LiDAR) system on a section of a two-lane rural highway. The collected data were manipulated by GIS to generate 3D models of highway subsections that were then visually examined for safety assessment.

Maps are products derived from the process of viewing, exploring, and analyzing spatial data. A crucial step for today’s mapping process is the representation of spatial information that needs to be displayed. It requires the intelligent use of mapping strategies as well as symbols that are usually differentiated by the six dimensions of visual variability: size, shape, orientation, texture, color hue, and color value (Bertin 1979). Varying these visual variables can highlight places of interest, show contrast, and identify patterns. When describing crash information, crashes can be represented by point symbol maps, on which different symbols and/or colors indicate different types of road crashes. Since crashes are always constrained to road networks, policymakers may prefer to make observations by line segments. Street segments can then be differentiated based on the number of crashes. Often, thicker lines show higher densities of crashes. For those who are more concerned with area-based data (such as local district councilors), crash densities or other ratios can be derived. Figure 24.1 is a map showing the percentage of crashes involving pedestrians by Tertiary Planning Unit (TPU) in Hong Kong from 2005 to 2007. It was made by using the choropleth mapping strategy that categorizes data values (share of pedestrian crashes) into several classes and assigns a unique color, shade, or texture to each interval. While traditional methods use two-dimensional (2D) symbols to delineate crash information, research in recent years has attempted to portray crash patterns in a 3D environment by using 3D symbols such as gray tone and height (Xie and Yan 2008). Most GIS platforms provide an array of options for users to create an effective map representation of crash information from simple mapping of crashes to showing statistical results after modeling the spatial trends.

Spatial Analysis

Spatial analysis refers the “general ability to manipulate spatial data into different forms and extract additional meaning as a result” (Bailey 1994). Many GIS functions allow users to do more than simply managing and displaying spatial data. Here we discuss four main related issues that are widely used in crash analysis and prevention: measurement, topological analysis, network analysis, and statistical spatial analysis.

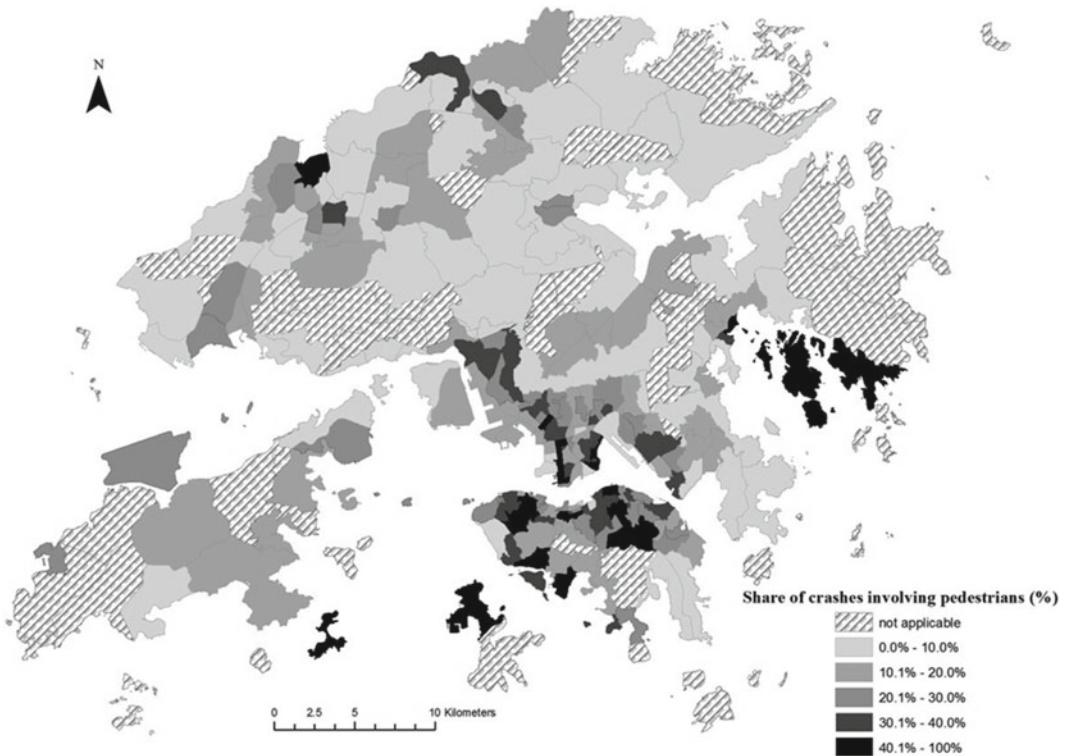


Fig. 24.1 A choropleth map showing the relative importance of pedestrian crashes at TPU level in Hong Kong, 2005–2007

Measurement

Measurement is important to all kinds of spatial analysis. If one wants to know the Euclidean distance between the site of a crash and the nearest hospital, the measurement function in GIS would be used to perform the calculation. When a road segment with exceptionally high number of crashes is of interest, the measuring tools can calculate the length and degree of curvature of the road, or even the gradient in a 3D environment.

Topological Analysis

Topological analysis is used to generate new information about spatial relationships among observations. A typical operation is to create buffers around points, lines, and areas. For instance, Subramanian (2009) examined the relationship between the locations of fatal road crashes in rural areas and their distance to urban areas. Also, to examine the extent to which bicycle crashes are related to bicycle tracks in Hong Kong, Loo and Tsui (2010) wanted to answer questions such as “What’s the number of bicycle crashes located within 100 m of bicycle tracks?”, “How about within 200 m?” and “How about 500 m?”. To answer these questions, they first created buffers around bicycle track centerlines by differing buffer distances. Figure 24.2 is an example of buffering bicycle tracks to generate the area within 100 m of the bicycle tracks in one part of Hong Kong. Next, the buffers and the bicycle crash datasets were overlaid. By using the topological operator “WITHIN,” the number of crashes in

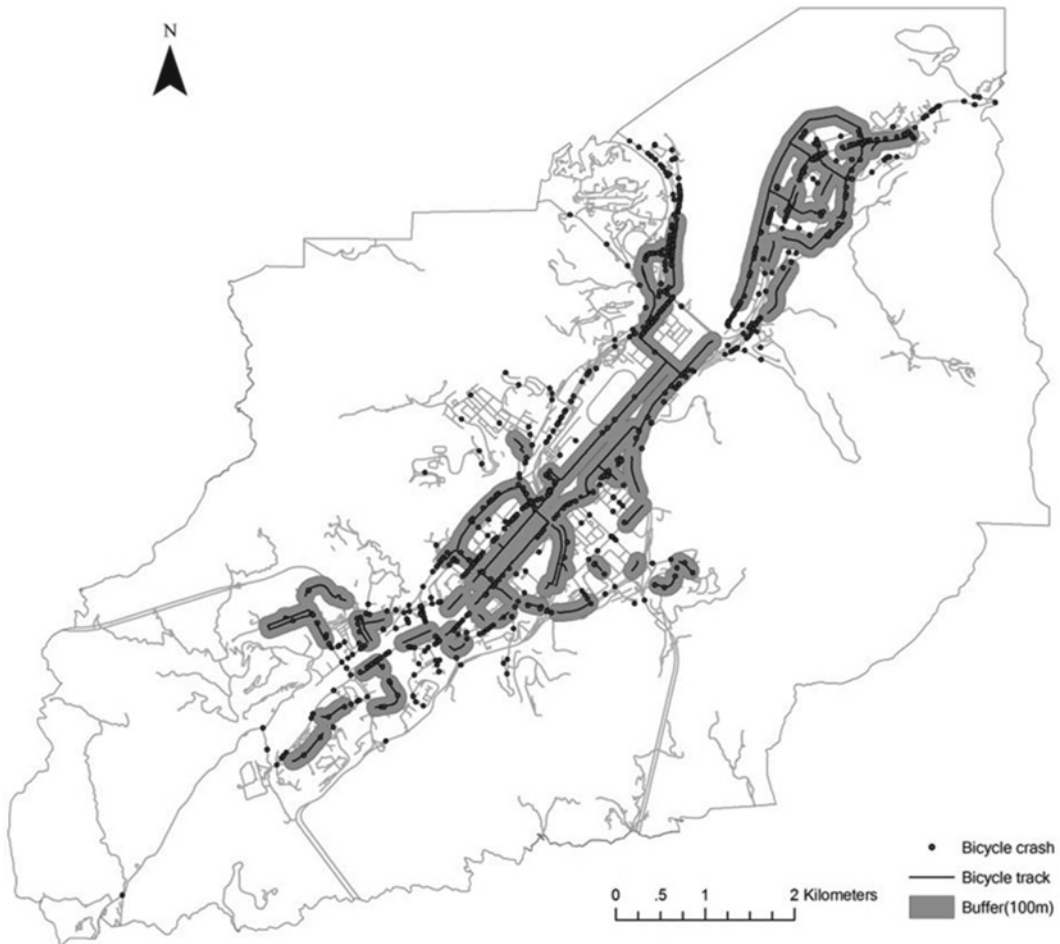


Fig. 24.2 Buffering lines representing bicycle tracks to show the area within 100 m of the tracks

each buffer was calculated. Apart from “WITHIN,” GIS provides a set of spatial query filters to create new information on spatial relationships. In the case of investigation of pedestrian crash rates at the TPU level (see Fig. 24.1), crash records do not typically contain relevant information on TPUs. The topological operator “INTERSECT” can be used to relate the crash counts with the local characteristics of different areas.

Network Analysis

Network analysis deals with flows through a network such as the road system that is modeled by nodes and links. In the emergency medical systems (EMS), network analysts can help the public health service providers to find the optimal route from the ambulance dispatch place to the location where the crash occurred (Derekenaris et al. 2001; Ganeshkumar and Ramesh 2010). Such analysis involves the shortest path algorithms that consist of nodes and arcs. Each arc connects two nodes and has its value representing the cost (weight) such as distance and travel time. These networks can be modeled by a vector GIS, in which users can even use the number of crashes as the weight to select

safest routes (Graettinger et al. 2005). While measurement and topological functions are generally available in all GIS packages, network analysis function is not a common feature. Nonetheless, many GIS providers have developed network analysis products that can be integrated into corresponding GIS platforms. For instance, ArcGIS supports a powerful network analysis extension named “Network Analyst,” whereby the police can quickly identify nearest medical facility locations for injuries or an ambulance driver can conveniently identify an appropriate path to reach the victims in need.

Statistical Spatial Analysis

Statistical spatial analysis is widely used in modeling spatial patterns or trends. The two major types are event-based and link-based analyses. In event-based analyses, crashes are represented as points. This kind of analysis can be further classified into distance-based methods that examine distances between events, and density-based methods that examine the crude density or overall intensity of a point pattern (O’Sullivan and Unwin 2003). Frequently used distance-based methods include nearest-neighbor distance and distance functions such as the G , F , and K (O’Sullivan and Unwin 2003). Of these, Ripley’s K -function has been utilized for crash analysis in many studies (Jones et al. 1996; Schneider et al. 2004). Loo and Tsui (2005) used the nearest neighbor analysis (NNA) to explore the crash clustering tendency. The Pythagorean theorem or Pythagoras’ theorem is used to calculate the distance of each crash to its nearest neighbor. Although conventional distance-based methods were originally developed for 2D space, researchers have extended them to one-dimensional (1D) where distances between events are more appropriately calculated using network measurements other than Euclidean distances. Examples of 1D methods for crash analysis include Okabe et al. (2006a, b), Yamada and Thill (2004, 2007), and Strauss and Lentz (2009). Notably, the Midwest Transportation Consortium conducted the network K -function on SANET with Iowa’s crash data and evaluated the degree and scale of their spatial distributions (Okabe et al. 2006a, b; Strauss and Lentz 2009). The alternative to distance-based methods is density-based measures. Quadrat count methods and density estimation belong to this type. The kernel-density estimation (KDE) methods are particularly promising in analyzing crash patterns (Anderson 2009; Delmelle and Thill 2008; Pulugurtha et al. 2007; Erdogan et al. 2008). Kernel density calculates the density of events in a region (or a kernel) around those events. Transforming from a distribution of discrete crashes to (planar) density estimates involves the generation of a continuous raster surface. A kernel is fitted over each crash. The kernel value is highest at the location of the crash and diminishes with increasing distance, reaching zero at or beyond the bandwidth (search radius distance) from the crash. The density at each output raster cell is calculated by summing up the values of all the kernels where they overlay the raster cell center. Figure 24.3 shows the crash density surface of Hong Kong in 2007, which was computed by density analysis tools of ArcGIS. The kernel function was based on the quadratic function with 1 km as the bandwidth. Higher density values indicate higher incidence of road crashes. From the map, one can get a better sense of hazardous road locations, which are displayed in darker color. Besides, the result could be further analyzed by GIS with other continuous surfaces such as air pollution and traffic density to evaluate environmental risks. Recently, density methods are also applied to network space. Pioneering work includes Xie and Yan (2008) and Okabe et al. (2009).

Since crashes can be aggregated by road segment or area, there are many studies focusing on link-attribute- or area-attribute-based analysis. Some spatial models that have been applied to these crash analyses investigate spatial autocorrelation, which evaluates spatial dependency between the value of a variable at a location and the value of the same variable at nearby locations. An important element in these spatial models is a matrix W containing weights w_{ij} that describe the spatial relationship (e.g., contiguity, proximity, or connectivity) between unit i and j . Moran’s I , Geary’s C , and

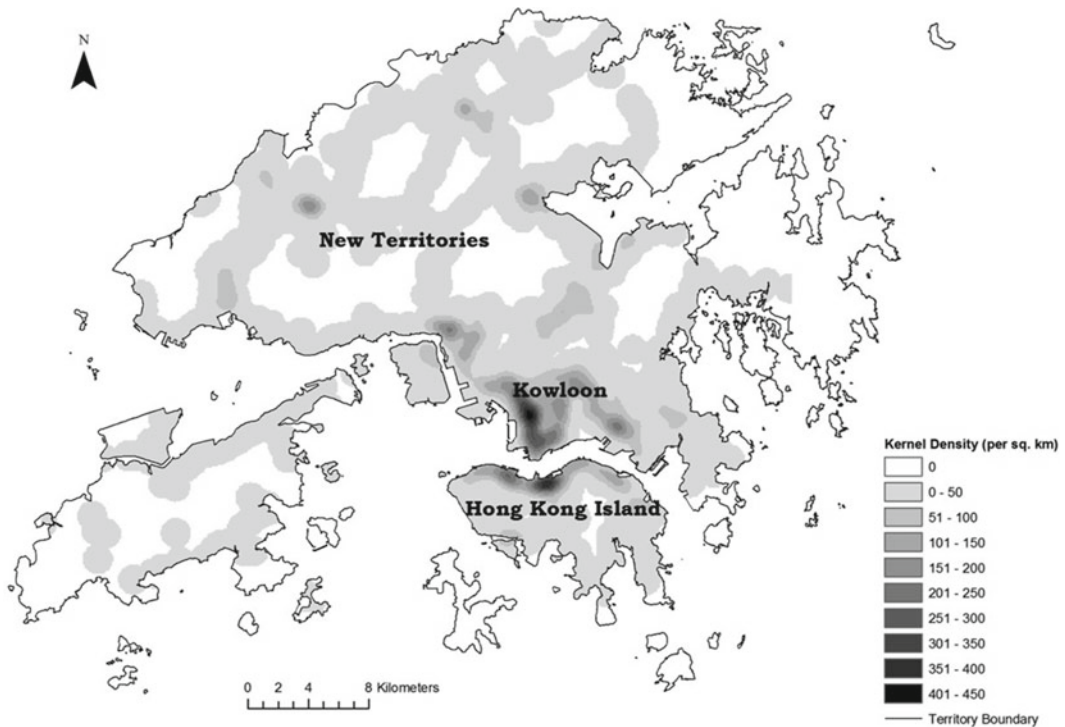


Fig. 24.3 Kernel density surface showing dangerous road locations in Hong Kong, 2007

Getis–Ord General G are three well-known statistics, whereby crash analysts can examine spatial autocorrelation at the global level (that is, the entire dataset) (Black 1991; Erdogan 2009; Hewson 2005). For instance, to show how the crash statistics were correlated among provinces in Turkey, Erdogan (2009) analyzed the clustering of crashes by Moran’s I, Geary’s C, and General G. Although these statistics are designed for global measures, their local versions can detect clustering tendency at the local level. These local indicators thus can be used for crash “hot zones” (or “black zones”) identification (Black and Thomas 1998; Flahaut et al. 2003; Flahaut 2004; Loo 2009; Yamada and Thill 2010). In addition, spatial dependence effects can enter into regression models, such as the spatial lag model (SLM), which introduces a spatial lag variable and the spatial error model (SEM) that incorporates a spatial error term. Applications to crash analysis include the study of demographic factors in pedestrian–vehicle collisions (LaScala et al. 2000), urbanization and crash rates (Wang and Kockelman 2007), and weather conditions on crash occurrence (Brijs et al. 2008).

Apart from spatial autocorrelation, spatial heterogeneity has also attracted much attention in recent years. Geographically weighted regression (GWR) is an attempt at exploring spatial heterogeneity (Brunsdon et al. 1996), which examines relationships among variables varying from location to location. Erdogan (2009) modeled crash and death rates by GWR under the assumption that there was q nonstationary spatial relationship between variables and found that the GWR model significantly improved model fitting over the OLS model.

Bayesian methods have become more and more popular in dealing with both spatial autocorrelation and heterogeneity effects. Extensive studies have been reported in safety research such as analyzing spatial–temporal patterns of motor vehicle crashes and ranking sites for safety improvements (Quddus 2008; Li et al. 2007; Miaou et al. 2003; Aguerro-Valverde and Jovanis 2006, 2008; Miaou and Song 2005). Some GIS packages have integrated statistical spatial functions such as Geostatistical

and Spatial Analyst provided by ESRI. Besides, there are also some alternative professional products for statistical spatial analysis, e.g., GeoDa developed by Luc Anselin, SpDep by Roger Bivand based on R, Spatial Econometric Toolbox by Lesage based on Matlab, *slm_panel* by J. Paul Elhorst based on Matlab, GWR by A. Stewart Fortheringham, and Winbugs (OpenBUGs).

Case Study of the Identification of Hazardous Road Locations

The identification of hazardous road locations is important before any site-specific safety improvement programs can be implemented. Since it involves spatial data manipulation, GIS technology has been widely used for this kind of analysis. For example, to identify dangerous “hot zones,” the first step is to identify exact crash locations. Next, the road network needs to be cut into small segments and the crash rate for each segment is calculated. After defining the threshold value, links having crash rates above the threshold value can be identified. Positive spatial autocorrelation between dangerous road segments nearby is then identified as hot zones. The whole process depends much on spatial information management and analysis. The following subsections illustrate the power of GIS in resolving “spatial” problems during the identification of hot zones. The procedures are also reported by Loo (2009).

Validation of Crash Locations

A high level of precision about the spatial location of a crash is vital for any meaningful spatial analysis, which ranges from simple visualization of locational patterns to complex modeling of spatial trends. It is, therefore, necessary to validate the location of road crashes before conducting any scientific spatial analysis. Using Hong Kong as an example, the crash database of Hong Kong, known as traffic accident data systems (TRADS), describes the police crash investigation data for every year. In particular, it stores five-figure grid references that could be transformed into projection coordinates by GIS as the precise location of crashes. Each crash could be plotted onto a digital map in a GIS environment based on the x and y coordinates. Meanwhile, the nonspatial road crash information such as severity, time, place, or surrounding environment could be stored as variables in an attribute table that is also managed by the GIS platform. When the crash database is linked to the traffic network, it is known as the traffic information system (TIS).

As a 1D phenomenon, traffic crashes should be constrained to a road network. However, for both technical and nontechnical reasons, they are unlikely to intersect with the centerlines of the road network (link–node system). Figure 24.4 delineates a tiny part of the crash map in Hong Kong in the year 2007. It is obvious that most crashes were not located on the road network. Taking the whole territory of Hong Kong as an example, almost all crashes (no less than 99%) did not intersect with the road links in the period 1993–2004 (Loo 2006). Hence, these crashes need to be snapped to the appropriate junctions or centerline of the road network. By using GIS topological tools, crashes are first snapped to the nearest crossings or line segments. Next, the name of the road on which a crash is located can be obtained from the attribute table of the road network database. It is then compared with road name information that is recorded in the crash attribute table such as landmark (INDE_FIR), precise location (PREC_LOCTN), first street name (FROAD), second street name (SROAD), and circumstances (HAPPEN). Figure 24.5 describes the crash distribution after snapping the original crashes (see Fig. 24.4) to the nearest road intersection or segment and the attributes used for road name matching. As shown in the figure, crash A has been moved to the nearest line segment and the name of the road on which it is located now is SOY Street. The system is then trying to find “SOY

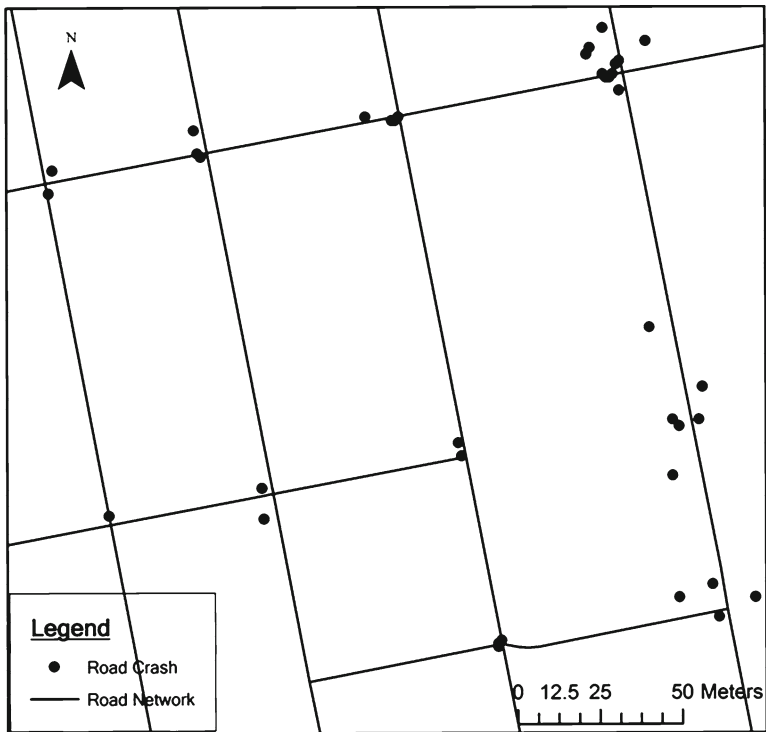


Fig. 24.4 Distribution of original crashes

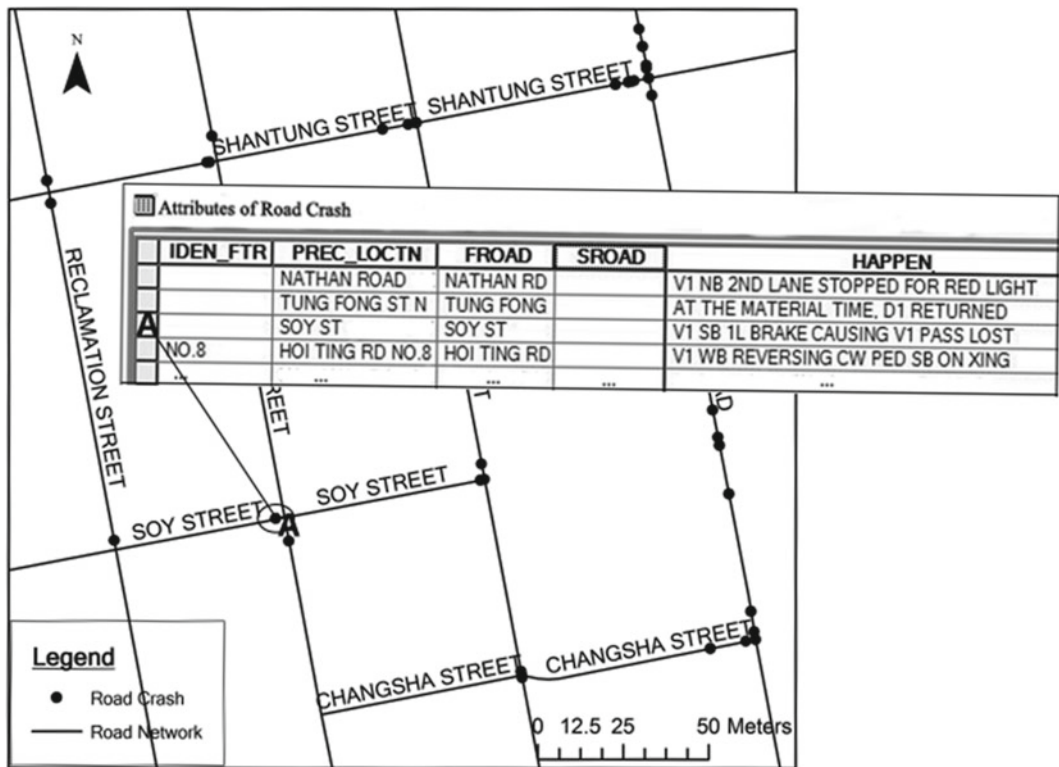


Fig. 24.5 Distribution of crashes after snapped to the nearest road junction or segment

Street” among the five variables in the attribute table of crashes. As “SOY Street” can be found in `PREC_LOCTN` and `FROAD`, the matching is successful. The precise location for crash A is thus validated. If the matching fails, the original crash will be snapped to the next nearest road junction or link and a new round of road name matching is then performed. For each unmatched crash, the next phase is to compare the road names stored in those five spatial variables with names in the road network database. If a match is found, the crash is then snapped to that “matched” road junction or link. Finally, missing road names or typo errors are checked manually. The rest of the crashes are snapped to these corrected road locations. Following this procedure, all crashes are located on the appropriate road junctions or segments. Details of GIS-based validation method discussed below can be found by Loo (2006).

Segmentation of Road Network

Although there is no clear indication of the best length of a dangerous road segment, most researchers recommend the use of a constant value. In this regard, the entire road network is always cut into segments with an equal interval, which are referred to hereafter as basic spatial units (BSUs). As a general rule, the length of a BSU is often defined as 100 m (Black 1991; Black and Thomas 1998; Flahaut et al. 2003; Flahaut 2004; Loo 2009). Theoretically, if the targeted road network of the study area, for instance, is about 1,000 km long, there should be 10,000 BSUs in total. However, the number of BSUs is always much higher, because the equal interval condition is likely to be violated near the end nodes of a road link. This effect is most obvious when an empirical link–node system is used, especially in those places where there are dense roads. For example, the entire road network system with annual traffic volume information in Hong Kong consisted of 6,445 links in GIS with a total length of 1,090 km. When this GIS Hong Kong road network is cut into BSUs of 100 m, 14,292 BSUs resulted. The number is 31% higher than that suggested by simply dividing the total length of roads by the length of a BSU. The shares of BSUs less than 25 and 50 m reach 12% and 24%, respectively.

Since such small segments are often not long enough for the identification of crash clusters, the links of the road network should first be dissolved to diminish the negative influence of short links. The main problem is that road network is complex and a link is always connected to more than one segment (i.e., they have an end node in common). As shown in the hypothetical road network in Fig. 24.6, a typical link has more than one neighbor. To determine which segment is dissolved, a priority sequence can be used for dissolving the network. Consider the road network structure in Fig. 24.6; we start with link 1. Link 1 can be dissolved with links 2, 5, and/or 6 because these three links are all connected to link 1 at the common end node B. Since the link–node road network system always stores information on road names, here we first dissolve the roads with same road names. Back to Fig. 24.6, since link 2 shares the same road name with link 1, the two links are first merged. Next, the GIS algorithm looks for continuous segments at end node C. As they share the same road name, link 3 is dissolved with link AC. Then, a new round of dissolving work begins at point D, but the road names of the three segments are different from each other. Under such circumstances, link AD may either be randomly dissolved with one segment or be merged according to a preset rule. For example, we may create a normal line for each of three links at end node D and calculate the angle between normal lines of link AD and link 4, as well as that of link AD and link 8. The one with a small angle will be picked out as the merged segment. Although this kind of dissolving task seems laborious, it can be programmed with the help of GIS by using measurement and topological functions. Moreover, the dissolving task will end once the length of a BSU reaches its maximum length (100 m in this example). After the above procedures, the empirical network database of Hong Kong consists of only 871 links and 11,401 BSUs. Moreover, only 4.5% are less than 50 m long. It is obvious that the merging work can dramatically reduce the negative effect of a dense road link–node system on the length of spatial units.

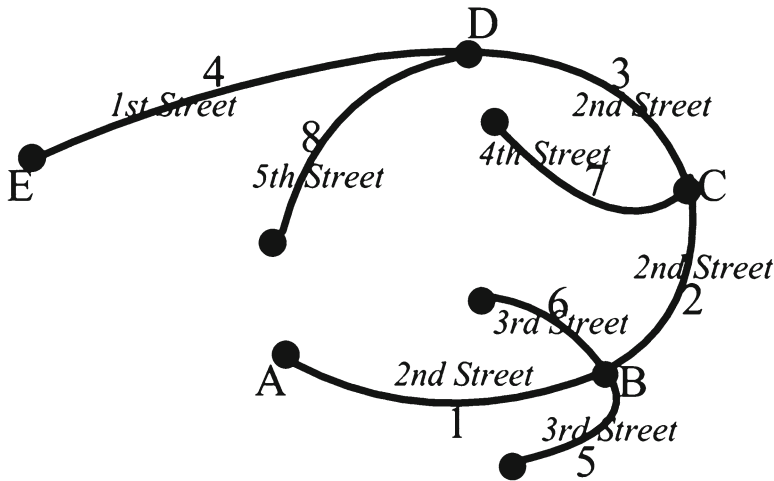


Fig. 24.6 Links in the hypothetical network

Calculation of Actual Crash Rates

After the generation of BSUs, the actual crash rate (which may be the crash number in a year or the average crash number for a period of 3 to 5 years) for each BSU needs to be calculated. The fundamental operation here is geometric intersection. With topological operators provided by GIS, the crash layer is intersected with the BSU layer. The number of crashes occurring on each BSU can then be computed. However, due to a great number of crashes located at road junctions, double counting of road crashes can happen frequently. To solve this problem, one may assign the crash to one of the BSUs by random selection or by a predefined rule. For example, GIS stores the location of a BSU by recoding a series of geographic coordinates. The minimum and the maximum x (x_{Min} and x_{Max}) and y (y_{Min} and y_{Max}) coordinates of each BSU can be calculated. The crash can then be assigned to one of the BSUs according to their locations, such as the left (smaller x_{Min}), the right (larger x_{Max}), the upper (higher y_{Max}), and the lower (smaller y_{Min}).

Determination of the Threshold Value

In general, there are three common definitions of the threshold value, above which the crash rate is considered “high” and worthy of further investigation. These three common definitions are the numerical, statistical, and model-based definitions (Elvik 2007). Simple numerical definitions are always favored by administrations, such as the official Norwegian definition, which denotes any site with a maximum length of 100 m where at least four injury crashes have been recorded during the last 5 years (Sorensen and Elvik 2007). A statistical definition refers to a normal number of a similar type of targeted location. For example, a site on an expressway is classified as a hot spot if its actual crash rate is significantly higher than the normal number of sites on that expressway. Model-based definitions are derived from crash prediction models, among which the EB method is found to perform best in differentiating the “true-positive” and the “false-positive” locations.

An alternative method, Monte Carlo simulation, which is widely used in defining threshold levels in crash analysis (Yamada and Thill 2007, 2010) can also be applied to the definition of crash threshold values for BSUs. Each simulation round includes randomly assigning the same total number of

crashes to the entire road network and computing the simulated crash rate for each BSU. Since it is not practical to allocate crashes to every possible location of the road network, we may use GIS to choose some representative points with an equal interval along the network in a similar manner to the geographical analysis machine (GAM) of Openshaw et al. (1987). If we denote m as the number of crashes and n as the number of representative points, m out of n points are randomly selected to simulate the crash pattern. Assuming that we repeat the simulation 1,000 times, for each BSU, the tenth largest value is then used as the threshold level at the significance level 0.01.

Modeling of the Spatial Pattern

As mentioned earlier, modeling spatial patterns or trends is a main advantage of GIS. By using spatial models, one can get more hints of the incidence of crashes. For instance, on the basis of local Moran's I method (Anselin 1995), Loo (2009) provides an indicator for detecting hot zones. Depending on the spatial relationships among BSUs, the indicator, $I_{(HZ)i}$, can be defined as:

$$I_{(HZ)i} = z_i \sum_{j=1, j \neq i}^n W_{ij} z_j, \tag{24.1}$$

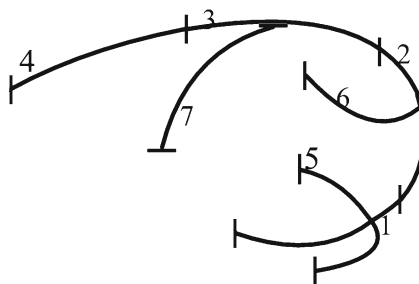
$$z_i = \begin{cases} 1 & \text{if } a_i \geq t_i \\ 0 & \text{otherwise} \end{cases}, \tag{24.2}$$

where n is the number of BSUs, t_i is the threshold crash rate of BSU i , a_i is the actual crash rate at the i th BSU, and W_{ij} is the network proximity matrix. As mentioned earlier, matrices are widely used in spatial analysis for representing spatial concepts such as distance, adjacency, interaction, and neighborhood. For hot zone identification, we concentrate on those contiguous BSUs with relatively high risks. Thus, W_{ij} is often denoted as a contiguity 0–1 matrix whose elements are only ones or zeros. Quantifying such relationship can be done by GIS. For instance, following the hypothetical structure of seven BSUs in Fig. 24.7, the weight matrix, W , can be calculated by GIS:

$$W = \begin{bmatrix} * & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & * & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & * & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & * & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & * & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & * & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & * \end{bmatrix}. \tag{24.3}$$

$I_{(HZ)}$ is computed with the assistance of GIS and the result is recorded as a variable in the attribute table of the BSU dataset. The value of $I_{(HZ)}$ is either positive ($I_{(HZ)}i = 1, 2, \dots, n - 1$) or equal to zero. A positive $I_{(HZ)}i$ value indicates that the actual crash rates of BSU i and its $I_{(HZ)}i$ neighboring BSUs, are no less than their threshold values. For hot zone studies, only these positive $I_{(HZ)}i$ values are of interest. Those BSUs with non-zero $I_{(HZ)}$ are then figured out and crash hot zones can be detected. Details of the GIS-based algorithm for implementing hot zone identification are reported by Loo (2009).

Fig. 24.7 BSUs in the hypothetical network



Display of the Analysis Results

The visualization and mapping functions of GIS can help display analysis results for different purposes. This is different from the simple mapping of crashes. Taking Hong Kong as an example, the road network with AADT is about 1,090 km long. Based on the procedures of geo-validation, 10,307 road crashes in the year of 2007 were snapped to the nodes or links. The road network was then cut into 100 m and 10,307 BSUs were generated. Following (24.1) and (24.2), 113 hot zones with 308 BSUs were identified if a threshold crash rate value of 4 is used to make the results comparable with the official blacksite definition (Loo 2009). It was observed that some hot zones comprised only two BSUs, whereas some contained more than ten units. The actual crash rate of a BSU varied from 4 to 26. To explore the most hazardous road locations, all the hot zones were categorized into four classes based on the total number of crashes. As shown in Fig. 24.8, the hot zones can be clearly portrayed using different symbol sizes. To further compare the actual crash rates of BSUs, a 3D map was produced in a three-dimensional environment using ArcScene of ESRI. Figure 24.9 shows 3D hot zones represented by black walls. The height of the wall indicates the actual crash rate of a BSU. From this map, one can not only identify hazardous zones on the road network but also find out the most dangerous locations with these zones.

GIS platforms such as ArcMap also enable users to produce graphs (e.g., histogram, vertical bar, horizontal line, and scatter plot) for some exploratory analysis. If a research team, for example, is interested in those dangerous locations where pedestrians are more likely to be involved, these crashes can be classified by collision type. These nonspatial data are recorded as attributes by GIS. Pie charts can then be added for illustrating the percentages of different types of crashes.

Conclusion and Ways Forward

To conclude, GIS is widely utilized in crash analysis. It is easy for GIS to manage a large amount of spatial and nonspatial data related with the road crashes, the surrounding environmental features such as road network and land uses, the vehicles such as vehicle type and age, and the road users such as driving experience and risky behavior such as driving under the influence. It also provides various methods for mapping and visualization in either 2D or 3D environment, such as the plotting of crash points using geographic coordinate system and displaying the results of crash analysis by different attributes (e.g., actual crash rate of a hot zone and type of collisions). More crucially, GIS is beneficial for dealing with spatial relationships between geographical features. Examples include measuring the distance between a crash and road network to examine the proximity, dissolving road links at an end node to take into account connectivity, and calculating weight matrix for hot zones to reflect contiguity.

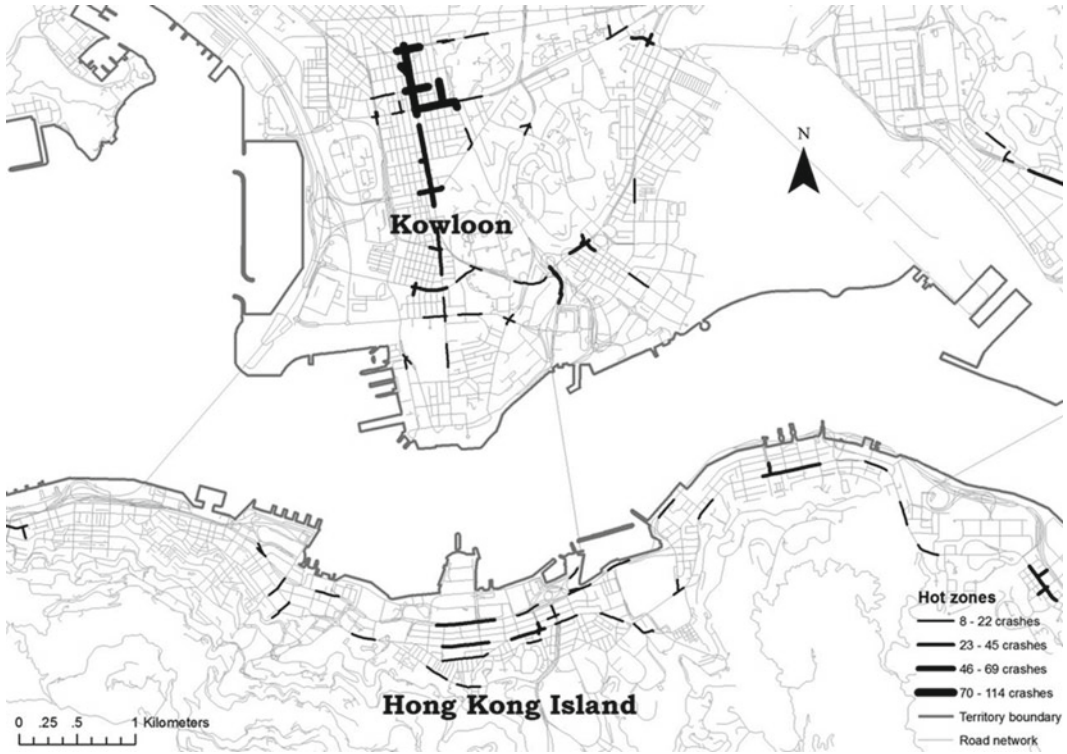


Fig. 24.8 2D map describing hot zones in part of Hong Kong, 2007

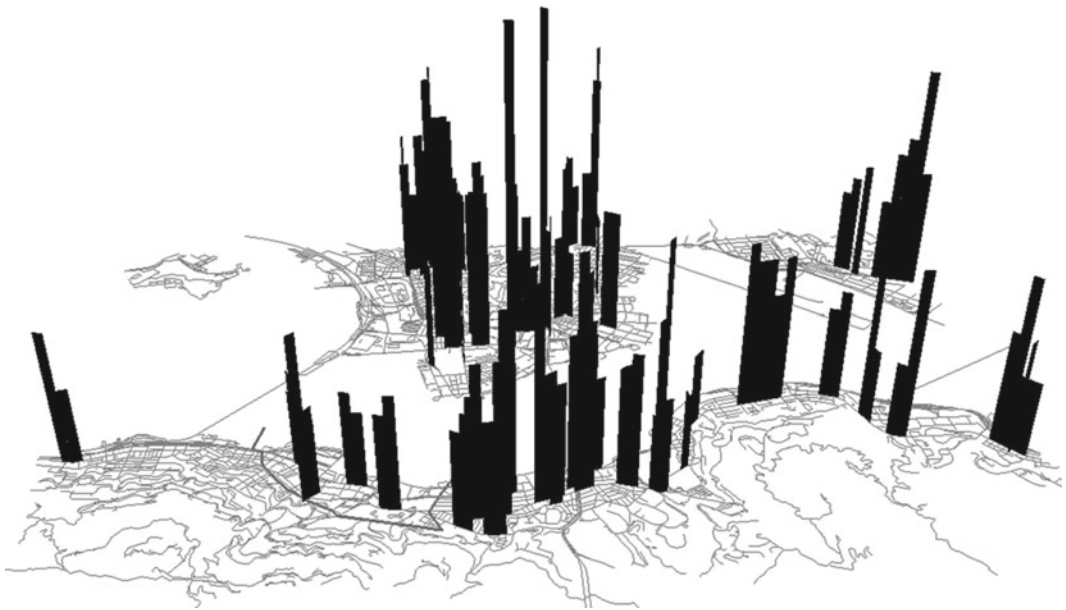


Fig. 24.9 3D map describing hot zones in part of Hong Kong, 2007

GIS has added great values to injury analysis, which is of high relevance to hospitals, policy makers, transport researchers, and the public. For instance, medical researchers could analyze spatial patterns of multiple traffic casualties or fatal and serious crash injuries by using geographical information methods. From the results of the identification of hazardous road locations, policy makers could offer strategic principles and proactive countermeasures to reduce crashes and injuries; engineers would install facilities such as traffic lights and pedestrian refuges, or modify vehicle design or road infrastructure to improve safety; the public could get better understanding of dangerous places and would hence be more likely to support policies and practical measures made by administrations. It is obvious that GIS is useful for preventing the future occurrence of road injury. However, GIS applications in crash analysis and prevention require a better conceptual understanding of spatial data and methods. The lack of this understanding is a barrier to the wider use and application of GIS in road safety research. More collaborative efforts between practitioners in agencies and researchers of GIS and other disciplines would be needed to address these problems.

References

- Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention*, 38(3), 618–625.
- Aguero-Valverde, J., & Jovanis, P. P. (2008). Analysis of road crash frequency with spatial models. *Transportation Research Record: Journal of the Transportation Research Board*, 2008, 55–63.
- Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention*, 41, 359–364.
- Anselin, L. (1995). Local indicators of Spatial Association – LISA. *Geographical Analysis*, 27(2), 93–115.
- Bailey, T. C. (1994). A review of statistical spatial analysis in geographical information systems. In S. Fotheringham & P. Rogerson (Eds.), *Spatial analysis and GIS* (pp. 13–44). Bristol, PA: Talyor & Francis.
- Bertin, J. (1979). Visual perception and cartographic transcription. *World Cartography*, 15, 17–27.
- Black, W. R. (1991). Highway accidents: a spatial and temporal analysis. *Transport Research Record*, 1318, 75–82.
- Black, W. R., & Thomas, I. (1998). Accidents on Belgium's motorway: a network autocorrelation analysis. *Journal of Transport Geography*, 6(1), 23–31.
- Brijs, T., Karlis, D., & Wets, G. (2008). Studying the effect of weather conditions on daily crash counts using discrete time-series model. *Accident Analysis and Prevention*, 40(3), 1180–1190.
- Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis*, 28, 281–298.
- Cromley, E. K., & McLafferty, S. L. (2002). *GIS and public health*. New York: Guilford.
- Delmelle, E. C., & Thill, J.-C. (2008). Urban bicyclists: spatial analysis of adult and youth traffic hazard intensity. *Transportation Research Record: Journal of the Transportation Research Board*, 2074, 31–39.
- Derekenaris, G., Garofalakis, J., Makris, C., Prentzas, J., Sioutas, S., & Tsakalidis, A. (2001). Integrating GIS, GPS and GSM technologies for the effective management of ambulances. *Computers, Environment and Urban Systems*, 25, 267–278.
- Elvik, R. (2007). *State-of-the-art approaches to road accident black spot management and safety analysis of road network* (Report No. 883). Oslo: Institute of Transport Economics.
- Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research*, 40, 341–351.
- Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. *Accident Analysis and Prevention*, 40(1), 174–181.
- Flahaut, B. (2004). Impact of infrastructure and local environment on road unsafety: logistic modelling with spatial autocorrelation. *Accident Analysis and Prevention*, 46(6), 1055–1066.
- Flahaut, B., Mouchart, M., Martin, E. S., & Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. *Accident Analysis and Prevention*, 35, 991–1004.
- Ganeshkumar, B., & Ramesh, D. (2010). Emergency Response Management and Information System (ERMIS) – A GIS based software to resolve the emergency recovery challenges in Madurai city, Tamil Nadu. *International Journal of Geomatics and Geosciences*, 1(1), 1–13.
- Gatrell, A. C., & Loytonen, M. (1998). GIS and health research: an introduction. In A. C. Gatrell & M. Loytonen (Eds.), *GIS and health research*. London: Taylor & Francis.

- Graettinger, A., Lindly, J. K., & Mistry, G. J. (2005). *Display and analysis of crash data*. Tuscaloosa, AL: Alabama University Transportation Center for Alabama.
- Graham, D. J., & Glaister, S. (2003). Spatial variation in road pedestrian casualties: the role of urban scale, density and land-use mix. *Urban Studies*, 40(8), 1591–1607.
- Han, K., Middleton, D., & Clayton, A. (2006). Enhancing highway geometric design: development of interactive virtual reality visualization system with open-source technologies. *Transportation Research Record: Journal of the Transportation Research Board*, 1980, 134–142.
- Hearnshaw, H. M., & Unwin, D. J. (1994). *Visualization in geographical information systems*. Chichester, UK: Wiley.
- Hewson, P. J. (2005). Epidemiology of child pedestrian casualty rates: can we assume spatial independence? *Accident Analysis and Prevention*, 37(4), 651–659.
- Jha, M. K., McCall, C., & Schonfeld, P. (2001). Using GIS, genetic algorithms, and visualization in highway development. *Computer-Aided Civil and Infrastructure Engineering*, 16(6).
- Jones, A. P., Langford, I. H., & Bentham, G. (1996). The application of K-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk. *England Social Science & Medicine*, 42(6), 879–885.
- Khattak, A. J., & Shamayleh, H. (2005). Highway safety assessment through geographic information system-based data visualization. *Journal of Computing in Civil Engineering*, 19(4), 407–411.
- LaScala, E. A., Gerber, D., & Gruenewald, P. J. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. *Accident Analysis and Prevention*, 32(5), 651–658.
- Li, L., Zhu, L., & Sui, D. Z. (2007). A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intra-city motor vehicle crashes. *Journal of Transport Geography*, 15, 274–285.
- Loo, B. P. Y. (2006). Validating crash locations for quantitative spatial analysis: a GIS-based approach. *Accident Analysis and Prevention*, 38(5), 879–886.
- Loo, B. P. Y. (2009). The identification for hazardous road locations: a comparison of black site and hot zone methodologies in Hong Kong. *International Journal of Sustainable Transportation*, 3(3), 187–202.
- Loo, B. P. Y., & Tsui, M. K. (2005). Temporal and spatial patterns of vehicle-pedestrian crashes in busy commercial and shopping areas: a case study of Hong Kong. *Asian Geographer*, 24(1–2), 113–128.
- Loo, B. P. Y., & Tsui, K. L. (2010). Bicycle crash casualties in a highly motorized city. *Accident Analysis and Prevention*, 42, 1902–1907.
- Loo, B. P. Y., & Yao, S. (2010). Area deprivation and traffic casualties: the case of Hong Kong. In A. Sumalee, W. H. K. Lam, H. W. Ho, & B. Siu (Eds.), *Transportation and urban sustainability: Proceedings of the 15th International Conference of Hong Kong Society for Transportation Studies (HKSTS)* (pp. 67–72). Hong Kong: HKSTS.
- MacEachren, A. M., Buttenfield, B., Campbell, J., DiBiase, D., & Monmonier, M. (1992). Visualization. In M. M. R. Abler & J. Olson (Eds.), *Geography's inner worlds: pervasive themes in contemporary American geography*. New Brunswick, NJ: Rutgers University Press.
- Miaou, S. P., & Song, J. (2005). Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*, 37(4), 699–720.
- Miaou, S. P., Song, J. J., & Mallick, B. K. (2003). Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics*, 6(1), 33–57.
- Murray, C. J., & Lopez, A. D. (1994). Quantifying disability: data, methods and results. *Bulletin of the World Health Organization*, 72(3), 481–494.
- O'Sullivan, D., & Unwin, D. J. (2003). *Geographic Information Analysis*. Hoboken, NJ: Wiley.
- Odero, W., Rotich, J., Yiannoutsos, C. T., Ouna, T., & Tierney, W. M. (2007). Innovative approaches to application of information technology in disease surveillance and prevention in Western Kenya. *Journal of Biomedical Informatics*, 40(4), 390–397.
- Okabe, A., Okunuki, K., & Shiode, S. (2006a). The SANET toolbox: new method for network spatial analysis. *Transactions in GIS*, 10(4), 535–550.
- Okabe, A., Okunuki, K., & Shiode, S. (2006b). SANET: a toolbox for spatial analysis on a network. *Geographical Analysis*, 38(1), 57–66.
- Okabe, A., Satoh, T., & Sugihara, K. (2009). A Kernel density estimation method for network, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, 23(1), 7–32.
- Openshaw, S., Charlton, M., Wymer, C., & Craft, A. (1987). Developing a mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information systems*, 1, 335–358.
- Pulugurtha, S. S., Krishnakumar, V. K., & Nambisan, S. S. (2007). New methods to identify and rank high pedestrian crash zones: an illustration. *Accident Analysis and Prevention*, 39(4).
- Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. *Accident Analysis and Prevention*, 40, 1486–1497.

- Schneider, R. J., Ryznar, R. M., & Khattak, A. J. (2004). An accident waiting to happen: a spatial approach to proactive pedestrian planning. *Accident Analysis and Prevention*, 36(2), 193–211.
- Sorensen, M., & Elvik, R. (2007). *Black spot management and safety analysis of road network-best practice guidelines and implementation steps* (TOI Report 919/2007 RIPCOR/ISEREST Project). Oslo: The Institute of Transport Economics.
- Strauss, T., & Lentz, J. (2009). *Spatial scale of clustering of motor vehicle crash types and appropriate countermeasures*. Ames, IA: Midwest Transportation Consortium.
- Subramanian, R. (2009). *Geospatial analysis of rural motor vehicle traffic fatalities*. Washington, DC: National Highway Traffic Safety Administration.
- Wang, X., & Kockelman, K. M. (2007). Specification and estimation of a spatially and temporally autocorrelated seemingly unrelated regression model: application to crash rates in China. *Transportation*, 34, 281–300.
- WHO. (2009). *Global status report on road safety*. Geneva: World Health Organization.
- Wier, M., Weintraub, J., Humphreys, E. H., Stebo, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis and Prevention*, 41(1), 137–145.
- Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5), 396–406.
- Yamada, I., & Thill, J.-C. (2004). Comparison of planar and network k-functions in traffic accident analysis. *Journal of Transport Geography*, 12, 149–158.
- Yamada, I., & Thill, J.-C. (2007). Local indicators of network-constrained clusters in spatial point patterns. *Geographical Analysis*, 39, 268–292.
- Yamada, I., & Thill, J.-C. (2010). Local indicators of network-constrained clusters in spatial patterns represented by a link attribute. *Annals of the Association of American Geographers*, 100(2), 269–285.

Chapter 25

Spatial Regression

Jurek Grabowski

Introduction

Geographically weighted regression (GWR) is a group of regression models developed by Fotheringham, Charlton, and Brunson (Brunson et al. 1996; Fotheringham et al. 2002) in which the β coefficients are calculated multiple times for the same set of factors over a defined geographic space. GWR uses the attribute data and coordinate data (e.g., latitude and longitude) of some event such as a homicide, suicide, vehicle collision, or other unintentional injury to generate a mathematical model that determines the relationship between two or more variables. However, GWR differs from a typical linear regression in that GWR allows for that relationship to vary over geographic space.

The steps for performing a GWR analysis are fairly straightforward and generally follow this progression:

1. Obtain and geocode event data using a geographic information system (GIS) software.
2. Define the size and geographic boundaries of the catchment window. A catchment window is the area (typically in the form of a circle) from which data will be collected and analyzed. The size is measured by its radius, called bandwidth. A critical issue in the GWR process is the selection of a bandwidth size and a determination of whether the bandwidth is allowed to vary in size or remain fixed during the GWR analysis.
3. Determine the placement of the catchment area. Since GWR is an iterative process, estimating β coefficients for the same factor(s) at different points on a map, a systematic method for moving the catchment window throughout the map must be defined. In general, researchers center the catchment area on the location of the event data itself or center the catchment area on the centroid of some predefined geographic zone such as a census tract, zip code, police district, congressional district, or county.
4. Perform the GWR analysis using all the data points that fall within the catchment area, move the catchment window to a new center point as defined in step 3 above, and rerun the GWR analysis while noting that catchment windows will overlap.
5. Assess the goodness-of-fit to judge how well the GWR models fit the data. The researcher can also compare the difference between the GWR goodness-of-fit test (R_{GWR}^2) and linear regression (R_{Linear}^2) to determine the amount of variability that is explained by the spatial relationships between data points.

J. Grabowski, PhD (✉)

AAA Foundation for Traffic Safety, 607 14th Street, NW, Suite 201, Washington, DC 20005, USA
e-mail: jgrabowski@aaafoundation.org

6. Present results. It is difficult to summarize the voluminous number of regression models created in a GWR analysis in a concise, informative, and unambiguous manner. For each variable in the regression model, a suitable method to display this complex data is to create an XYZ graph, where X represents the longitude coordinate of each catchment window center, Y represents the latitude coordinate, and Z is the estimated β coefficient and then allow for different colors to represent the range of Z values.

The remainder of this chapter focuses on the theory and background of GWR, as well as the statistical underpinnings for using GWR methodology.

Background

The definition of epidemiology is often stated as the study and control of health-related events. While the methodologies for examining the determinants of health are well developed, methodologies for studying the effects of the spatial distribution of the determinants are less well established. Historically, the role of geographic studies in epidemiology has been important in identifying the etiology of disease by delineating the geographic variations in disease rates. However, by examining only the geographic variation in disease frequency, epidemiologists are limited to three broad types of geographic studies. First is a descriptive study in which the distribution of incidence cases of a disease or injury (with respect to an individual and the place of occurrence) is presented on a map. A classic example of this kind of study is John Snow's examination of the source and distribution of cholera cases in Soho, England's 1845 outbreak (Snow 1854). Second is an ecological study where aggregate measures of countries or populations, rather than individuals, are correlated to a disease or injury. Finally, there are studies that delineate the health effects that migration, whether international or internal, has on the migrants themselves (or their children), the areas from which they leave, regions they journey through, or countries in which they ultimately settle. The strength of these three types of geographical studies is that the results can be succinctly presented on a map. Unfortunately, most mapped results impel readers to speculate about etiology and rarely provide unbiased empirical evidence for a relationship between factors and location.

To accurately measure the relationship between two or more variables, a multiple regression analysis is traditionally performed. Conventionally, the dependent variable is the variable of interest that represents some measure of injury that is described by other measured factors called independent or predictor variables. To mathematically model the magnitude of the relationship between the dependent and multiple independent variables, researchers usually perform a multiple regression analysis. The output obtained from a standard regression analysis is referred to as a parameter estimate and is represented in a regression formula as β . Thus, the general expression of a regression model with k independent variables is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon, \quad (25.1)$$

where Y is a measure of the dependent variable, $\beta_1 \dots \beta_k$ are the estimated regression coefficients, $X_1 \dots X_k$ represent a measure of the independent variable, and ε represents an error term that most simplistically can be thought of as a representation of the effect of unmeasured variables that were omitted from the equation or some combined effect of the omitted variables (Pratt and Schlaifer 1984). Of special note is that the parameter estimates are constant across a geographic space, which means that any spatial variation in the model is measured by the error term.

There is also a practical problem resulting from the use of linear regression on data that can be categorized by geographic zones or areas. For example, say a researcher conducts a simple study identifying significant factors that increase the probability of automobile crashes in young drivers

residing in the state of Colorado. Not surprisingly, the researcher finds that for every 100 gallons of liquor sold in the state, the fatal crash rate increases by 1%. However, when the researcher stratifies the state into three regions (west, central, and east), she notices a particular phenomenon: in the west region, for every 100 gallons sold, crash rates increase by 2%, while in the central region, for every 100 gallons sold, there was no significant increase or decrease in crash rate, and in the east region, for every 100 gallons sold, there was a protective effect where crash rates decrease by 1%. Moreover, knowing that rural and urban areas within each region may be fundamentally different, she further stratifies the state into the 64 individual counties, reruns the same regression analysis, and finds that groups of counties within each region demonstrate wildly different results. The astute researcher realizes that she has encountered a classical example of Simpson's paradox (also known as reversal paradox and amalgamation paradox). This phenomenon occurs when an aggregate data set is stratified into two or more heterogeneous groups and reanalyzed, resulting in two (or more) opposing conclusions (Simpson 1951; Colin 1972). The researcher is faced with a dilemma when asked by her state legislators if her study supports the idea of curbing alcohol sales as a mechanism to decrease crash rates. The state-level analysis does support this idea, whereas the more localized analysis offers a different set of conclusions. All the conclusions the researcher made were statistically correct, but they reflected answers to subtly different questions that ultimately relied on how the event data were geographically categorized.

From the above hypothetical example, several inferences can be made. First, the β estimate from the researcher's first regression model can be termed a global measure in that it assumes to represent the situation in every part of the state of Colorado. Conversely, each of the β estimates from the individual counties can be termed a local measure. The relationship between global and local measures can be stated simply: as the spatial variation of local measures increases, the ability of global measures to accurately represent them decreases. Hence, by geographically stratifying the data, it is clear that the β estimate of the state model does not accurately represent the reality found in each region or each county. This concept that the relationship depends in part on the location where the measurement was taken is referred to as spatial nonstationarity. Unfortunately, identifying that spatial nonstationarity exists in a spatial data set is easier than identifying what causes it. Contextual effects of traffic safety appear to be well discussed and documented, for example, in a compendium examining the relationships between traffic safety culture and various risk factors (speeding, not wearing a safety belt, driving while intoxicated), whereas safety culture can be viewed as intrinsically different across geographical space shaped in part by variations in people's attitudes, behaviors, beliefs, values, preferences, and other contextual issues, such as politics and perceived history (AAA Foundation for Traffic Safety 2007).

This leads to the second inference: global measures are easy to comprehend since there is only one regression model with one set of β estimates to interpret. When the researcher in the above example treated her local statistic as a spatial disaggregation of the global statistic, she calculated the β estimates for 64 different models (one for each county). Understanding the relationship and the variability between the β estimates from two (let alone 64) models is difficult, if not impossible.

The third inference is that aspatial data (attribute data) and spatial data (attribute data for a particular location, typically measured by latitude and longitude) are fundamentally different in that each observation in an aspatial data set can be collected in a way that it is independent of other observations. In spatial data sets, two points on a map will be related to one another by a constant (the distance between the two points). This concept is referred to as spatial autocorrelation, and the data are commonly measured by Moran's I or Geary's C statistics that analyze the degree of dependency among events in a defined geographic space or map. The concept of spatial autocorrelation was succinctly summarized by Waldo Tobler in what is commonly known as the first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970; Sui 2004). If two points located near one another on a map are more related to each other than two points far from each other, a statistician would correctly conclude that the assumption

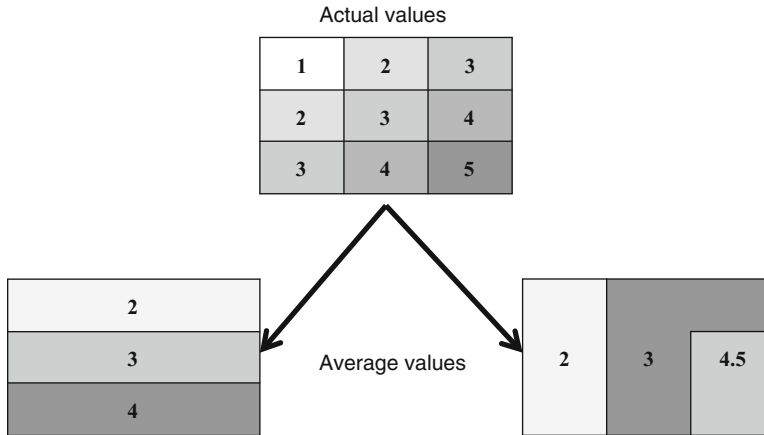


Fig. 25.1 Modifiable aerial unit problem: average values change as geographic boundaries change

of independence, in its strictest form, was violated when linear regression was used to analyze spatial data.

Finally, as the researcher changed analysis from the state level to the county level, the shape of the study area changed from a large rectangular shape (the state of Colorado) to multiple smaller amorphous shapes (the various shapes of each county). If the boundaries for analysis were to change within the map of Colorado, and if the researcher models automobile crashes and purchased alcohol volume stratified by census tract rather than by county, different patterns would likely emerge from the same spatial data. These resulting patterns may result solely from the aggregation procedure and nothing else; alternately, real relationships may be concealed by this aggregation process (Fig. 25.1). This is called the modifiable aerial unit problem, and it characterizes a potential source of bias when spatial data points are aggregated to bounded aerial units (or zones “artificially” defined by human construct through the process of drawing boundaries between two areas on a map such as census tracts or police districts).

Modeling Spatial Regression

Taking the aforementioned issues and knowing that the parameter estimates obtained in the linear regression are constant over space and that any spatial variations in a model being examined are measured within the error term, the challenge becomes: how a researcher can model the relationship between an outcome variable and the independent variables that are known to change over geographic space. Traditionally, the researcher would have mapped the regression residuals in an attempt to identify or describe spatial patterns with various autocorrelation statistics. However, Fotheringham et al. (2002) addressed the issue of spatial nonstationarity directly by allowing the relationships that were being measured to vary over a defined geographic space or map within a regression procedure called GWR. Thus, they let:

$$y_i = \beta_{0(u_i, v_i)} + \sum \beta_{k(u_i, v_i)} X_{ik} + \varepsilon_i,$$

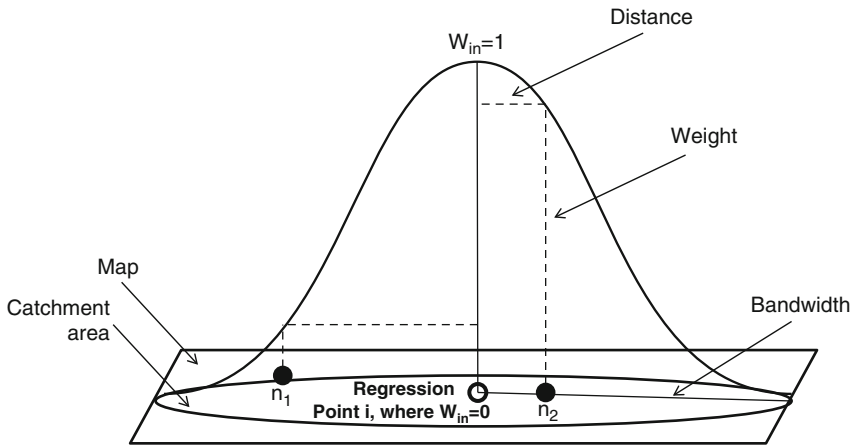


Fig. 25.2 A map with three data points depicting a spatially variable Gaussian weighting scheme for point i

where y_i represents the dependent variable, (u_i, v_i) represents the coordinates (e.g., longitude and latitude) of the i th point in space, $\beta_{k(u_i, v_i)}$ represents the value of the k th parameter at location i , and X_{ik} represents the k th independent variable in the i th location, and ϵ_i represents an error term for location i . Fotheringham obtained the parameters for a linear regression model using ordinary least squares and letting $\beta = (X^T X)^{-1} X^T Y$. Taking the first rule of geography into account, they then create $W(i)$, a matrix of weights for location i such that points on a map nearer to i are given more weight than points further away. Specifically:

$$W(i) = \begin{matrix} & W_{i1} & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ & 0 & W_{i2} & \cdot & \cdot & \cdot & \cdot & 0 \\ & 0 & 0 & W_{i3} & \cdot & \cdot & \cdot & 0 \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & W_{in} \end{matrix} ,$$

where W_{in} is the weight of data point n for the estimate of the local β parameter at location i (graphically represented in Fig. 25.2). Various methods can be used to calculate W_{in} , including a spatially adaptive Gaussian function, $W_{in} = \exp(-R_{in}/h_{in})$, where R is the ranked distance of all n points and h is the bandwidth size (Fig. 25.2). Similarly, if d_{in} is taken to be the distance between the regression point and data point n , a spatially variable bisquare function can be used to weight, where $W_{in} = [1 - (d_{in}^2/h^2)]^2$ if a data point is the N th nearest neighbor of point i and $W_{in} = 0$ if a data point is not a N th nearest neighbor. GWR is comparatively indifferent to the choice of weighting function so long as bandwidth distance decay is controlled for by selecting an optimal value of either h or n by minimizing a cross-validation score (e.g., jackknifing) or by use of the Akaike information criterion:

$$AICc = 2n \text{Log}_e(\hat{\sigma}) + n \text{Log}_e(2\pi) + n \{n + \text{tr}(\mathbf{S}) / n - 2 - \text{tr}(\mathbf{S})\},$$

where n is the number of data points, $\hat{\sigma}$ is the standard deviation of the error term, and $\text{tr}(\mathbf{S})$ is the trace of a matrix \mathbf{S} , $\text{tr}(\mathbf{S})$ (the sum of the diagonal elements of \mathbf{S}) defined by the relation: $y' = \mathbf{S}y$

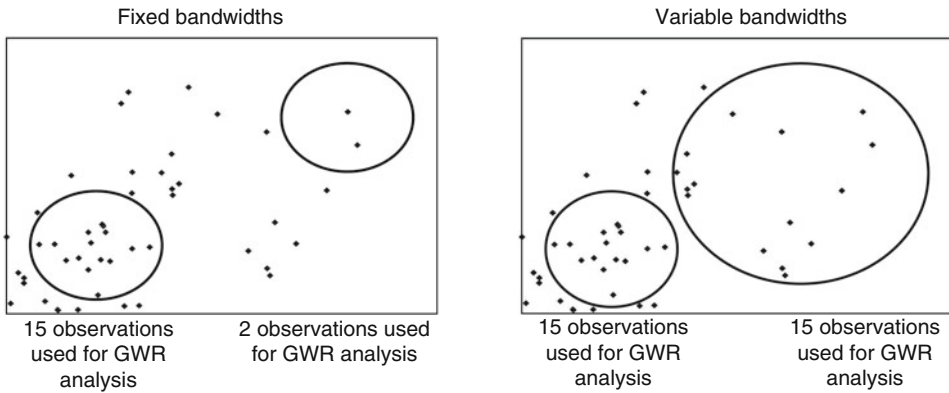


Fig. 25.3 A map of event data demonstrating catchment areas with fixed and variable bandwidths

(Akaike 1974; Fotheringham 2002). It is worth noting that a Gaussian function or a bisquare function with a *fixed* bandwidth can also be used in cases where the investigator wants the regressions to have relatively similar weighting structure, so that GWR will converge to the standard ordinary least squares. However, since the bandwidth is directly related to the influence of neighboring points in a spatial area, a map where events occur in irregular patterns, the fixed bandwidth function will encompass more points in dense (i.e., urban) areas of the map and fewer points in areas in which sampling stations are rare, such as rural areas (Fig. 25.3). Thus, the standard errors from various regression models will not be comparable. To avoid this, fixed bandwidths can be replaced by adaptive (also called variable) bandwidths described above.

While not described in this chapter, methods to calculate local standard errors, t statistics, goodness-of-fit measures, leverage measures, p values for tests of significance for spatial variation in estimated β (based on a Monte Carlo procedure), and tests to compare local GWR models to global regression models are readily available and explained in most GWR software packages. Moreover, it is worth noting that the GWR methodology can be extended to model count data using Poisson (log-linear) regression and to binary data using logistic regression by means of an iteratively reweighted least squares (IRLS) for model fitting. These methods are called geographically weighted Poisson regression (GWPR) and geographically weighted logistic regression (GWLR), respectively (Nakaya et al. 2005; Fotheringham 2002).

Linear Regression Versus GWR: An Example

Services provided by departments of emergency medicine are recognized as an important factor of overall population health in the USA and thus are often characterized as the nation's health-care safety net. In recent years, this safety net has been under strain due to emergency department (ED) overutilization by nonemergent cases, ED closures, and ED understaffing. This overutilization may manifest itself in the physical form of ED overcrowding. The majority of research and policy efforts to promote proper utilization of the health-care delivery system and eliminate disparities in ED health care have focused on individual-level patient characteristics and costs, resulting in an

abundant information on the association between ED health-care affordability and utilization rates. However, surprisingly little is known about how other issues impact emergency department utilization rates and population health. One of these issues is the spatial accessibility of EDs that can now be more aggressively addressed with advances in GIS and spatial regression statistics. For example, in addition to individual patient-level behavior to describe spatial accessibility to EDs, the socially conditioned neighborhood utilization behavior can be further studied in detail (Li et al. 2003; Li 2004).

For a patient, a number of issues can retard or promote progression from potential to realized ED utilization. Penchansky and Thomas (1981) grouped these issues into five dimensions: accessibility, availability, acceptability, accommodation, and affordability. Guagliardo (2004) further categorized acceptability, accommodation, and affordability as aspatial barriers to health care (reflecting the health-care financing arrangements and cultural aspects of these variables) and accessibility and availability as spatial in nature, where accessibility was defined as travel impedance (distance or time) between patient location and health-care locations, and availability was referred to as the number of service points from which patients can choose.

While the distinction between these aspatial and spatial variables can be useful in understanding the multidimensional concept that describes a population's ability to use ED services, these two types of variables should be considered simultaneously in the same statistical model to fully investigate neighborhood level factors that may increase or decrease ED utilization. The following example will compare and contrast two methods, a traditional multivariate regression model and a GWR model, to better understand the strengths and weaknesses of each.

In this example, the researcher was interested in determining if neighborhood block characteristics such as population density, proportion of vacant houses, and distance to the emergency department were associated with ED utilization. First, similar to the steps outlined in the beginning of the chapter, the study area was defined on a map, and then event data and the ED location were geocoded on the map. To avoid selection bias, the researcher picked a county within which one hospital provides all the emergency services to the population. The county was divided into 159 blocks defined by the US Census Bureau's national census. The study ED and each census block were plotted on a digital map created using GIS software. The centroid (the geometric center of the block's shape) for each block was plotted, and the straight line distance (in kilometers) from each centroid to the study ED was calculated using tools within the software. Population (number of people living on the block) and housing (number of occupied and vacant housing units) attribute data from Census 2010 data were merged into the map for each census block.

The study ED, with an annual volume of 57,000 patients, serves a population from both a western suburban area and an eastern rural area. Data on patients' home addresses were collected from the ED's administrative billing database containing information on all patients treated at the study ED in the year 2010. Included in the study were patients who resided within the same county as the hospital and who were 18 years or older. Excluded from the analysis were patients who reported a post-office-box address or a home address outside the study area or who did not specify an address. Patient home addresses were geocoded to longitude and latitude coordinates and plotted on the map containing all census blocks.

The ED utilization rate (per 100 population) for each centroid was calculated by dividing the number of ED visits from each block by the population living on that block times 100. To account for its skewed distribution, ED visit rates were transformed to logarithms when fitting the linear regression models. For each centroid, population density and proportion of vacant houses were calculated. Since the Census Bureau reports land area as square kilometer (to the nearest tenth), population density per square kilometer was calculated by dividing the total population living on that block by the land area. The proportion of vacant housing units was calculated per block by dividing the number of vacant housing units by the total number of housing units.

Analysis

Multivariate linear regression was used to assess the global relationships between neighborhood characteristics and ED utilization fitting the model:

$$\text{Ln(ED utilization rate)} = \beta_0 + \beta_1(\text{population density}) + \beta_2(\text{proportion of vacant housing units}) + \beta_3(\text{distance to the ED}) + \varepsilon.$$

To explain the local spatial relationships of neighborhood contextual factors of ED utilization rates, geographically weighted regression was modeled so that:

$$\begin{aligned} \text{Log}_e(\text{ED utilization rate}) = & \beta_{0(u_i, v_i)} + \beta_{1(u_i, v_i)}(\text{population density}) \\ & + \beta_{2(u_i, v_i)}(\text{proportion of vacant housing units}) \\ & + \beta_{3(u_i, v_i)}(\text{distance to the ED}) + \varepsilon. \end{aligned}$$

Using AICc, it was determined that data from each block's 23 nearest neighbors would be used for estimating β coefficients. Thus, using a moving catchment area centered over each of the 159 block centroids and calculating a regression model using the data from 23 of its nearest neighbors, 159 regression models were built. The results are displayed in Table 25.1.

We can assess the goodness-of-fit by comparing the R^2 of the regressions procedures, where a higher value indicates that the model fits the data better, for this example, by simply accounting for spatial relationships among the centroids; GWR better explains the variability between the outcome and independent variables.

Plotting the centroids' longitude and latitude and estimated β coefficients on a graph with XYZ axis, maps of estimated β coefficients from the spatial regression models were produced showing areas of positive β versus negative β . These maps can help the detection of neighborhoods that exhibit positive regression slopes versus communities revealing negative regression slopes (Fig. 25.4).

Both regression models identified that distance to an ED was a significant barrier to health-care access in this county. Specifically, the linear regression model indicated that for every kilometer that a patient traveled to the ED, there is an 8% decrease in the ED utilization rate, which implies that the spatial barriers to health-care utilization are explained by a global distance decay model: as time, cost, and effort of traveling to a health-care facility increase, patients' willingness and ability to travel decrease, resulting in decreasing utilization rates. In contrast, the GWR also indicates that distance to the ED decreased ED utilization rates in some census blocks, while in others, the opposite was true. The present example demonstrates that global measures of health-care (ED) utilization erroneously suggest that ED utilization follows the distance decay model. However, more variation in ED utilization can be explained by accounting for spatial relationships and Tobler's first law of geography.

To examine the spatial variability of only injury-related ED visits, a GWR model was created where injury-related ED utilization (per 100 population) for each centroid was defined as the dependent variable, and the same dependent variables (population density, proportion of vacant housing units, and distance to ED) were used as the independent variables. None of the β estimates was significant. This result may be explained by at least two issues. First, ED patients were geocoded by their home address, not by the geographic location of the injury occurrence. Hence, neighborhood factors may only weakly explain spatial variations in regions that have a high proportion of injuries due to traffic crashes or work/recreational place-related injuries. To address this issue, actual injury location data should be collected and analyzed. Thus, future studies comparing results using residence location versus event occurrence location will provide useful insights to the field of injury research.

Table 25.1 Linear versus geographically weighted regression

	Linear regression		Geographically weighted regression	
	β	p value	β range for 159 models	p value ^a
Intercept	2.61	0.04	0.99–4.88	0.01
Population density	0.01	0.15	–0.02 to 0.02	0.07
Proportion of vacant housing units	0.42	0.71	–1.23 to 1.12	0.46
Distance to ED	–0.08	0.05	–0.29 to 0.04	<0.001
R^2 (goodness-of-fit test)	0.16		0.28	

^aTests based on a Monte Carlo significance test procedure

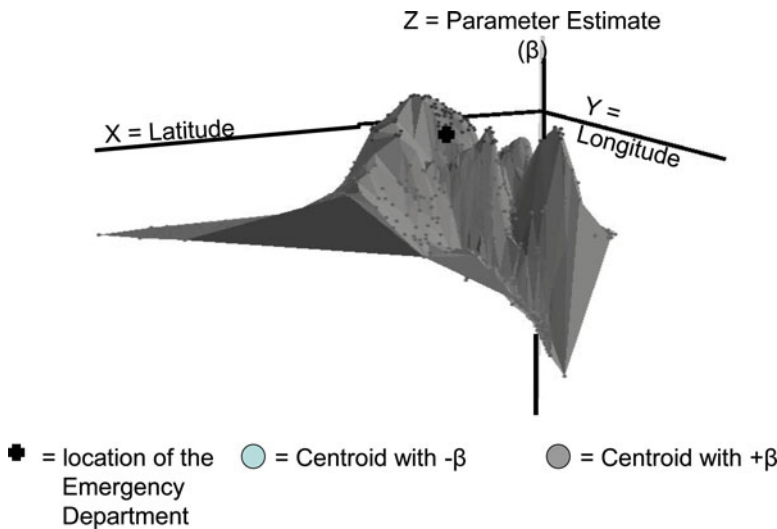


Fig. 25.4 Neighborhood level map of the estimated β coefficients from spatial regression models

Another issue that deserves consideration when interpreting these insignificant results is selection bias: persons living and injured in “rural” areas of the map may have a greater likelihood of dying at the scene, or they may not seek medical attention for minor injuries. Hence, they never made it to the ED and consequently were not included in analysis. More effort in injury research to capture location data on the full spectrum of injury severity will benefit future GWR analysis.

Current State of GWR in Injury Research

GWR offers a unique methodological insight to spatial data, signaling an important advance in injury epidemiology and research. Originally proposed in 1996, the GWR method was slow in being adopted into the medical literature. An online search of the National Library of Medicine’s PubMed database for the phrase “geographically weighted regression” identified only 40 publications that were made available from 1997 to mid-2011. Of these, only four were related to injury or violence as compared to the 18 published on the topic of various chronic diseases, eight on botany/geography/environment, four on infectious disease, three on health services research, two describing the GWR method, and one miscellaneous editorial. Upon closer inspection of the methods sections section of the 40 identified studies, ten were not related to GWR, including one of the injury studies. Of the

three injury papers that explicitly used GWR, two addressed the topic of traffic safety and one dealt with homicide in immigrant populations across Chicago neighborhoods (Hadayeghi et al. 2010; Erdogan 2009; Graif and Sampson 2009). What is apparent is that, to date, GWR has been underutilized in the study of injury. More disconcerting is a lack of methodological papers that test the merits of GWR in relation to other methods in the injury field. One exception was a paper published by Waller et al. (2007) that compared and contrasted the results of a Poisson GWR model to results of a spatial random effects model by examining the spatially heterogeneous effects of illegal drug arrests and alcohol distribution on violent crime rates in 439 Houston, Texas census tracts. They concluded that while both models provided similar model estimates, the Poisson GWR model was less robust in providing an inferential basis for analysis of spatially referenced data. They also identified that the impact of covariate collinearity and parameter correlation on these two models are unknown, thus warranting additional methodological investigation. Although the GWR method has been underutilized in the study of injury, increased use of this technique in the future will undoubtedly add a unique perspective to the field of injury research.

References

- AAA Foundation for Traffic Safety. (2007). Improving traffic safety culture in the United States: the journey forward. <http://www.aaafoundation.org/pdf/SafetyCultureReport.pdf>.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Brunsdon, C., Fotheringham, S., & Charlton, M. (1996). Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis*, 28(4), 281–298.
- Colin, R. B. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364–366.
- Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research*, 40(5), 341–351.
- Fotheringham, S. A., Brunsdon, C., & Charlton, M. 1 edition (October 21, 2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Chichester: Wiley. ISBN-10: 0471496162.
- Graif, C., & Sampson, R. J. (2009). Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates. *Homicide Studies*, 13(3), 242–260.
- Guagliardo, M. F. (2004). Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics*, 3, 3.
- Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis and Prevention*, 42(2), 676–688.
- Li, G. (2004). Biological fallacy. *Academic Emergency Medicine*, 11(1), 119–120.
- Li, G., Grabowski, J. G., McCarthy, M. L., & Kelen, G. D. (2003). Neighborhood characteristics and emergency department utilization. *Academic Emergency Medicine*, 10(8), 853–859.
- Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, 24(17), 2695–2717.
- Penchansky, R., & Thomas, J. W. (1981). The concept of access. *Medical Care*, 19(2), 127–140.
- Pratt, J., & Schlaifer, R. (1984). On the nature and discovery of structure. *Journal of the American Statistical Association*, 79, 9–21.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B*, 13, 238–241.
- Snow, J. (1854). *On the mode of communication of cholera* (2nd ed.). London: Churchill Livingstone.
- Sui, D. Z. (2004). Tobler's first law of geography: a big idea for a small world? *Annals of the Association of American Geographers*, 94(2), 269–277.
- Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), 234–240.
- Waller, L. A., Zhu, L., Gotway, C. A., Gorman, D. M., & Gruenewald, P. J. (2007). Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21(5), 573–588.

Chapter 26

Social Network Analysis

Paul D. Juarez and Lorien Jasny

Introduction to Social Networks and Public Health

In public health, much of what is studied is inherently relational: disease transmission, peer influence on risky behavior, diffusion of information through coalitions, etc. Social network analysis is a paradigm, grounded in empirical data, mathematical and computational models, and graphs and matrices that offer public health researchers a new set of theoretical assumptions, methods, nomenclature, and software programs for examining complex public health issues. It provides researchers with a set of tools for describing and testing hypotheses about the ways in which social structures and relations between people affect health behaviors, knowledge, attitudes, and outcomes and the conceptualization of innovative health promotion interventions. Specifically, social network analysis provides the opportunity to advance our understanding of public health in three distinct ways: (1) It provides structural descriptions of transmission networks, (2) it enables investigators to use simulations and models that provide more accurate predictions of the course of health issues, and (3) it is being used to develop and evaluate health promotion interventions based on a relational approach.

Social Networks as a Methodological Tool

A social network approach, whether applied to individuals, organizations, or other units of analysis, includes how networks operate, the types and levels of relations between actors in a network, and how information and resources flow through the network. Individuals often operate in different social networks and/or occupy different positions within each network. Positions in a network are defined by the number and types of links individuals have to others and the probability of different types of members adopting a particular behavior. Social network analysis differs from traditional social research methods, which are more likely to focus on the relationships between personal attributes, behaviors, and outcomes.

P.D. Juarez, PhD (✉)

Department of Family and Community Medicine, Meharry Medical College, Room 3015,
3rd Floor Old Hospital, 1005 Dr. D.B. Todd Boulevard, Nashville, TN, USA
e-mail: pjuarez@mmc.edu

L. Jasny, MA, PhD

Department of Sociology, University of California at Irvine, Irvine, CA, USA
e-mail: ljasny@uci.edu

The primary objective of network analysis is to measure and represent structural relationships among nodes (e.g., actors, organizations) in a network, explain why they occur, and analyze their effects on outcomes. Unlike traditional social and behavioral research methods which focus on attributes of the individual, the unit of analysis in social network analysis is found in attributes of the *relations* or *ties* between an “ego” and its “alters” such as family members, friends, or organizations in the network. Better understanding of the type and level of support and resources provided by different types of alters offers us new ways for conceptualizing strategies and interventions for promoting physical health and psychological (mental and emotional) well-being through changes in knowledge, attitudes, and behaviors.

Types of Networks

Social network theory can be used to study social relationships among networks of individuals at a microlevel and among organizations and communities at a macrolevel. In public health, social network analysis has been used to study a number of different types of networks: *disease transmission networks*, *HIV/AIDS* (Auerbach et al. 1984; Klovdahl 1985), *pneumonia* (Meyers et al. 2003), and other *STDs* (Wylie and Jolly 2001; De et al. 2004); *social contagion (teen smoking)* (Valente et al. 2005; Christakis and Fowler 2008), *breast cancer screening* (Allen et al. 1999; Keating et al. 2010), *mental health* (Fowler and Christakis 2008; Rosenquist et al. 2011), *alcohol abuse* (Ormerod and Wiltshire 2009; Rosenquist et al. 2010), *substance abuse* (Bauman et al. 1994; Valente et al. 2004), *delinquency/intentional injuries* (Dijkstra et al. 2010; Radil et al. 2010), *bullying* (Moultapa et al. 2004), and *obesity* (Bahr et al. 2009; Mulvaney-Day and Womack 2009); *diffusion of information* (Valente 2005; Wandersman et al. 2008), *social support* (Kawachi and Berkman 2000; Srinivasan et al. 2003), *social capital* (Lin 2001), and *organizational networks* (Burt 2000; Borgatti and Foster 2003).

Describing a Network

A social network can represent many different types of interactions between people, including communication, social support, intimate relationships, etc. Social network analysis provides methods for translating social and behavioral concepts into formal definitions and empirical expressions of the structural properties (*nodes* and *ties*) of social relationships. In social network analysis, relations are not the properties of individuals or nodes but of network attributes that connect pairs of nodes into larger relational systems.

The principal types of data used in social network analysis are “attribute” and “relational”: *Attribute* data are measured as values of both actors (income, occupation, education, etc.) and networks (measures of centrality, such as degree, closeness, betweenness, etc.), while *relational* data are properties of the contacts, ties, and connections which relate one node to another. In general, social network analysis is more concerned with attributes of the relationships between actors than characteristics of the actors themselves. The topology of a network is key to understanding the nature of relationships between actors in a network.

Types of Networks

There are two major types of social networks: “*complete networks*” and “*egocentric networks*.” The major distinction between *complete networks* and *egocentric networks* is in how data are collected.

Complete network analysis is conducted when all of the relationships among all respondents in a network are known, such as all employees of a given company or students in a classroom. This type of data collection uses a defined population and shows links between the actor and each known network member. Egocentric network methods, in contrast, involve assessing the attributes of relations in a personal network (size, diversity, etc.) or relating attributes of an ego with attributes of its alters (*homophily*). *Ego network analysis* is usually conducted when the composition of a participant's network is unknown and relies on the ego to identify the participants using a *name-generation* method. Egocentric network methods usually yield considerably less information about network structure but are less costly than complete network methods and allow for easier generalization from the sample to some larger population.

Attributes of a Network

Actors and *relations* are the primary elements of a social network. *Actors* can be individuals, organizations, or even countries, while *relations* are the specific kinds of contacts, connections, and ties that occur between them. An actor (e.g., ego) derives different types of social support and resources (e.g., emotional support, financial support, instrumental support) from its *alters* (e.g., a spouse, relative, colleague, friend, or neighbor). The *relation* or a *tie* between an *ego* and its *alters* is derived from the social attributes of both participants. When one person gives social support to a second person, there are two relations at play: giving support and receiving support. Each *tie* potentially provides the ego with direct access not only to its alters but also to all network members to whom their alters are connected. Social network measures of ties include *direction* (extent to which the tie is from one actor to another), *indirect* link (the path between two actors is mediated by one or more others), *frequency* (how often the link occurs), *stability* (existence of the link over time), *strength* (amount of time or intensity of tie), *symmetry* (extent to which the relationship is bidirectional), and *multiplexity* (extent to which two actors are linked together by more than one relationship).

Individual Attributes

Investigators typically are interested in correlating relational attributes of an actor with others in the network. Social network measures of individual actors in the network include *degree* (number of direct links with other actors), *in-degree* (number of incoming links from other actors), *out-degree* (number of outgoing links to other actors), *closeness* (extent to which an actor can easily reach all other actors in the network), *betweenness* (extent to which an actor falls in between any two other actors), *centrality* (extent to which an actor is central to the network), and *prestige* (the extent to which actors are the object rather than the source of relations).

Network Attributes

Knowledge of a network's attributes can help us understand how the various processes through which information, influence, social support, disease, etc., spread from one or more nodes to other nodes in the network and the outcomes of those processes (e.g., new behaviors adopted). Understanding the structure of a network is useful for aggregating microprocesses into macro outcomes.

Network analysts have developed their own specialized language to describe the structure and contents of networks they observe. *Content*, *direction*, and tie *strength* are three of the most commonly

studied attributes of social networks: *Content* examines the nature of resource exchange, *direction* is concerned with the flow of resources, and *strength* relates to intensity of relationships. Other social network components that have been frequently studied include *centrality* of networks, *density* of networks, *cliques*, *positions*, *roles*, and *clusters* of relationships. Similarly, a vocabulary has evolved to describe relations between actors including *directed* or *undirected*, *shared*, or *asymmetrical*. All of these terms typically utilize mathematical models and graphical constructs to examine information exchange and diffusion and to display results as network diagrams. Other important network concepts are described by Durland and Fredericks (2005) and will not be reviewed here.

Data Analyses

Social network analysis is more of a branch of “mathematical” sociology than of “statistical or quantitative analysis.” Mathematical approaches to network analysis treat observations as a population of interest, not as a “sample.” Because network observations are almost always nonindependent, conventional inferential statistics is of limited value for analyzing network data. In recent years, nonlinear models have been developed that allow for the systematic analysis of stochastic processes affecting the rates and patterns of relations over time. Specialized social network analysis software has been developed to analyze unique features of matrices.

Levels of Analyses

There are four distinct conceptual *levels of analysis*: egocentric, dyadic, triadic, and complete networks.

Egocentric Network

The first level of analysis is the *egocentric network* (with and without alter connections), which includes the ego, all of its alters, and any direct relations among the alters. Egocentric networks are used to find out such things as the number of connection nodes and the extent to which these nodes represent close-knit groups. While some properties, such as overall network density, can be reasonably estimated with egocentric data (e.g., prevalence of reciprocal ties, cliques), many network properties such as distance, centrality, and various kinds of positional equivalence cannot.

Dyadic Network

The second level of analysis is the *dyadic network* which addresses the ties that exist between two actors in a network and characteristics of the ties between them (e.g., *intensity*, *duration*, and *strength of the relation*).

Triadic Network

The third level of network analysis is *triadic relations* and includes all possible combinations of relations among any three actors in a network.

Complete Network

The fourth level of network analysis is of the *complete network*. Complete network methods require information to be collected about each actor's ties with all other actors. Complete network methods use a census approach to gathering information about all ties in a population of actors – rather than a sample. Because information is collected about ties between all pairs or dyads, complete network data provide a complete picture of relations in the population. Most of the special approaches and methods of network analysis were developed to be used with complete network data. Complete network data are necessary to properly define and measure many of the structural concepts of network analysis (e.g., betweenness).

Statistical and descriptive uses of network analysis require concepts and analytic procedures that are different from traditional statistics and data analysis. Stochastic modeling of a social network improves on existing methods by allowing statistical uncertainty in the social space to be quantified and graphically represented. It recognizes that networks are dynamic and change over time; where existing actors attain new relations, new actors join the network, while others may leave. Stochastic modeling can be generalized to allow for multiple relationships, ties with varying strengths (using generalized linear models), and time-varying relations (by modeling the latent positions as stochastic processes). To date, however, only a handful of public health studies have utilized stochastic (Morris and Kretzschmar 1995) or longitudinal methods (Valente 1995, 2005).

Youth Violence as Public Health Issue

Historically, youth violence has been thought of as a criminal justice or sociological problem. Interpersonal violence, however, also is a major public health issue as evidenced by a wide array of statistics: Homicide is the second leading cause of death for young people between the ages of 10 and 24. In 2007, 5,764 young people aged 10–24 nationally were murdered – an average of 16 each day. Over 656,000 violence-related injuries in young people ages 10–24 were treated in emergency rooms in 2008 (CDC 2010). In a 2005 nationwide survey, 36% of high school students reported being in a physical fight during the past 12 months. An estimated 30% of kids between 6th and 10th grade report being involved in bullying. New approaches for understanding and preventing youth violence are sorely needed.

Adolescent Development

The onset of puberty is marked by a sequence of changes to the physical, psychosexual, and social growth and development of children and hails the onset of adolescence. Adolescent development is frequently broken down into three levels: early adolescence (ages 12–14), mid-adolescence (ages 15–16), and late adolescence (ages 17–21). Each stage of development contains unique challenges to teens relative to their physical, psychological, social, and spiritual development.

Marshall and Tanner (1969, 1970) describe adolescence as five stages of normal pubertal maturation (i.e., Tanner stages) consisting of predictable changes in secondary sexual characteristics that all girls and boys go through. In contrast, Erikson (1959) identified adolescence by the set of major developmental tasks that face teens: (1) personal identity formation, (2) becoming independent, (3) achieving a sense of competency, (4) establishing social status, (5) experiencing intimacy, and (6) determining sexual identity. Bronfenbrenner (1979), meanwhile, described child and adolescent development as being shaped by the context of roles, norms, and rules of four types of nested

ecological systems: the *microsystem* (family, classroom), *mesosystem* (interaction of two microsystems), *exosystem* (external environment), and *macrosystem* (sociocultural context).

During their teenage years, as adolescents begin to develop their sense of personal identity and autonomy and formulate their own principles of right and wrong, it is not uncommon for them to see themselves one way when they are with parents and teachers and another way when they are with their peers. Nor is it uncommon for adolescents both to see themselves and to act differently within various peer groups. While adults remain essential to their continuing development as caregivers, role models, educators, and mentors, the frame of reference of teens expands during adolescence, from family to peers and other adults with whom they increasingly have more contact. The successful transition into adult roles (e.g., work, relationships, parenting) appears to reduce involvement in violence and other behaviors that increase risk for poor health and social outcomes.

Adolescents frequently look to different types of people in their lives for different types of social support. For most adolescents, family support is the most important element in their lives (Schwarzer and Leppin 1991; Taylor 2007). Inadequate support and guidance from their parents increases the probability of poor academic performance, inadequate interpersonal skills, and engagement in risk-taking behaviors (Grunbaum et al. 2004). While parents tend to provide more emotional and instrumental support, other caring adults, such as a teacher, coach, and counselor, also can be significant providers of social support to teens, helping them cope with many issues and choices they face during this transitional period in their lives as they mature and move on toward adulthood. In the absence of support from parents and other caring adults, teens often turn to peers for information, emotional support, and guidance. However, reliance on peer networks can be unhealthy when they reinforce behaviors that are, in themselves, harmful. While peer support is an essential factor in an adolescent's social network, if peers support harmful behaviors, an adolescent may be more likely to engage in such harmful behaviors.

Risk and Protective Factors

Individual and family dysfunction has been commonly identified as risk factors for youth violence (Zingraff et al. 1993; Farrington 1989; Lipsey and Derzon 1998). Community-level risk factors for youth violence that have been identified include weak social controls/social bonds (Hirschi 1977; Hawkins et al. 1998), lack of social capital (Sampson and Lauritsen 1994), community deterioration or disorganization, and low levels of neighborhood and organizational collective efficacy (Sampson et al. 1997; Perkins and Long 2002; Perkins et al. 1996).

There also is an abundant literature demonstrating a positive association between the risk behavior of adolescents and that of their peers. Adolescents appear to be particularly susceptible to peer influence during mid-adolescence, including behavioral constraints that pull them toward or away from delinquent behavior. Their position within their peer networks provides different opportunities for peer interaction, resulting in varying exposure to delinquent behaviors, communication of delinquent norms, access to information on delinquency opportunities, and opportunities for rewards or deterrents for participation in delinquent behaviors.

Although prior research establishes that adolescents are likely to behave in a manner consistent with their friends, it has yet to incorporate the network structure of friendship relations into empirical models. Social network analysis provides an alternative framework to social control and differential association theory, the two dominant theories for studying delinquent behavior. While differential association theory focuses on the effects of peer networks and social control theory on adolescents' attachment to friends, neither has considered characteristics of the networks themselves (Krohn 1986). To date, there have only been a handful of studies that have applied a network perspective to the study of delinquency (Krohn et al. 1986; Haynie 2001; Sarnecki 1990, 2001; Baerveldt and Snijders 1994; Snijders and Baerveldt 2003; Tita et al. 2005; Radil et al. 2010).

A network perspective assumes that the structure of a network has consequences for its individual members and for the network as a whole, over and above effects of characteristics and behaviors of the individuals involved (Klovdahl 1985, p. 1204). Network methods assume that patterns of friendship ties structure the flow of information, social norms, and social support and potentially provide linkages for the transmission of delinquent behavior (Ennett et al. 1999). By providing a methodologically rigorous approach that allows for the analysis of identifiable structural properties of peer/friendship networks, it is uniquely suited to addressing and measuring the behaviors of youth (Ennett and Bauman 1996).

Preliminary Findings

The current study builds upon a previous *Community Asset Mapping study* (Juarez et al. 2009) that was undertaken to map the relationship between social capital and acts of youth violence. This phase of the research was designed as an exploratory study employing qualitative methods to explore whether and how youth perceived community resources as possible protective factors in preventing youth violence. The youth were asked to identify “places” that they thought of as “safe places,” “a place they would go to for mentoring or advice,” or “somewhere they would go to find a job.” The responses were identified as three types of social capital. The procedures for identifying community assets were developed with input from a community coalition called the *Nashville Community Coalition for Youth Safety (NCCYS)* which was established in 2006 to serve as the steering committee for the Nashville Urban Partnership Academic Center of Excellence (NUPACE), a research center funded by the NCIPC/CDC to engage communities in the prevention of youth violence (see <http://nupace.org/coalition.html>). The NCCYS was established to provide input on research projects funded by NUPACE and to participate in the development of new research proposals, provide leadership in youth violence prevention planning and surveillance activities, coordinate and implement community-level, youth violence prevention interventions, and disseminate information on research findings and best practices to the broader community, including policymakers. In a 2-day strategic planning session undertaken by the NCCYS (which included youth participation), youth violence prevention was operationalized as (1) safety/safe places, (2) caring adults/mentors, and (3) job and career training/work opportunities. It was unanimously agreed upon by Coalition members that its focus would be to increase the number of safety/safe places, caring adults/mentors, and job and career training/work opportunities for the youth. The operational definition of youth violence prevention adopted by the NCCYS was used for this study.

This initial stage of the study employed structured focus groups and GIS/mapping and was carried out in two parts. In the first stage, 105 15–18-year-old youth were recruited at community agencies to participate in structured focus groups. The youth were asked to assist with the identification of the types of assets in their communities relating to safe places/safety, caring adults/mentors, and job and career training/work opportunities they saw as important for promoting safety. Ten focus groups of 8–12 youth each were conducted at youth-serving organizations in north, south, east, and southeast Nashville (the four areas with the highest rates of youth homicides in the city/county according to police data). Trained research assistants used a prepared script organized as introductory, transition, key, closing, and summary questions to guide the focus group discussions in each of the three areas of interest: safe places/safety, caring adults/mentors, and job and career training/work opportunities. Focus groups were conducted in English at community-based organization and took 60–90 min to complete. Proceedings of focus groups were recorded, transcribed, and analyzed. The Meharry IRB determined this project as exempt as no sensitive information was collected and information was recorded in such a manner that subjects could not be identified from responses. Qualitative analysis of focus group transcriptions, field notes, and debriefing discussions obtained by the field team were conducted after all focus groups had been conducted. NUDIST software and manual

qualitative analysis methods were used to identify and code common themes that emerged from focus group discussions. Specific types of safety/safe places, caring adults/mentors, and job and career training/work opportunities identified by youth were classified into broader categories, such as grocery stores and fast-food restaurants (for employment opportunities). A final list of community assets was reviewed and finalized with youth input.

This list of community assets was used by youth to identify safety/safe places, caring adults/mentors, and job and career training/work opportunities when they conducted the street-by-street walkabouts of four targeted neighborhoods during the second phase of the study. The youth were trained to use preprogrammed GPS devices to capture data and geocode the locations of identified community assets. Walkabout teams were comprised of two youth and one adult from the community organization who were familiar with the neighborhood. A 3×3 analytic framework was developed to guide the identification of community assets. The first dimension was comprised of safety/safe places, caring adults/mentors, and job and career training/work opportunities. The second dimension was comprised of people, places, and opportunities. When a community asset was identified, youth completed a log of the type of asset, name, location, and latitude/longitude coordinates.

A list of homicide victims of youth 14–24 years of age occurring in Nashville/Davidson County in calendar years 2005, 2006, and 2007 was obtained from Metro Nashville Police Department. Data were merged into a database, mapped, and analyzed using ArcView 9.2 (see <http://www.imnashville.com/home/youth-violence/surveillance-data/>).

Results found that the youth were much more likely to identify a person instead of an agency or place: I would go to my auntie, or an uncle, or a teacher. Based on youth responses, these investigators decided to change our approach from trying to map (using GIS) the relationship between social capital and youth violence to exploring the social networks of youth as protective factors in preventing youth violence.

Methodology

This second phase of the study was designed to identify the relationship between egocentric networks of youth and their risk for youth violence. Three distinct egocentric networks were identified for each youth. These included persons that the youth would turn to for assistance or information around (1) safety, (2) social/emotional support, and (3) job training or work opportunities. For the purpose of this study, nodes were defined as individual youth, and ties were defined as the relationships between youth and other persons within and outside of their immediate neighborhood in each of the three areas.

This study expands on prior research in two important ways. First, peer networks were defined more rigorously as adolescents' egocentric social support networks and are measured with network data on friendship nominations provided by adolescents. Second, it expanded upon the conceptualization of social capital as community resources to one that clearly recognized the importance of individuals in their lives whom they turn to for information, advice, or comfort. In this phase of the study, social network analyses were conducted to assess attributes of network structures and relations of youth and those persons in their lives whom they turn to (or would turn to) when concerned about "safety," were in need of social/emotional support, or sought information or guidance about job training or work opportunities.

Social network analysis allows us to explore the types and degree of social capital available to youth who live in areas with high rates of youth violence and, in particular, to assess who and where they are most likely to turn to for safety, mentoring, and employment opportunities. Social network analysis is concerned with relational patterns among individuals and their social networks and, thus, can be used to determine the level and types of social capital utilized by individuals (Borgatti et al. 1998).

The ties in their networks were assessed on a number of dimensions including the degree of closeness of connections, importance, cohesion, and centrality in a network.

Research Questions

(1) What are the characteristics of social networks that youth access to help them with their needs around safety, mentoring/caring adults, and job training and work opportunities? (2) Do social networks protect youth against interpersonal violence? (3) What attributes of a youth's social network are associated with risk for interpersonal violence?

Procedures

One hundred high school students participated in ten small groups and responded to a questionnaire that measured incidents of violence experienced in the previous year, their school performance, and information on exposure to individual, family, school, and community risk and protective factors. Youth participated in a systematic process in which they developed a geographically anchored map of their social networks. They identified persons to whom they would turn for personal safety, mentorship/personal support, and job training/work opportunities, three factors youth had previously identified as important for preventing violence.

To answer these questions, the youth were asked to complete sociograms (pencil and paper) depicting persons they turn to relative to safety, social support, and employment and/or job training. *Sociograms* were used to provide a graphic representation of the links that a youth has with other persons in his or her social network. The sociogram was used to diagram the structure and attributes of network links and to assess patterns of relationships, subgroup organization types, and the nature, types, and strength of interrelationships. Individual sociograms were completed by youth to depict whom or where they turn to for safety, mentoring/caring adults, and job training or employment opportunities.

Students were recruited through community agencies that are members of the NCCYS and participated in the previous study on mapping community assets. Written authorization from the agency director was obtained. A trained research assistant met with agency staff to explain the project and to recruit youth to participate in the study. For youth who were eligible and agreed to participate, a parental consent form was sent home for signature and returned to the counselor. Parental consents and student assents were collected for all students who participated in the study.

The youth were administered three instruments in a small group of 8–12 youth with two trained research assistants. Completion of the three instruments took 1–1½h. Students were administered a 20-item *Personal Safety Survey* to identify demographic information and exposures to interpersonal violence and other risk and protective factors. Students were then administered a second survey, *Student Social Network Survey*, where they were asked to identify persons whom they would turn to for personal safety, caring adults/mentors, and job training/work opportunities, three factors the youth had previously identified as important for preventing violence. The youth were asked to systematically identify persons in each of the three areas. Categories of persons for each area included: (1) immediate family, (2) other relatives, (3) school personnel, (4) minister or other church official, (5) community agency staff, (6) police/security, (7) friend, and (8) online friend. The youth were asked to write the first name and initials of all persons who were identified, the category of relationship associated with each identified person, demographic characteristics of each, and the neighborhood in which they live and to respond to each of nine items on the Social Network Survey. The youth then were asked to identify the type of personal contact they most frequently had with each of the persons they identified in their Social Network Survey.

Each youth was then given large pieces of blank butcher paper with three concentric circles (like an archery target) to identify who they would most likely relate to around issues of personal safety, social/emotional support, and job training/employment opportunities. They were asked to put a sticker representing themselves in the center of the paper and to place a red dot for everyone in their social network they identified as a family member in the appropriate concentric circle, a blue dot for everyone they identified as a friend, and a yellow dot for all other persons with the initials of each of the people identified on the Social Network Survey on the appropriate dot. They were asked to place stickers in one of the three concentric circles with persons who they were most likely to contact in the circle closest to the center and so on. Students were then asked to draw a green line between themselves and each of the persons they would turn to for safety (questions 1–3), an orange line to each of the persons they would turn to for job training/work opportunities information, and a purple line between themselves and each of the persons they identified for social/emotional support. There could be three different lines to each ego and their alters. Finally, the youth were asked to draw a line between the persons in their network who knew each other and to write how well they know each other on the line using a scale from 1 to 3, with one being a little, two being pretty well, and three being very well. Students were given a \$20 Walmart gift certificate for participating in the study.

Instrumentation

Individual sociograms for each youth were generated to assess their access to social capital in their neighborhoods and community relative to safe places, mentoring, and job training/work opportunities. Sociograms used distinct shapes and colors indicating the type of resource/opportunity it provides (safe place, mentoring, or job training/work opportunity).

Analyses

This study sought to test a methodology for examining the content, direction, and tie strength of youth of persons they relate to around concerns about personal safety, social support, and jobs and careers and the relationship between attributes of social networks and risk for youth violence. Data were analyzed using descriptive statistics, correlations, regression analysis, and social network analysis to analyze spatial, mathematical, and substantive dimensions of the social structures and ties formed between youth and persons whom they relate to that serve as protective factors for exposure to interpersonal violence either as a victim or offender.

Often when we use network data, we want to examine the relationship between node-level indices of network structure (degree, betweenness, etc.) or graph-level indices (size, centralization, etc.) and other covariates. In this pilot study, for example, we could hypothesize that those who engage in criminal and antisocial behavior do so because they have smaller or larger friendship networks than those who do not. When comparing individuals tied in one network, we cannot use traditional statistical methods to answer this question because the variables of network structure are not independent. For example, my level of centrality is not independent of those who are connected to me. On the other hand, egonet data can be analyzed with more standard statistical methods than with complete networks because when sampled appropriately from a larger population, we can assume that the network statistics drawn from each egonet are independent of each other. Thus, we could use linear regression or other methods that require these assumptions to predict network variables. In both cases, complete and ego network data, variables of network structure can be included as independent variables.

In the correlation table (see Table 26.1), the significance levels are determined through a nonparametric permutation test. This is more appropriate for a number of reasons: Given the sampling, it is quite possible that many of these students know each other and are included in each others' networks. Additionally, the skewness of the distributions and the nonlinear nature of many of the scales is the

Table 26.1 Correlation of network measures with ego attributes

	Alter variables (same for each relationship network)										
	Average relationship strength					Employment network		Social support network		Safety network	
	Network size	Family size	Friends size	Family %	Average relationship strength	Average degree	Ego betweenness	Average degree	Ego betweenness	Average degree	Ego betweenness
Age	-0.13	-0.2	0	-0.13	0.05	0.04	-0.02	-0.06	0.15	0.11	0
Grade	0.09	-0.21	0.21*	-0.17	0.12	0.03	0.14	-0.13	0.07	0.07	-0.1
Gender	0.03	0.05	-0.01	0.04	0.02	0.09	0.02	0.11	-0.01	0.13	-0.09
Physical fight	-0.07	-0.08	0	0	-0.01	-0.04	-0.06	0.05	-0.04	-0.04	-0.02
Threatened or injured	0.04	-0.14	0.11	-0.13	-0.12	-0.07	-0.1	-0.03	0.23*	-0.03	0.02
Unsafe at school	0.34*	-0.02	0.21	-0.15	0.02	-0.14	-0.04	-0.15	-0.08	-0.05	-0.06
Physical hurt by boyfriend/girlfriend	0.13	0.13	0.05	0.01	0.03	0.07	0.1	-0.09	0.05	0.09	-0.1
Carried weapon to school	-0.1	-0.01	-0.11	0.08	-0.18	0.15	0.11	0.12	-0.02	0.07	0.05
Uses marijuana	-0.03	0	-0.03	-0.01	-0.04	-0.02	0.01	-0.02	-0.11	0.07	0.02
Drinks alcohol	-0.01	0.07	-0.06	0.03	-0.16	0.04	0.01	0.18	-0.11	0.28*	-0.05
Incident reports at school	-0.06	0.03	-0.06	0.08	0.01	0.07	-0.03	0.09	-0.05	-0.13	-0.06
Disciplined by parent	-0.03	0.02	-0.06	0.07	-0.01	-0.05	-0.11	-0.11	-0.05	0.04	-0.11
School performance	-0.13	-0.04	-0.11	0.01	0.02	-0.15	-0.18	-0.13	-0.08	0.03	-0.09
Truant	-0.07	0.03	-0.07	0.09	-0.1	-0.12	-0.09	0.09	-0.11	0.03	-0.09
Holds driver's license permit	-0.02	0.21*	-0.13	0.24*	-0.04	-0.17	-0.06	-0.05	-0.12	-0.17	0.04

* $p < 0.05$

Table 26.2 Regression models of feeling unsafe at school

		Model 1	Model 2	Model 3
Ego covariates	Intercept	-0.074 (0.082)	-0.199 (0.113)	-0.166 (0.114)
	Age	0.040 (0.041)	0.065 (0.039)	0.084 (0.039)*
	Age ²	-0.005 (0.005)	0.008 (0.005)	-0.010 (0.005)*
	Gender (female)	0.030 (0.026)	0.030 (0.025)	0.035 (0.025)
	Race (black)	0.012 (0.030)	0.027 (0.029)	-0.004 (0.032)
	Race (Asian)	0.004 (0.087)	0.034 (0.084)	0.015 (0.085)
	Race (other)	0.018 (0.090)	0.076 (0.086)	0.026 (0.089)
Network covariates	Size		0.009 (0.003)**	0.012 (0.003)**
	Family size		-0.008 (0.006)	-0.006 (0.007)
Network structure covariates	Average relationship strength		0.005 (0.034)	0.014 (0.039)
	Average degree: employment			-0.008 (0.008)
	Ego betweenness: employment			0.028 (0.023)
	Average degree: social support			-0.006 (0.007)
	Ego betweenness: social support			-0.18 (0.009)*
	Average degree: safety			-0.001 (0.006)
	Ego betweenness: safety			0.002 (0.012)
<i>R</i> ²		0.03	0.18	0.28
BIC		-94.68	-95.00	-79.00

Significance codes: * $p < 0.05$, ** $p < 0.01$

difference between getting into one fight (coded as 1) and getting into two or three fights (coded as 2), the same as the difference between getting into two or three fights and getting into four or five fights (coded as 3). This is not a new problem in statistics by any means, but it is interesting to point out that many of the approaches that handle these typical problems can adequately address the dependence created by networks for simple analyses.

Further, network variables can be included in the set of independent variables in a regression equation without nullifying the results. This is not true of the dependent variable, which means that regression can only be used when we wish to predict some independent external covariate with network structure. In our case, this would mean that teens are more likely to engage in certain behaviors because of their network properties, and not that their network properties are also influenced by their behavior. Of course, frequently this is exactly what we believe to be going on, hence the large amount of work by methodologists in models of network autocorrelation and exponential random graph models that attempt to appropriately test these kinds of ideas.

An example of a regression including variables of network structure is presented in Table 26.2. We have taken one of the significant relationships from the correlation table, that of feeling unsafe at school and the size of the ego network, and have explored it further using standard OLS regression. The first model looks at the relationship between feeling unsafe at school with the ego's attributes. The second includes a description of network size, the number of family members included in the network, and the average closeness to each individual the respondent reported. The final model incorporates two measures of network structure, average degree and ego betweenness, for each of the three relationships measured.

The independent variables are broken into three groups. Ego covariates are those properties solely of the ego: age (we have also included age squared), gender (male is the reference category), and race (white is the reference category). Network covariates are the size of the network, the number of the family members in the network, and the average closeness the ego reported feeling to each member. Covariates of network structure are the average degree and ego betweenness for each of the three relationship types (employment, social support, and safety). The average degree is a measure of network density that is independent of network size. The ego betweenness is a measure of how central the ego perceives himself or herself to be in their own network (Everett and Borgatti 2005).

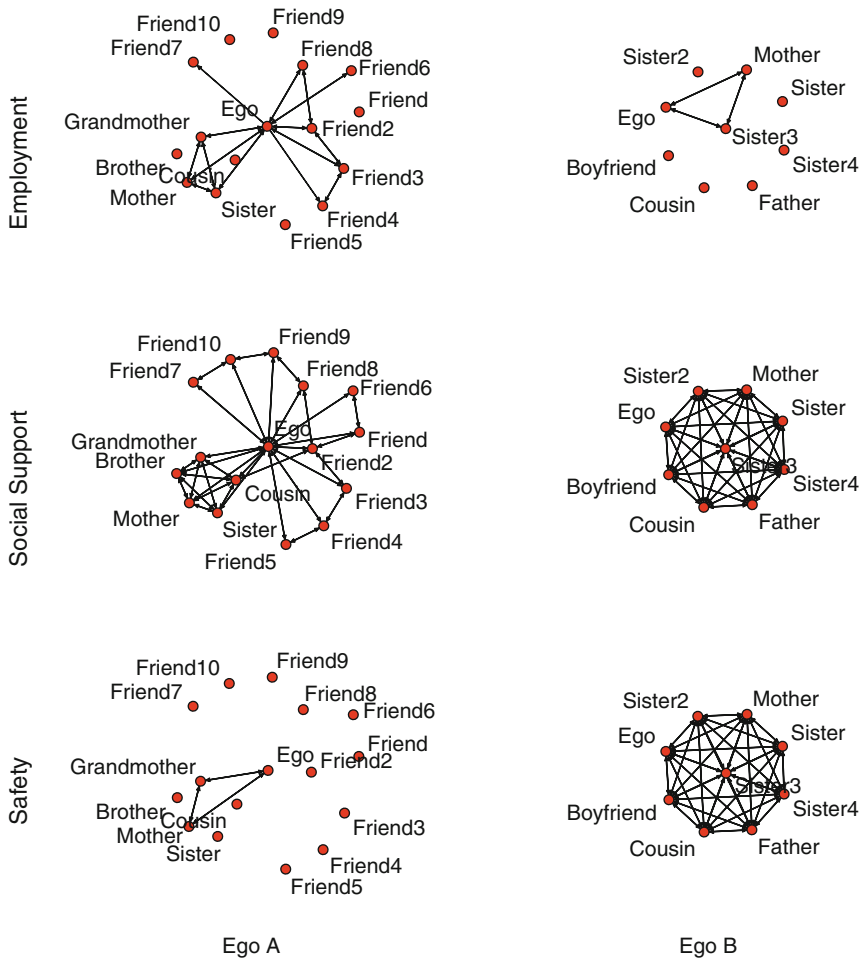


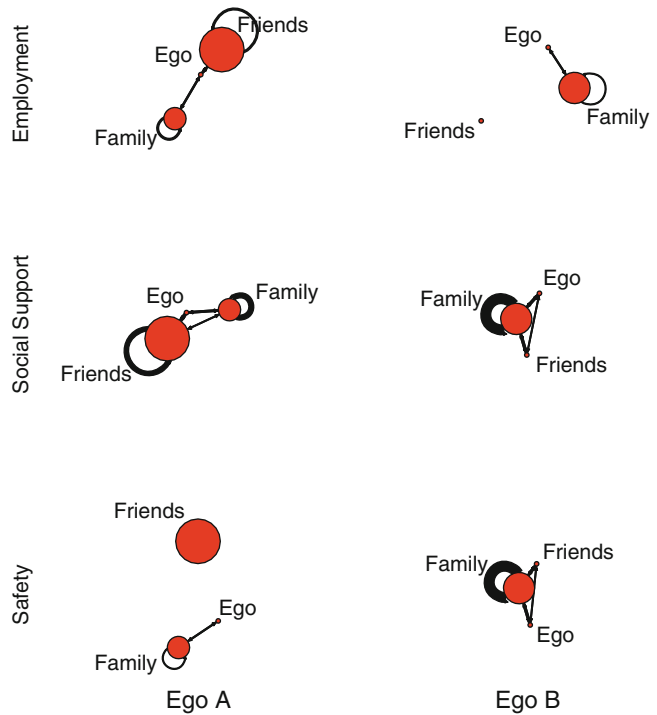
Fig. 26.1 Two ego networks with three relationships

In other words, this is inversely related to how often alters in the network have relationships with each other and thereby do not rely on the ego.

From the results, we see that feeling unsafe at school is poorly explained solely by ego attributes (model 1). This explains almost no variance in the dependent variable (R^2 is 0.03). Model 2 includes the network covariates. Individuals who named more alters (network size) significantly predict feelings of being more unsafe at school. Although this explains more variance (R^2 is 0.18), the added covariates penalize the model, so the Bayesian information criterion (BIC) is lower (-94.68 to -95), but not by much. The third model includes the measures of network structure. We see that this model explains much more of the variance (R^2 is 0.28) to the extent that even though there are more covariates, the BIC is higher (-79), indicating that by this measure, model 3 is better. Under this model, age is positively correlated with feeling unsafe at school, but this relationship decreases as age increases since age squared is negatively correlated with feeling unsafe at school. Network size remains highly positively correlated, and ego betweenness is significantly negatively correlated. While this was not found in the raw correlation table, here it indicates that, given all the other covariates, individuals whose alters have more relationships with each other (resulting in lower ego betweenness) are correlated with a decrease in ego’s feeling unsafe at school.

In Fig. 26.1, plots of two different individual’s ego networks are shown. For simplicity, each relationship (employment, social support, and safety) is plotted on a different graph. An actor sends

Fig. 26.2 Blockmodels of ego networks by alter type



a tie to another if the respondents said they relied on support from the recipient of the tie. We have kept the position of each vertex the same across each plot, so we can more easily compare the three networks. With Ego A, we see a denser social support network and a sparser safety network. Additionally, Ego A has a larger network with many more friends than Ego B. Ego B has a sparse employment network consisting solely of a triad with ego's mother and third sister, but Ego B's social support and safety networks are complete cliques. This means that everyone in the network is tied to everyone else. Thus, the structure of A and B's networks is very different, and we hypothesize that this could have different effects on the behaviors of each ego.

In Fig. 26.2, a reduced blockmodel plot of the same networks as in Fig. 26.1 is presented. Here, friends and family have been collapsed into two respective groups. This allows us to see many different structural features about these groups that are easily overlooked in the full network. First, it is obvious how many more friends Ego A has named than Ego B, as the vertices are sized by the number of individuals in each group. Additionally, the edges are sized by the number of ties between members of each group, for example, from Ego's friends to Ego's family. The loops represent the number of ties within a group, for example, the number of family members who are tied to each other in the full network. Notice that Ego B's friends are not named in the employment support network and Ego A's friends are not named in the safety network. From this, we can more easily see that there are differences between the people Ego A and Ego B rely on. Also, we can see that in Ego A's safety network, and in Ego B's social support and safety network, friends and family are tied. This would unlikely happen in families where the parents did not like the friends of their child. Thus, this type of blockmodeling can show us interesting properties of the network by given alter attributes.

Discussion

Study results indicate that all of the network characteristics – density (average degree measures), centrality (ego betweenness), and popularity (network size) – may be strongly correlated with measures of risk for youth violence, either as risk or protective factors. While the power of ego networks is limited, the results of this study show that network effects may play an important role in understanding how the relations between youth and others in their social networks increase or decrease their risk for youth violence victimization and/or perpetration. What we have shown serves as a proof of concept; different measures of network structure are significantly correlated with increased participation in and suffering from harmful behavior. Understanding the mechanisms and causal relationships behind this finding is beyond the scope of current data but is a crucial project for the field as a whole.

By incorporating a social network approach, these efforts contribute to research and theory on the conditions under which differential associations are maximized. Overall, study findings present a picture more complex than that previously provided by social control theory and differential association theory alone and suggest that a network perspective can provide a coherent and powerful framework for addressing adolescent delinquency. If we apply some contextual data, such as the onset dates of symptoms, we also may be able to further increase our understanding of the transmission structures and relations associated with youth violence victimization and perpetration. For instance, to date, few previous studies have attempted to understand geospatial embeddedness of the structure and relations of social networks (Radil et al. 2010). In this study, we also obtained data on the geographic locations of persons in the networks of youth participants to identify the potential geospatial dimension of the different networks in which the youth participate. Visual observation of the sociograms that the youth completed for this study suggests that in addition to social attributes, geographic attributes of social networks also should be examined. Our findings suggest that introducing geographical locations into social network analysis may provide a good way not only to better understand the role that those locations play in the transmission of social contagion but also to identify potential bridges between networks.

Social network analysis provides a new paradigm for analyzing complex, dynamic, and longitudinal public health issues such as youth violence and for conceptualizing interventions that can be tailored to address the unique circumstances of different subpopulations. While the analyses of the attributes of egocentric networks are still limited, it is difficult to envision many scenarios under which data on the complete networks of youth might be gathered. In addition, new techniques and measures using stochastic and multidimensional scaling are continually being developed and may eventually increase our ability to analyze and visualize dynamic and longitudinal data that are not limited by the lack of independence found in social networks. This study attempted to examine the independent impact of the social structures and relations of social networks that teens engage with family, peers, and other adults, reflecting the real-life complexities that face teens and shape their behavior. Results suggest that while network analysis can provide important new insights into the relationship between single networks and youth violence, there also is a need to extend social network analysis to multilevel analysis that can better explain social networks within a geospatial and temporal context.

References

- Allen, J. D., Sorensen, G., Stoddard, A. M., Peterson, K. E., & Colditz, G. (1999). The relationship between social network characteristics and breast cancer screening practices among employed women. *Annals of Behavioral Medicine, 21*(3), 193–200.
- Auerbach, D. M., Darrow, W. W., Jaffe, H. W., & Curran, J. W. (1984). Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. *American Journal of Medicine, 76*(3), 487–497.

- Baerveldt, C., & Snijders, T. A. B. (1994). Influences on and from the segmentation of networks: hypotheses and tests. *Social Networks*, *16*, 213–232.
- Bahr, D. B., Browning, R. C., Wyatt, H. R., & Hill, J. O. (2009). Exploiting social networks to mitigate the obesity epidemic. *Obesity*, *17*(4), 723–728.
- Bauman, K., Ennett, E., & Susan, T. (1994). Peer influence on adolescent drug use. *American Psychologist*, *49*(9), 820–822.
- Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: a review and typology. *Journal of Management*, *29*(6), 991.
- Borgatti, S. P., Jones, C., & Everett, M. G. (1998). Network measures of social capital. *Connections*, *21*(2), 27–36.
- Bronfenbrenner, U. (1979). *The ecology of human development: experiments by nature and design*. Cambridge, MA: Harvard University Press.
- Burt, R. S. (2000). The network structure of social capital. In R. I. Sutton & B. M. Staw (Eds.), *Research in organizational behavior*. Greenwich: Jai.
- Centers for Disease Control & Prevention (CDC). (2010). Youth violence facts at a glance. <http://www.cdc.gov/injury/publications/index.html>. Accessed 2 Mar 2011.
- Christakis, N. A., & Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *The New England Journal of Medicine*, *358*, 2249–2258.
- De, P., Singh, A. E., Wong, T., Yacoub, W., & Jolly, A. M. (2004). Sexual network analysis of a gonorrhoea outbreak. *Sexually Transmitted Infections*, *80*, 280–285.
- Dijkstra, J. K., Lindenberg, S., Veenstra, R., Steglich, C. E. G., Isaacs, J., Card, N. A., et al. (2010). Influence and selection processes in weapon carrying during adolescence: the role of status, aggression, and vulnerability. *Criminology*, *48*, 187–220.
- Durland, M., & Fredericks, K. A. (2005). The use of network analysis in program evaluation: trends, techniques, and applications. In M. Durland & K. A. Fredericks (Eds.), *New directions for program evaluation* (issue 107). San Francisco, CA: Jossey Bass.
- Ennett, S. T., & Bauman, K. E. (1996). Adolescent social networks: school, demographic, and longitudinal considerations. *Journal of Adolescent Research*, *11*(2), 194–215.
- Ennett, S. T., Bailey, S. L., & Federman, E. B. (1999). Social network characteristics associated with risky behaviors among runaway and homeless youth. *Journal of Health and Social Behavior*, *40*, 63–78.
- Erikson, E. H. (1959). *Identity and the life cycle*. New York: International Universities Press.
- Everett, M. G. & Borgatti, S. P. (2005). Extending centrality. In: P. J. Carrington, J. Scott, & S. Wasserman, (Eds.), *Models and methods in social network analysis*. Cambridge, UK: Cambridge University Press.
- Farrington, D. P. (1989). Early predictors of adolescent aggression and adult violence. *Violence and Victims*, *4*, 79–100.
- Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ*, *337*, a2338.
- Grunbaum, J. A., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Lowry, R., et al. (2004). Youth risk behavior surveillance – United States, 2003. *MMWR Surveillance Summary*, *53*, 1–96.
- Hawkins, J. D., Farrington, D. P., Catalano, R. F., Hawkins, J. D., Farrington, D. P., & Catalano, R. F. (1998). Reducing violence through the schools. In D. S. Elliott, B. A. Hamburg, & K. R. Williams (Eds.), *Violence in American schools: a new perspective* (pp. 188–216). New York: Cambridge University Press.
- Haynie, D. L. (2001). Delinquent peers revisited: does network structure matter? *American Journal of Sociology*, *106*, 1013–1057.
- Hirschi, T. (1977). Causes and prevention of juvenile delinquency. *Sociological Inquiry*, *47*, 322–341.
- Juarez, P. D., Bess, K., Padgett, D., Samaniego, V., & Hill, B. (2009). *Social networks as protective factors in preventing youth violence*. Washington, DC: Society for Prevention Research.
- Kawachi, I., & Berkman, L. F. (2000). Social cohesion, social capital, and health. In L. F. Berkman & I. Kawachi (Eds.), *Social epidemiology* (pp. 174–190). New York: Oxford University Press.
- Keating, N. L., O'Malley, A. J., Murabito, J. M., Smith, K. P., & Christakis, N. A. (2010). Minimal social network effects evident in cancer screening behavior. *Cancer*. Early release. <http://dx.doi.org/10.1002/cncr.25849>.
- Klov Dahl, A. S. (1985). Social networks and the spread of infectious diseases: the AIDS example. *Social Science & Medicine*, *27*(11), 1203–1216.
- Krohn, M. (1986). The web of conformity: A network approach to the explanation of delinquent behavior. *Social Problems*, *33*(6), 81–93.
- Lin, N. (2001). *Social capital. A theory of social structure and action*. Cambridge, UK: Cambridge University Press.
- Lipsey, M. W., & Derzon, J. H. (1998). Predictors of violent or serious delinquency in adolescence and early adulthood. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: risk factors and successful interventions* (pp. 86–105). Thousand Oaks, CA: Sage.
- Marshall, W. A., & Tanner, J. M. (1969). Variations in pattern of pubertal changes in girls. *Archives of Disease in Childhood*, *44*(235), 291–303.

- Marshall, W. A., & Tanner, J. M. (1970). Variations in the pattern of pubertal changes in boys. *Archives of Disease in Childhood*, *45*(239), 13–23.
- Meyers, L. A., Newman, M. E. J., Martin, M., & Schrag, S. (2003). Applying network theory to epidemics: control measures for *Mycoplasma pneumoniae* outbreaks. *Emerging Infectious Diseases*, *9*(2), 204–210.
- Morris, M., & Kretzschmar, M. (1995). Concurrent partnerships and transmission dynamics in networks. *Social Networks*, *17*, 299–318.
- Moultapa, M., Valente, T., Gallaher, P., Rohrbach, L. A., & Unger, J. B. (2004). Social network predictors of bullying and victimization. *Adolescence*, *59*(154), 315–335.
- Mulvaney-Day, N., & Womack, C. A. (2009). Obesity, identity and community: leveraging social networks for behavior change in public health. *Public Health Ethics*, *2*(3), 250–260.
- Ormerod, P., & Wiltshire, G. (2009). ‘Binge’ drinking in the UK: a social network phenomenon. *Mind & Society*, *8*(2), 135–152.
- Perkins, D. D., & Long, D. A. (2002). Neighborhood sense of community and social capital: a multi-level analysis, 291–318. In A. Fisher, C. Sonn, & B. Bishop (Eds.), *Psychological sense of community: research, applications, and implications*. New York: Plenum.
- Perkins, D. D., Brown, B. B., & Taylor, R. B. (1996). The ecology of empowerment: predicting participation in community organizations. *Journal of Social Issues*, *52*, 85–110.
- Radil, S. M., Flint, C., & Tita, G. (2010). Spatializing social networks: using social network analysis to investigate geographies of gang rivalry, territoriality, and violence in Los Angeles. *Annals of the Association of American Geographers*, *100*(2), 307–326.
- Rosenquist, J. N., Murabito, J., Fowler, J. H., & Christakis, N. A. (2010). The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, *152*(7), 426–434.
- Rosenquist, J. N., Christakis, N. A., & Fowler, J. H. (2011). Social network determinants of depression. *Molecular Psychiatry*, *16*(3), 273–281. doi:10.1038/mp.2010.13.
- Sampson, R. J., & Lauritsen, J. (1994). Violent victimization and offending: individual-, situational-, and community-level risk factors. In A. J. Reiss & J. A. Roth (Eds.), *Understanding and preventing violence: vol. 3, Social influences* (pp. 1–114). Washington, DC: National Academy Press.
- Sampson, R. J., Raudenbush, S., & Earls, F. (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science*, *277*, 918–924.
- Sarnecki, J. (1990). Delinquent networks in Sweden. *Journal of Quantitative Criminology*, *6*(1): 31–50.
- Sarnecki, J. (2001). *Delinquent networks: youth co-offending in Stockholm*. Cambridge, UK: Cambridge University Press.
- Schwarzer, R., & Leppin, A. (1991). Social support and health: a theoretical and empirical overview. *Journal of Social and Personal Relationships*, *8*, 99–127.
- Snijders, T. A. B., & Baevelde, C. (2003). A multilevel network study of the effects of delinquent behavior on friendship evolution. *Journal of Mathematical Sociology*, *27*, 123–151.
- Srinivasan, S., O’Fallon, L. R., & Dreary, A. (2003). Creating healthy communities, healthy homes, healthy people: initiating a research agenda on the built environment and public health. *American Journal of Public Health*, *93*(9), 1446–1450.
- Taylor, S. E. (2007). Social support. In H. S. Friedman & R. C. Silver (Eds.), *Foundations of health psychology* (pp. 145–171). New York: Oxford University Press.
- Tita, G., Cohen, J., & Enberg, J. (2005). An ecological study of the location of gang “set space”. *Social Problems*, *52*(2), 272–299.
- Valente, T. W. (1995). *Network models of the diffusion of innovations*. Cresskill, NJ: Hampton.
- Valente, T. W. (2005). Models and methods for innovation diffusion. In P. J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis*. Cambridge, UK: Cambridge University Press.
- Valente, T. W., Gallaher, P., & Moultapa, M. (2004). Using social network to understand and prevent substance use: a transdisciplinary perspective. *Substance Use & Misuse*, *39*, 1685–1712.
- Valente, T. W., Unger, J., & Johnson, A. C. (2005). Do popular students smoke? The association between popularity and smoking among middle school students. *Journal of Adolescent Health*, *37*, 323–329.
- Wandersman, A., Duffy, J., Flaspohler, P., Noonan, R., Lubell, K., Stillman, L., et al. (2008). Bridging the gap between prevention research and practice: the interactive systems framework for dissemination and implementation. *American Journal of Community Psychology*, *41*, 171–181.
- Wylie, J. L., & Jolly, A. (2001). Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. *Sexually Transmitted Diseases*, *28*(1), 14–24.
- Zingraff, M., Leiter, J., Myers, K., & Johnson, M. (1993). Child maltreatment and youthful problem behavior. *Criminology*, *31*, 173–202.

Part V
Approaches to Injury Reduction

Chapter 27

Legal Approach

Tom Christoffel

Introduction

The classic public health approach is one that involves surveillance, epidemiological analysis, and intervention aimed at diminishing the likelihood and severity of an injury or disease problem. Intervention can take the form of education, engineering, or enforcement – the “three Es.” Education strives to reduce high-risk behaviors by providing information regarding risks in a way that will alter people’s attitudes and persuade them to modify their behavior. Engineering works to modify the natural and built environments in ways that will provide passive protection from risk or, at least, make behavior change easier. Enforcement seeks to use legal requirements and prohibitions (a) to alter behavior and reduce risk and (b) to bring environmental changes into being. Obviously, society relies on a combination of these three approaches, and each of the three can complement the other two. Subsequent chapters in this book will discuss educational and engineering approaches to injury prevention. This chapter outlines the ways in which the legal system can be used to reduce the likelihood and severity of injury.

The chapter explains the authority of the federal and state governments to enact laws to reduce injury. Both public law (statutes, regulations, and ordinances) and private law (tort litigation) are addressed. The tension between such legal strictures, on the one hand, and legally protected individual rights, on the other, is discussed, as are the ways in which laws can interact with corporations and the products they market. The chapter concludes with discussion of the injury prevention professional’s involvement with law and with the public policy process.

Using Law to Reduce Injury

It is government, at the local, state, and federal levels, that is responsible for protecting the public’s health and safety. Law has long been an important part of public health practice. Be it childhood immunization laws, quarantine laws, restaurant inspection programs, or bans on toxic chemicals or

T. Christoffel, JD (✉)
Boulder, CO, USA
e-mail: chrgui@indra.com

dangerous products, there is a long, clear history of using law to achieve public health goals. Lawrence Gostin's widely cited definition of public health law is:

the legal powers and duties of the state, in collaboration with its partners (e.g., health care, business, the community, the media, and academe), to ensure the conditions for people to be healthy (to identify, prevent, and ameliorate risks to health in the population), and of the limitations on the power of the state to constrain for the common good the autonomy, privacy, liberty, proprietary, and other legally protected interests of individuals. The prime objective of public health law is to pursue the highest possible level of physical and mental health in the population, consistent with the values of social justice.

(Gostin 2008, p. 4)

When it comes to reducing the occurrence and severity of injury, law is a tool that can be used in many ways. Laws can require certain behaviors, such as seatbelt use. Laws can prohibit other behaviors, such as drinking and driving. And laws can establish regulatory agencies and programs, such as the National Highway Traffic Safety Administration. Some laws apply to people (i.e., the behavior of individuals), some apply to things (e.g., automobiles or children's toys), and some apply to places (e.g., requiring residential sprinklers or breakaway signposts along roadways). Additionally, laws are often enacted to appropriate funding for injury prevention programs (Christoffel and Gallagher 2006, Chap. 9).

Compliance with laws can be uncertain, but the assumption behind most laws, particularly those requiring or prohibiting particular behaviors, is that people will abide by laws not only out of a commitment to good citizenship but also to avoid incurring negative consequences (the deterrence effect). Leon Robertson has suggested that a law is most likely to be obeyed if detection is easy, there are – independent of the law – social expectations to meet the requirements, and few exceptions to compliance are allowed (Robertson 1998, Chap. 6).

There are a great number of injury prevention laws on the federal, state, and local level. These laws cover a broad range of injury areas. Some states have over 100 state and local laws aimed at preventing injuries to children. These laws deal with such matters as child abuse reporting, child restraints in motor vehicles, and the private possession and use of fireworks (Christoffel and Gallagher 2006, pp. 212–213).

On the federal level, laws have created programs to deal with youth suicide prevention, emergency medical services, child abuse prevention, and other injury prevention concerns. In addition, the federal government has established regulatory agencies such as the National Highway Traffic Safety Administration, the Occupational Safety and Health Administration, and the Consumer Product Safety Commission. The federal Centers for Disease Control and Prevention includes a National Center for Injury Prevention and Control.

The Soundness of Injury Prevention Law

Injury is highly preventable. Governments have an important and legitimate role in reducing the occurrence, the severity, and the costs of intentional and unintentional injury. Laws can be quite effective to this end. Certainly there are many notable success stories in which law played a role in reducing injury (e.g., motorcycle helmet use, children's toys, fire safety, and Federal Motor Vehicle Safety Standards).

Whether to enact and enforce a specific injury prevention law may be a complex question on which reasonable minds may disagree. Using the power of law often restricts individual autonomy and threatens economic interest. Therefore, injury prevention laws – both those on the books and those proposed – can generate controversy. Three types of objections are often raised against proposed or existing injury prevention laws. The first objection is that the law would be legally invalid, that it violates a constitutionally protected right or suffers some other legal deficiency. This is a legal

question. The second objection is that even if legally valid, the law will not be effective in achieving its aim. This is a research question. And the third objection is that even though legally valid and potentially effective, the law conflicts with ideological or economic values, such as individual autonomy or economic frugality. This is a political question. The pages that follow address these three arguments.

Legal Validity

An assertion that an injury prevention law is legally invalid can be based on the argument that government lacks the authority to enforce the law. Or it can be based on the argument that some greater legal principle, often to be found in the US Constitution's Bill of Rights, takes precedence over the law. Or a due process objection can be raised, arguing that while government may have the authority to enact and enforce such a law, the failure to have followed proper procedural rules when enacting or enforcing the law invalidates the final result.

The majority of injury prevention and other public health laws are found at the state level. State governments (and, by delegation, cities, towns, counties, and other governmental subdivisions) possess the authority to enforce public health and safety laws under what is known as the "police power," which can be understood as the functions historically carried out by governments in regulating society. Laws enacted under the police power will be upheld if they are found to be reasonable attempts to protect and promote the public's health, safety, or general welfare. The classic definition of state public health authority was provided by the US Supreme Court in 1905 in *Jacobson v. Massachusetts*, 197 US 11, 1905. *Jacobson* upheld the use of the police power to penalize an individual for failure to comply with a compulsory smallpox immunization statute. The Court held that:

The authority of the state to enact this statute is to be referred to what is commonly called the police power – a power which the State did not surrender when becoming a member of the Union under the Constitution....[The police power provides] the authority of a state to enact quarantine laws and 'health laws of every description;' indeed, all laws that relate to matters completely within its territory and which do not by their necessary operation affect the people of other states. According to settled principles the police power of a state must be held to embrace, at least, such reasonable regulations established directly by legislative enactment as will protect the public health and the public safety....The mode or manner in which those results are to be accomplished is within the discretion of the state, subject, of course, so far as Federal power is concerned, only to the condition that no rule prescribed by a state, nor any regulation adopted by a local governmental agency acting under the sanction of state legislation, shall contravene the Constitution of the United States, nor infringe any right granted or secured by that instrument....

In fact, even when it comes to individual rights protected under the US Constitution, state governments have been afforded wide latitude to protect public health and safety. As noted in *Jacobson*:

But the liberty secured by the Constitution of the United States to every person within its jurisdiction does not import an absolute right in each person to be, at all times and in all circumstances, wholly freed from restraint. There are manifold restraints to which every person is necessarily subject for the common good. On any other basis organized society could not exist with safety to its members.

It is not that states have absolute *carte blanche* in terms of how the police power can be implemented. To pass judicial scrutiny, such restraints on the individual must be shown to be necessary, must not go beyond what "is reasonably required for the safety of the public," and must not be arbitrary.

Jacobson was decided over a century ago. Recently Gostin observed that:

If the Court today were to decide *Jacobson* once again, the analysis would likely differ – to account for developments in constitutional law – but the outcome would certainly reaffirm the basic power of government to safeguard the public's health.

Courts have generally been quite willing to uphold compulsory public health measures and they have done so not only in respect to communicable diseases but also for non-contagious health problems, such as the mandating of fluoride in public drinking water or requiring vision and hearing tests for school children. Where there is a clear risk of disease or injury and the proposed response is supported by relevant public health expertise, the courts have afforded government extensive authority to intervene.

In a 1946 decision the Supreme Court noted that “The police power is one of the least limitable of governmental powers....” The Court did so while upholding the constitutionality of a municipal ordinance that required automatic fire sprinkler systems in commercial buildings. The Court found the city’s interest in protecting the safety of its citizens to be superior to any property interests (*Queenside Hills Realty Co., Inc. v. Saxl, Commissioner of Housing and Buildings of the City of New York*, 328 US 80).

In *State of Iowa v. John Hartog*, 440 N.W.2d 852 (Iowa, 1989) *cert. denied*, 493 US 1005 (1989) rehearing denied, 493 US 1095 (1990), the Supreme Court of Iowa held that Iowa’s mandatory seat belt law was “a proper exercise of the state’s police power and does not violate the due process provisions of the federal and Iowa constitutions.” They did so in response to the defendant’s argument: “...that the purpose of the statute is to protect the individual from his own folly and, consequently, such purpose has no relation to the public health, safety, or welfare. Implicit in Hartog’s argument is that the decision whether to wear a seat belt is a personal one affecting him only; therefore, he should be able to make that decision free of state interference....” In upholding the seat belt law, the court cited the following commentary with approval:

The government provides roads as a service to its citizens, and part of that service is assuring that these roads will be safe and efficient. The motorist is not being overly imposed upon when asked to comply with minimal standards of behavior designed to reduce the dangers of his driving to other drivers. It is also difficult to object to the state’s attempt to stop an individual from making the rest of society pay for the consequences of his risk-taking. Under a system of laissez-faire one could argue that a person’s risk-taking would be his own business, but...our government provides services from the ambulance that delivers the injured motorist to the hospital to disability insurance. Having to buckle up may be inconvenient, but it is not an unreasonable price to pay for the use of public roads.

There is not a federal equivalent to state police power authority. Under the US Constitution, the federal government is given specifically enumerated powers, but the protection of public health and safety is not one of them. Yet during the past half-century more and more of the use of law to reduce injury has shifted to the federal government. So where does the federal government get its authority to enact injury prevention programs such as those dealing with consumer product safety or youth suicide prevention? The US Supreme Court has construed the federal government’s specifically enumerated powers to also include those things necessary to advance those powers. Two enumerated powers – the power to regulate interstate commerce and the power to tax and spend – have been broadly interpreted to support a wide range of federal public health efforts, including injury prevention. The Consumer Product Safety Act and the Occupational Safety and Health Act have been found to fall under the federal government’s interstate commerce authority. Other federal safety provisions are based on the spending power. State governments are routinely induced by Congress to use their police power authority to enact laws, including injury prevention laws, as a *quid pro quo* for receiving federal funding in an area related to the enacted laws. Of course federal laws, like those enacted on the state level, must not infringe upon individual rights protected under the US Constitution.

There is a worrisome cloud on the legal validity horizon, and surprisingly it is one that relates to trade policy. By joining the World Trade Organization and by signing international trade agreements, the federal government has surrendered some of its authority to enforce public health protections. The agreements have the legal potential to negate federal (as well as state and local) injury prevention (and other public health) protections. Despite being of proven effectiveness and otherwise

legally valid, a needed public health measure may be attacked and potentially invalidated by trade partners as an impermissible restraint on trade (Stier et al. 2007, pp. 522–523).

Finally, the ways in which laws are enacted and/or implemented can affect legal validity. Particularly, in the regulatory area, governmental authority must be exercised properly, affording due process in the application of the law. For example, the federal Administrative Procedures Act, 5 U.S.C. Sec. 551 et seq, spells out basic procedural safeguards to be followed by regulatory agencies. These include such things as notice, opportunity for input, hearings (in some, not all, cases), the opportunity to challenge regulations, and judicial review. In other words, government must do it right or its efforts may be for naught.

Effectiveness

The fact that an injury prevention law is constitutionally valid does not necessarily mean that it is a good law. The statute books certainly contain laws that are dubious, ineffective, or even counterproductive. Can we assess the effectiveness of an injury prevention law (or, for that matter, any public health law)?

It has long been clear that evaluating injury prevention laws is important. A quarter of a century ago, the Committee on Trauma Research recommended that:

Laws and regulations aimed at controlling injuries should be scientifically evaluated. The separate influences of degree of enforcement, severity of punishment, and speed of administration of punishment should also be researched.

(Committee on Trauma Research 1985, pp. 46–47)

This remains an important goal. But evaluating the effectiveness of injury prevention laws is a difficult research task. For example, if an injury problem is a high priority – e.g., drinking and driving – the response is usually not limited to enacting a single law (nor should it be). There will be a variety of different laws. Levels of enforcement will vary. There will also be a variety of educational campaigns. Moreover, once an injury problem is identified, laws are likely to be enacted in many different jurisdictions, in many different versions. Separating out the impact of one particular law in one particular jurisdiction will be challenging, yet differences in laws and their enforcement from jurisdiction to jurisdiction can make aggregating data problematic. And certainly randomization will rarely be possible (Christoffel and Teret 1993, Sect. IV).

Nevertheless, these methodological and other challenges can and are being met. Although much more needs to be done, a number of injury prevention laws have been evaluated. One noteworthy approach has been to identify the key components of a law that may vary from state to state and to measure their presence or absence in relation to differences in injury rates. A study of graduated driver licensing (GDL) in 43 states (36 with GDL, 7 with no GDL) identified seven separate components of such laws (e.g., restrictions on nighttime driving or carrying passengers). The researchers then compared the fatal crash rates of 16-year-old drivers in states with varying numbers of components versus states with none of the components. The authors found that fatal crash rates were lowered by 18% with any five of the seven GDL components and by 21% with six or seven components (Baker et al. 2006).

A recent cataloging of “systemic reviews” of the effectiveness of 52 public health laws included 16 injury prevention laws. Of these 16, 15 were found effective and one “not determined.” Among those found effective were safety belt and safety seat laws, safety belt primary enforcement laws, GDL laws, and red-light camera laws. The authors of the review concluded that: “Much remains to be done... more primary studies of the effectiveness of public health laws, systematic reviews of those studies, and initiatives to make the results available to public policy makers.” (Moulton et al. 2009).

In addition to methodological difficulties, there are other reasons more evaluation is not being done. Obviously, in this day of shrinking local, state, and federal budgets, government funding is extremely hard to come by. Moreover, legislators are politicians, and they may shy away from having laws or programs they supported subjected to evaluation out of fear that the evaluation will show little or no positive impact.

Fortunately, this rather haphazard approach to evaluation is changing. In 2009, the Robert Wood Johnson Foundation began a major initiative to expand the field of public health law *research*. The Foundation noted that:

As public health practitioners, policymakers and others consider how laws influence the public's health, they need evidence to inform questions such as: How does law influence health and health behavior? Which laws have the greatest impact? Can current laws be made more effective through better enforcement, or do they require amendment? The purpose of RWJF's *Public Health Law Research* program is to answer such questions by building a field of research and practice in public health law....

(Robert Wood Johnson Foundation 2010)

Some of the leaders in the RWJF program have described the task at hand in improving public health law research. "The better we understand how law works, the better we can use it, replicate its successes across jurisdictions, and extend its approach to other kinds of health risks." They suggest the need for mapping studies (What laws are out there?), implementation studies (How well are laws being implemented?), intervention studies ("[E]valuate the intended and incidental effects of legal interventions on health outcomes..."), and mechanism studies ("[H]ow the law has the effects it has.") (Burris et al. 2010).

There are three reasons that the current surge in studying, compiling, and evaluating public health laws is so important. First, it allows us to ascertain whether what we are doing is effective. Second, it affords public health professionals in other jurisdictions the opportunity to learn and duplicate laws that work. And, finally, it provides assurance to the public. As Mariner, Annas, and Glantz write:

The public will support reasonable public health interventions if they trust public health officials to make sensible recommendations that are based on science and where the public is treated as part of the solution instead of the problem.

(Mariner et al. 2005)

Political Considerations

Injury prevention laws that are constitutionally valid and have evidence of efficacy may still be objected to for ideological and/or economic reasons. It is certainly possible that an otherwise supportable legal approach to an injury problem would be rejected due to its considerable financial cost or because it is viewed as going too far in restricting some valued aspect of individual autonomy. Thus, there may be a situation in which it could be reasonably argued that a proposed law not be enacted even though government could do so without running afoul of constitutional limitations and even though research has demonstrated its potential effectiveness.

But if such objections are raised it is wise to look first at the self-interest of those objecting. Ours is not a classless society. The disadvantaged suffer from higher rates of injuries than the more well-to-do (Zarzaur et al. 2010). As Quick (1991) has noted, the risk of injury "is one of the clearest instances of health inequality in our society." At the same time, efforts to prevent injury often challenge the economic interests of the very powerful. This means that efforts to use the law to reduce injury may run into opposition that is both well financed and highly motivated. A classic example is the decades-long effort to require the installation of air bags in automobiles. It took over 20 years, 60 rulemaking notices, plus a Supreme Court decision, to finally get in place the regulations requiring

that airbags be installed in new passenger vehicles. This type of delay and opposition should not be surprising; injury prevention efforts most often involve attempting to change the status quo.

Gostin (2008, p. 514) notes that “it is important to recall that public health, and the law itself, are highly political, influenced by strong social, cultural, and economic forces.” This is certainly true of injury prevention. Not only laws specifically aimed at reducing injury but also laws in other areas can have a critical impact on injury rates.

Early in his legal career, Charles Evans Hughes (Chief Justice of the US Supreme Court, 1930–1941) famously noted: “We are under a Constitution, but the Constitution is what the judges say it is” Justices come and go, usually reflecting the ideology of the Presidents who appoint them. Unfortunately from an injury prevention perspective, we currently have a Court with an extremely conservative majority. This is perhaps best reflected in its Second Amendment decisions.

Until fairly recently, the Supreme Court had interpreted the Second Amendment to be a protection afforded collectively to the states and their militias and police, rather than providing an individual right to possess firearms. The Amendment did not significantly limit federal or state gun control laws. This has changed dramatically. In recent decades, the gun lobby has become a powerful political force, pushing the idea that the Second Amendment did afford individuals a right to bear arms. This view was so at odds with settled law that in an interview after leaving the Court, Chief Justice Warren Burger (appointed by Richard Nixon and considered a conservative) referred to the gun “rights” position as “one of the greatest pieces of fraud, I repeat the word ‘fraud,’ on the American public by special interest groups that I have ever seen in my lifetime” (MacNeil/Lehrer 1991).

Yet two 5–4 decisions have upended this state of affairs. First in a decision that applied to federal gun laws, *District of Columbia v. Heller*, 554 U.S. 570, 2008, and then in a follow-up decision extending this ruling to state and local gun laws, *McDonald v. Chicago*, 130 S.Ct. 3020, 2010, the conservative majority asserted their conviction that the original intent of the Second Amendment was indeed to protect an individual right to firearm possession. They did this despite the fact that there is little historical support for such a view. In fact, in *Heller* 15 of 16 academic historians who joined friend of the Court briefs took the position that the Second Amendment was about protecting state militias. A similar near consensus of historical evidence was presented to the Court in *McDonald*.

The extreme conservatism and judicial activism of the current Court’s five-member majority had a critical impact on injury prevention in the gun cases. These same five Justices have also formed the majority in a variety of decisions that have enhanced the legal status of corporations and businesses – decisions that have the potential to indirectly affect public health laws.

Those objecting to statutory or regulatory controls have a very big megaphone and they have been quite successful in recent decades in pushing for “deregulation,” arguing that laws such as product safety laws interfere with the “free market,” are inefficient, and ultimately hurt the economy. The alternative they propose is self-regulation in the form of voluntary codes and standards. Yet such voluntary self-regulation is usually less rigorous than governmental regulation. Moreover, the truly bad apples can simply ignore such programs and their voluntary standards. The call for less governmental regulation has come in particular from business groups such as the US Chamber of Commerce, from business-funded think tanks, and from the businesses most likely to benefit financially from a loosening of governmental regulatory programs. Their collective efforts to modify laws have influenced the legislative and administrative arenas of government as well as the courts.

In late 2010, the *New York Times* reported on a study that the paper had commissioned in which scholars at Northwestern and University of Chicago Law Schools examined 1,450 decisions of the Court since 1953. The study found that

the percentage of business cases on the Supreme Court docket has grown in the Roberts years, as has the percentage of cases won by business interests. The Roberts court, which has completed five terms, ruled for business interests 61 percent of the time, compared with 46 percent in the last five years of the court led by Chief Justice William H. Rehnquist, who died in 2005, and 42 percent by all courts since 1953.

The *Times* also reported that the US Chamber of Commerce had significantly increased its filing of briefs to the point of being involved in most major business cases before the Court:

The side it supported in the last term won 13 of 16 cases. Six of those were decided with a majority vote of five justices, and five of those decisions favored the chamber's side. One of them was *Citizens United*, in which the chamber successfully urged the court to guarantee what it called "free corporate speech" by lifting restrictions on campaign spending.

(Liptak 2010; Epstein et al. 2010)

None of this bodes well for public health law, particularly injury prevention law. The extremely conservative Court majority has displayed no sympathy for measures aimed at protecting the public. Moreover, they are in harmony with a well-funded anti-regulatory, anti-government faction within Congress, the political parties, and the media (Mayer 2010; Lichtblau 2011). However, this need not lead to hopelessness and despair. Rather, it should lead to enhanced efforts to educate and advocate for new and improved public health and safety laws and programs. How this can be accomplished will be the focus of the concluding section of this chapter.

Tort Law

It would be nice if, when patterns of injuries caused by dangerous products were discerned, manufacturers routinely responded by improving and/or recalling the products. This does not always happen; hence the need for statutory programs. But in the face of industry lobbying, it is not always possible to deal with such a problem through legislation or regulation. Another avenue is private law, litigation between private parties (individuals, corporations, groups, etc). In part because the courts are somewhat protected from corporate pressure, civil litigation lawsuits have become an alternative (albeit imperfect) approach to dealing with some injury problems.

One area of private law, tort law, allows a party who has been harmed by another to sue for damages. Only certain categories of harm are compensable, one of which is unintentionally inflicted harm where it can be shown that negligence – the failure to exercise due care – was involved. Negligence has been defined as:

The failure to exercise the standard of care that a reasonably prudent person would have exercised in a similar situation; any conduct that falls below the legal standard established to protect others against unreasonable risk of harm....

(Black's 2009, p. 1133)

Tort litigation has been used successfully against a variety of threats to the public's health, most notably tobacco, but also toxic substances and environmental contamination. Tort litigation can also play an important role in injury prevention, particularly injuries involving dangerous products.

Public health law experts have long argued that product liability lawsuits can be an effective tool in injury prevention (Vernick et al. 2003, 2004; Leonard 2007). Gostin writes that:

Powerful interest groups...can thwart regulation through the political process. Consumers themselves may rise up in revolt against regulation and taxation of the products they desire.....Where direct regulation through the political process fails, tort law can become an essential tool in the arsenal of public health advocates.

(Gostin 2008, p. 202)

Product liability lawsuits can result from defects in design, from defects in manufacturing, or from failure to adequately warn of product hazards. Product liability lawsuits have been brought successfully against makers of unstable hot water vaporizers, against the manufacturers of particularly dangerous farm machinery, and against BB gun makers for failure to provide mechanisms to indicate whether a gun is loaded. It is worth noting that prior to the 1960s, automobile manufacturers and

the courts both took the view that motor vehicle injuries were the result of driver error, that if a crash occurred injury was inevitable, and that failure to design “crashworthy” vehicles incorporating known technologies was not negligence on the part of the manufacturers. Tort law has advanced since then.

Product liability lawsuits put manufacturers on notice that injured customers could potentially win significant damage awards. A classic product liability decision held that:

public policy demands that responsibility be fixed wherever it will most effectively reduce the hazards to life and health inherent in defective products that reach the market...It is to the public interest to discourage the marketing of products having defects that are a menace to the public. If such products nevertheless find their way into the market, it is to the public interest to place the responsibility for whatever injury they may cause upon the manufacturer, who even if he is not negligent in the manufacture of the product is responsible for its reaching the market.

(*Escola v. Coca Cola Bottling Co. of Fresno*, 24 Cal.2d 453, 150 P.2d 436, 1944)

This statement describes what are considered the two classic functions of tort lawsuits: compensation and deterrence. *Compensation* is rather obvious. If a person is injured through no fault of his own, thereby incurring medical expenses, lost wages, etc., and if the party responsible for causing the injury did so by acting in a negligent manner when they knew or should have known that preventable harm would result, the courts shift the financial burden to the negligent party. The compensation function of tort law is particularly important in the USA (as compared, for example, to most European countries) because the USA has a relatively weak social support network. A party harmed through no fault of his own will often have extensive medical expenses and may lose the ability to work. As T.R. Reid observes, “...every year...some twenty thousand Americans died because they couldn’t get health care. That doesn’t happen in any other developed country. Hundreds of thousands of Americans go bankrupt every year because of medical bills. That doesn’t happen in any other developed country either” (Reid 2009, p. 2). If little of the financial burden is assumed by government, courts can become the default social support system by shifting the financial burden to defendants.

The *deterrence* function of tort law is that of setting an example. By making negligent actions costly, courts are sending a message that will hopefully dissuade others from acting in similarly negligent fashion in the future. In some instances, where compensation for the plaintiff’s harm seems trivial as compared to the defendant’s assets, courts may impose additional damage awards (known as punitive damages) to further punish the negligent party and thus strengthen the deterrent message of the litigation.

Put more positively, tort damage awards can encourage individuals and corporations to act responsibly and with due care. They can take a defective product off the market, redesign the product to eliminate the dangerous defect, or – for unavoidably dangerous but useful products – provide adequate warning. The deterrence function of tort law becomes increasingly important when government regulatory programs are weak. To the extent that consumers are protected from negligence by governmental programs, this deterrence function is less important.

It should be noted that product liability lawsuits are an imperfect mechanism. They can take years to make their way through the courts, thus delaying and muting any harm prevention potential they offer. Moreover, manufacturers’ insurance coverage can limit their impact. And the general fear of such lawsuits, while it hopefully encourages more responsible business practices, can also lead to less openness and sharing of information by manufacturers. Yet despite such weaknesses, there can still be a positive impact of tort litigation. For example, in 1982 Stephen Teret and Edward Downey published an article in *Trial* magazine arguing that it was negligence on the part of automobile manufacturers to refuse to install air bags in cars after their lifesaving potential had been demonstrated. The article prompted a willingness among personal injury lawyers to litigate such negligence lawsuits. It was only after some of these lawsuits led to multimillion dollar settlements that the automobile companies began offering air bags as an option (Teret and Downey 1982;

Teret 1986). Eventually the Federal Motor Vehicle Safety Standards were strengthened, so that by the end of the 1990s dual airbags were mandatory in all new automobiles.

Because product liability lawsuits are directed at some of the biggest corporations in the country, it should be no surprise that a variety of efforts have been made to block or limit them. Corporate America has mounted a campaign to promote the idea that there is a “litigation crisis” that – according to their campaign – is harming the economy. But there is no litigation crisis. Courts and juries are not awarding damages wildly. The stereotype of greedy plaintiff’s attorneys and out-of-control juries escalating product liability awards is not borne out by the facts. In terms of the impact on the national economy, this is not a significant issue. But the stereotype has been pushed effectively enough to result in a variety of legislative restrictions on product liability litigation (Glaberson 2001; Bogus 2001; Cohen 2009). To the extent that the “litigation crisis” myth succeeds in limiting the tort system’s classic compensation and deterrence functions, society as a whole will suffer.

Conclusion

Most injury prevention professionals are not lawyers. But because law can be a useful tool in achieving and supporting injury prevention goals, injury prevention professionals need to be alert to the role of law in their field and ready to interact with lawyers and legislators. This might include carrying out research that might aid policymakers by providing data on injury problems or suggesting the most promising preventive interventions. It might mean helping in the process of enacting new legislation by providing data, ideas, and testimony. It might include assisting in the development of the regulations needed to implement a law. And it might include advising on improvements in enforcement efforts (Christoffel and Gallagher 2006, Ch. 14; Lopez and Frieden 2007). As Burriss et al. have written:

Researchers are often isolated from the policy process and disconnected from policymakers and public health practitioners, making it difficult for them to identify salient topics for study and to produce knowledge that can both respond to policymakers’ concerns and drive policy agendas toward evidence-based innovation....

(Burriss et al. 2010)

Attending to the policy process can lead injury prevention professionals toward advocacy, which is not a bad thing. In fact, it could be argued that advocacy is as much a part of public health as data collection, epidemiology, and program design. This may require adjusting to a changed professional role. As Susan Baker has pointed out, “The role of advocate does not come easily to many scientists. Yet often it is only by taking on this role that we can turn our special knowledge about the causes of injury into public policies that will prevent injury.” (Baker 1989).

Working to advance injury prevention laws and programs means many things. It means understanding the self-interests and power relationships that often hamper injury prevention efforts. It means working to strategize as to where and how to use leverage points, how to frame issues, and how to bring counterforces to bear in the face of opposition to injury prevention efforts. It means spreading the message that injuries are preventable, that they are not random events. It means emphasizing the fact that injury prevention is cost effective, that it can save society money. It means developing and working with allies (of which injury victims and their families can be a particularly effective group). And it means maintaining ongoing relationships with legislators and other key decision makers.

The entire process of bringing new laws into being includes multiple steps. Larry Berger has suggested that the injury prevention professional involved in this process:

be thoroughly convinced that the bill addresses a strikingly important issue. One should have evidence that the bill’s actions can be effective; support from judges and police officers that the law can be enforced expeditiously;

economic estimates that excessive costs will not be involved; legal counsel confirming the constitutionality and compatibility of the proposed law with existing legislation and ordinances; and broad-based support from constituents.

(Berger 1981)

Although government employees and public health agencies must avoid direct involvement in partisan politics, this does not prevent public health experts from educating the public, stakeholders, advocacy groups, and decision makers. Public health has always included challenges and often achieved successes. It is all part of fighting the good fight (Wallack et al. 1993, 1999).

References

- Baker, S. (1989). On receiving the Charles S Dana award for pioneering achievements in health and higher education, New York City. http://www.traumaf.org/advocates/q_and_a2.shtml. Accessed 30 Oct 2011.
- Baker, S. P., Chen, L., & Li, G. (2006). *National evaluation of graduated driver licensing programs*. Washington, DC: National Highway Traffic Safety Administration (DOT HS 810614).
- Berger, L. R. (1981). Childhood injuries: recognition and prevention. *Current Problems in Pediatrics*, 12(1), 12–24. *Black's law dictionary* (9th ed.). (2009). St. Paul, MN: West.
- Bogus, C. J. (2001). *Why lawsuits are good for America: disciplined democracy, big business, and the common law*. New York: New York University Press.
- Burris, S., Wagenaar, A. C., Swanson, J., Ibrahim, J. K., Wood, J., & Mello, M. M. (2010). Making the case for laws that improve health: a framework of public health law research. *Milbank Quarterly*, 88, 169–210.
- Christoffel, T., & Gallagher, S. S. (2006). *Injury prevention and public health: practical knowledge, skills, and strategies* (2nd ed.). Sudbury, MA: Jones and Bartlett Publishers.
- Christoffel, T., & Teret, S. P. (1993). *Protecting the public: legal issues in injury prevention*. New York: Oxford University Press.
- Cohen, T. H. (2009, November). Civil justice survey of state courts: Tort bench and jury trials in state courts, 2005. *Bureau of Justice Statistics Bulletin*. NCJ 228129. <http://bjs.ojp.usdoj.gov/content/pub/pdf/tbjtsc05.pdf>. Accessed 24 Jan 2011.
- Committee on Trauma Research. (1985). *Injury in America: a continuing public health problem*. Washington, DC: National Academy Press.
- Epstein, L., Landes, W. M., & Posner, R. A. (2010). Is the Roberts Court pro-business? <http://epstein.law.northwestern.edu/research/RobertsBusiness.pdf>. Accessed 21 Jan 2011.
- Glaberson, W. (2001, August 6). A study's verdict: jury awards are not out of control. *New York Times*, Section A, Page 9. <http://query.nytimes.com/gst/abstract.html?res=FB0B17FF3A580C758CDDA10894D9404482>. Accessed 16 Jan 2011.
- Gostin, L. O. (2005). *Jacobson v Massachusetts* at 100 years: police power and civil liberties in tension. *American Journal of Public Health*, 95(4), 576–581. doi:10.2105/AJPH.2004.055152.
- Gostin, L. O. (2008). *Public health law: power, duty, restraint*. Berkeley, CA: University of California Press.
- Leonard, E. W. (2007). Beyond compensation: using torts to promote public health. *Journal of Health Care Law & Policy*, 10. <http://ssrn.com/abstract=939328>. Accessed 24 Jan 2011.
- Lichtblau, E. (2011, January 19) Advocacy group says justices may have conflict in campaign finance cases. *New York Times*. <http://www.nytimes.com/2011/01/20/us/politics/20koch.html>. Accessed 25 Jan 2011.
- Liptak, A. (2010, December 18). Justices offer receptive ear to business interests. *New York Times*. http://www.nytimes.com/2010/12/19/us/19roberts.html?_r=4&pagewanted. Accessed 28 Jan 2011.
- Lopez, W., & Frieden, T. R. (2007). Legal counsel to public health professionals. In R. A. Goodman, R. E. Hoffman, W. Lopez, G. W. Matthews, M. A. Rothstein, & K. L. Foster (Eds.), *Law in public health practice* (2nd ed.). New York: Oxford University Press.
- MacNeil/Lehrer Newshour. (1991, December 16). Interview with Charlayne Hunter-Gault.
- Mariner, W. K., Annas, G. J., & Glantz, L. H. (2005). *Jacobson v Massachusetts*: it's not your great-great-grandfather's public health law. *American Journal of Public Health*, 95(4), 581–590. doi:10.2105/AJPH.2004.055160.
- Mayer, J. (2010, August 30). Covert operations: the billionaire brothers who are waging a war against Obama. *The New Yorker*, p. 45. http://www.newyorker.com/reporting/2010/08/30/100830fa_fact_mayer. Accessed 25 Jan 2011.
- Moulton, A. D., Mercer, S. L., Popovic, T., Briss, P. A., Goodman, R. A., Thombly, M. L., et al. (2009). The scientific basis for law as a public health tool. *American Journal of Public Health*, 99(1), 17–24. doi:10.2105/AJPH.2007.130278.

- Quick, A. (1991). *Unequal risk: accidents and social policy*. London: Socialist Health Association.
- Reid, T. R. (2009). *The healing of America: a global quest for better, cheaper, and fairer health care*. New York: Penguin.
- Robert Wood Johnson Foundation. (2010). Public health law research: making the case for laws that improve health. http://www.rwjf.org/files/applications/cfp/cfp_PHLR2010Rapid.pdf. Accessed 18 Jan 2011.
- Robertson, L. S. (1998). *Injury epidemiology: research and control strategies* (2nd ed.). New York: Oxford University Press.
- Stier, D. D., Mercy, J. A., & Kohn, M. (2007). Injury prevention. In R. A. Goodman, R. E. Hoffman, W. Lopez, G. W. Matthews, M. A. Rothstein, & K. L. Foster (Eds.), *Law in public health practice* (2nd ed.). New York: Oxford University Press.
- Teret, S. P. (1986). Litigating for the public's health. *American Journal of Public Health*, 76(8), 1027–1029.
- Teret, S., & Downey, E. (1982). Air bag litigation: promoting passenger safety. *Trial*, 18(7), 93–99.
- Vernick, J. S., Mair, J. S., Teret, S. P., & Sapsin, J. W. (2003). Role of litigation in preventing product-related injuries. *Epidemiological Review*, 25, 90–98.
- Vernick, J. S., Sapsin, J. W., Teret, S. P., & Mair, J. S. (2004). How litigation can promote product safety. *Journal of Law, Medicine & Ethics*, 32, 551–555.
- Wallack, L., Dorfman, L., Jernigan, D., & Themba, M. (1993). *Media advocacy and public health: power for prevention*. Newbury Park, CA: Sage.
- Wallack, L., Woodruff, K., Dorfman, L., & Diaz, I. (1999). *News for a change: an advocate's guide to working with the media*. Thousand Oaks, CA: Sage.
- Zarzaaur, B. L., Croce, M. A., Fabian, T. C., Fischer, P., & Magnotti, L. J. (2010). A population-based analysis of neighborhood socioeconomic status and injury admission rates and in-hospital mortality. *Journal of the American College of Surgeons*, 211(2), 216–223.

Chapter 28

Public Policy

David Hemenway

This chapter is divided into three sections. The first discusses the rationale for public policy – when “the market” might not work well enough to ensure the optimal level of prevention and protection against injury. The second section describes three main aspects of safety regulations: the rules themselves, the monitoring of those rules, and the penalties for noncompliance. The third section briefly describes the public health approach to policy, differentiating it from the medical and criminal justice approaches. Many success stories in injury prevention (Hemenway 2009) are depicted in section “The Rationale for Public Policy.” In section “Three Aspects of Regulation,” the Brady gun law is used to illustrate the three main aspects of regulations. Section “The Public Health Approach to Policy” describes the public health approach to suicide prevention.

The Rationale for Public Policy

In the USA, the economists’ framework has become the dominant one for analyzing public policy initiatives. A key construct in economic theory is the model of perfect competition, in which decentralized decision making (“the market”) leads to an optimal efficiency outcome in the long run. For economists, the only efficiency reason for government intervention in the market is if (1) there are deviations from the perfectly competitive model (“market failure”) and (2) government intervention would improve the outcome (i.e., there is not too much “government failure”). A second rationale for government intervention involves equity rather than efficiency – if there are income or wealth inequalities and redistribution is deemed desirable; for example, someone through no fault of his own may become disabled and unable to support himself.

For economists, “the market” is also broader than the buying and selling of goods but refers to any private decisions, such as deciding how fast to drive, or whether to hit back after someone hits you. It is important to note that in the economic approach the default is “the market.” The burden of proof is on anyone advocating government action; they must show that the market is not working well, and that government intervention could help. Markets, of course, often work well in providing incentives for private individuals and institutions to act in ways to reduce injury.

D. Hemenway, PhD (✉)

Harvard Injury Control Research Center, Harvard School of Public Health, Kresge 309,
677 Huntington Avenue, Boston, MA 02115, USA
e-mail: hemenway@hsph.harvard.edu

When Markets Work

When markets are working well, they provide strong financial incentives for innovators to create products which are safer and less costly. In well-functioning product markets, producers are continually searching for ways to innovate and improve the product and to reduce its cost, because that is the way to increase their sales, market share, and profits. Consumers are the clear beneficiaries of the improved and less expensive products. Improvements in residential smoke detectors and ski boots/bindings are examples of two of the many areas where market forces have helped reduce injury.

Residential Smoke Detectors

The most dangerous residential fires typically occur when people are asleep. By providing early warning of fire, it is estimated that a working smoke detector reduces the risk of residential fire death by almost 50%. The first battery-powered smoke detector was patented in 1969. Soon smoke alarms were a familiar presence in American homes, increasing from being present in 5% of homes in 1970 to 75% in 1985, to 96% in 2008. The main reason for this increase was the low price. While the cost of protecting a three-bedroom household with professionally installed alarms was about \$1,000 in 1970, the price is now about \$10 per alarm, or less than \$50 for the entire house. The National Fire Protection Association says that “smoke alarms are the residential fire safety success story of the past quarter century” (National Fire Protection Association 2005).

Ski Boots and Bindings

Lower extremity injuries, such as sprained ankles and fractured tibia commonly occurred to snow skiers in the 1960s. Over the next 20 years, progressive improvements were made to the ski boot (such as providing increased support for the ankle) which, combined with a reduction in the torque release of the bindings (which reduced twist-related problems) led to a marked decrease in lower leg injuries from skiing between the 1960s and 1980s (Johnson and Ettingler 1982; Matter et al. 1987). “In the 1960s ambulances used to line up at ski areas to shuttle the injured skiers to local medical facilities. By the 1980s, these lines of ambulances were a thing of the past, a relic of a more dangerous era” (Hemenway 2009, 70).

When Government Action Can Help

Economic theory suggests that governmental intervention in the market may improve outcomes when the assumptions of the perfect market are violated. In this section we will focus on three of these assumptions: (1) perfect information, (2) no externalities, and (3) consumer rationality. A fourth section explains the case for “public goods,” which are sometimes included in the broad area of market failure due to externalities.

Imperfect Information

One aspect of a perfect market is that both buyers and sellers have excellent information about price, product quality and safety, and alternative products. A major problem in most retail markets is that most buyers have little knowledge or understanding of the quality or safety of what they are buying.

Given the large number of products purchased by any household, and the small quantities involved in each purchase, most consumers are “rationally ignorant.” By contrast, large firms buying in bulk hire purchasing agents whose job is to become expert about the products the firm may buy. However, for the average retail consumer, it is typically not worthwhile to spend the time and money trying to obtain large amounts of such information about any one item. As an example, readers can look at the clothes they are wearing and consider what they know about their level of flammability.

When buyers do not have good information about safety, the government can help provide that information (e.g., the National Highway Safety Administration does crash testing and makes the data public). Another approach for the government is simply to mandate a minimum level of safety for products; this approach can be much more cost effective if the overwhelming majority of buyers would want, at a minimum, that level of safety. For example, before the 1967 model year, the steering column in cars was a rigid pole ending in a narrow hub. It was like a spear pointed at the driver. In the 1960s, the General Services Administration began requiring improved steering assemblies in government-purchased vehicles. These reduced injuries in frontal crashes. Instead of trying to educate motorists about the safety benefits of the energy-absorbing (“collapsible”) steering column, and letting the market work, the National Highway Safety Administration sensibly mandated energy-absorbing steering columns for all vehicles. These reduce the risk of driver fatality in a frontal crash by 12% and serious injury by 17% (Kahane 1981). Most motorists would prefer to have this inexpensive safety feature in cars, but remain blissfully unaware of the entire issue; the costs of trying to educate them all about all the many currently mandated safety features in automobiles would be enormous, and probably not very successful.

In the early 1980s, handheld hair dryers caused an average of 18 deaths per year in the USA. The typical victim was a young child in the bathtub. These were very rare events in a population of more than 280 million people, but they were devastating and avoidable. Standards and certifying organizations (e.g., Underwriters Laboratories), the industry, and the government [i.e., Consumer Product Safety Commission (CPSC)] worked together to address the problem. First they required a pictorial warning on the hair dryers, showing that it was dangerous to drop a dryer in water. A second and more important step was to write voluntary standards that were incorporated into the National Electric Code requiring that all new hair dryers provide protection against electrocution. This effectively eliminated these tragic deaths. The enormous cost of trying to educate hundreds of millions of consumers was eliminated; consumers are not currently offered the option of a dangerous, but marginally less expensive, product.

Externalities

Economists tout the benefits of free trade, in large part because they expect that voluntary trades will be mutually beneficial to both parties – or the parties would not have voluntarily agreed to the transaction. However, the exchange may impose costs, or provide benefits, to third parties whose preferences are not fully taken into account by the buyer and seller (e.g., “happy hour” at the bar may be mutually acceptable to both customer and bar owner, but imposes a cost on the neighboring community in increased traffic crashes). The effects on third parties are “externalities,” and when externalities exist, even the outcomes of competitive markets can be extremely inefficient. A preferred economic solution is to “internalize the externality” through taxes or subsidies, ensuring that the costs and benefits of third parties are included in the utility calculus of the buyer and seller. In the real world, if the transaction or activity is seen as providing little benefit, but large external costs, an approach often used is to make the activity illegal and to set the punishment (e.g., fine) at an appropriate level to internalize the externality.

Driving a car is dangerous, not only to oneself but also to others. Drunk drivers are especially prone to crash and cause injury; alcohol-impaired driving is a leading cause of traffic fatalities in

most countries. When weighing their own costs and benefits when deciding whether or not to drink and drive, motorists may not sufficiently take into account the costs they impose on others. In 2002, Japan decided to try to reduce the amount of drunk driving by dramatically increasing the financial penalty – from about \$425 to \$4,250, a tenfold increase. In addition, the law made bartenders and passengers culpable along with the arrested drivers. The law reduced both alcohol-impaired traffic fatalities and serious injury by over 35% (Nagata et al. 2008).

Drivers also impose costs on others from their vehicles' emissions. For more than 50 years, American cars used leaded gasoline, which improved engine performance, but spewed lead into the air. Leaded gasoline was a major contributor to childhood lead poisoning – which increased the likelihood of learning disabilities, lower IQ, hyperactivity, and antisocial behavior, including violence. In 1972, the Environmental Protection Agency launched an initiative to phase out leaded gasoline, and by 1986 the phaseout was largely complete. Average lead blood levels in young children fell over 75%, due to change to unleaded gasoline, along with legislation banning lead from paint and plumbing supplies (U.S. Environmental Protection Agency 2000).

Rather than banning goods, or making activities illegal, American economists would generally prefer to use taxes and subsidies to internalize the externalities. For example, in the nineteenth century, matches made of white phosphorus caused a terrible disease (phossy jaw) among match workers. Most European countries outlawed the production of white phosphorus matches. By contrast, the USA put a tax on white phosphorus matches, making them more expensive than matches made of nonpoisonous phosphorus sesquisulfide. This achieved the same result (the elimination of white phosphorous matches, and phossy jaw) while not actually prohibiting its manufacture (Myers and McGlothlin 1996).

Consumer Nonrationality

The economic model assumes that people are rational – they are able to carefully weigh the costs and benefits of various courses of action and make the wise decision. Yet parents, and others, know that their children are not fully rational – children have had little experience and their brains are far from fully developed. The economic model also assumes that people are born with well-defined preferences that are stable over time, yet evidence suggests that people's preferences are often ill defined and quite malleable. Finally, the economic model assumes that people are never tired, lazy, or stupid, and that their decisions are carefully calculated and not affected by emotion. A whole new branch of economics, “behavioral economics,” has been developed to analyze the many real-world situations where the rationality assumption is violated. Below we discuss three types of policies: those which (a) protect children, (b) shape tastes, and (c) help protect against human frailties such as procrastination and inertia.

Children: It is generally agreed that society has a responsibility to protect its children. Many injury prevention successes have come from governmental policies designed specifically to reduce injury to minors. Two examples from the motor vehicle arena are child safety seats and graduated licensing. In the 1960s, it was common for parents to hold infants in their lap while traveling. During a crash, the physical forces would make it impossible to hold on to the child, who would crash into the car interior or through the windshield. In the mid-1970s, a small group of Tennessee pediatricians successfully lobbied for a law that would require that very young children be restrained in the car. In 1978, Tennessee became the first state to require child safety seat use; within 7 years every state in the union had passed such a law. Infants and toddlers in crashes who are not in car seats are more than ten times more likely to die in a crash than those who are restrained (Kalbfleisch and Rivara 1989).

Data showed that 16 year olds had three times the crash risk of older teenagers, with the first few months of driving having the highest rates of injury. Research also showed that especially dangerous

times were at night and when another teen was the right front passenger. A graduated licensing system was developed to provide experience to young drivers, while minimizing the risk of collision. The system requires a period with supervised driving, followed by an intermediate phase that permits unsupervised driving only in less risky situations (e.g., not at night, or with an adolescent front seat passenger). In 1997, Michigan became the first state to adopt such a system; within a decade all states had followed suit. These state-mandated graduated licensing programs reduced young driver crash risk by 20–40% (Shope 2007).

Shaping tastes: The economic model assumes that people have stable, consistent, and well-defined preferences. In the real world, preferences are imprecise and malleable. And it is well known that not only do preferences influence behavior, but behavior can also affect preferences and social norms. For example, seatbelt laws change use, and then preferences (Steptoe et al. 2002). Once people begin to wear seat belts, they get used to them. Change is always hard, but soon people become accustomed to wearing the belts. Adults used to not wearing seat belts saw them as confining. By contrast, children who have been in safety restraints since infancy feel uncomfortable without them.

In the second half of the twentieth century, Sweden began changing laws concerning the corporal punishment of children. Child abuse was always illegal, but soon spanking was also banned. There was no intent to criminalize parents for spanking, and the law was never going to be strongly enforced. Instead the main goal was to alter public attitudes, to acknowledge children's rights as autonomous individuals (Durrant 1996). The ban was so well publicized that 99% of Swedes were familiar with the law, a level of knowledge unmatched "in any other study on knowledge about law in any other industrialized society" (Ziegert 1983). And the law did indeed change attitudes. Whereas a majority of Swedes believed in the necessity of corporal punishment in 1965, only 11% supported its use 30 years later. "The Swedish corporal punishment law has been very effective in shaping a social consensus regarding the rejection of corporal punishment in childrearing" (Durrant 1996).

Nonrational frailties: Psychologists and behavioral economists have documented many systematic "mistakes" that people make, consistently and repeatedly, that violate the rules of rationality. These "mistakes" have such names as the endowment effect (people place a higher value on objects they own than objects that they do not), status quo bias, presentism (the belief that our feelings tomorrow will be exactly like today), and loss aversion (Thaler and Sunstein 2008). For example, because of the status quo bias – inertia – the "default option" has a disproportionate effect on people's decisions. In France, for instance, close to 100% of French citizens voluntarily agree to be organ donors, compared to fewer than 20% in the UK. This result appears to have almost nothing to do with differences in beliefs or altruism. Instead French motorists have to check a box to *not* become a donor, while in the UK, the motorists have to check a box *to* become a donor. Mostly, no one in either country takes the time to check a box.

Government can help determine the defaults which are set in the public and private sectors, and the default can affect injury rates. The tap water burn story illustrates a governmental intervention that reduced injury by changing a private "default," while maintaining consumer choice. Household tap water can be a major source of serious injury. Exposure to water at 140°F can severely burn a young child's skin in less than 5 s. Unfortunately, in the 1970s, the factory preset temperature for water heaters was 140–150 degrees. Various state laws now mandate that the factory preset be 120–125 degrees. Adults can easily change the temperature setting, but almost no one does. This change in the default setting dramatically reduced child tap water burn injury rates (Erdmann et al. 1991).

Public Goods

In the economic model, government is expected to provide "public goods" – goods such as clean air or national defense – which everyone can enjoy and individuals cannot efficiently provide for

themselves. For public goods, each individual's consumption of the good does not subtract from other individuals' consumption of the good. A second aspect of pure public goods is that it is impossible to exclude any individuals from consuming the good – even if they are unwilling to help pay for it.

While there are few “pure” public goods in the real world, many goods such as police protection, public parks, and highways are considered close enough that it typically makes sense that they be provided by government. Goods that are government built, such as playgrounds and roads, can be more or less safe. Major advances came when highway safety efforts moved away from improving drivers' habits (the “nut behind the wheel” approach) to building safer cars and roads (the “forgiving roadside” approach). There are many injury prevention success stories attributable to government road building policies which followed this change in philosophy (e.g., crash cushions, guardrails, and roundabouts) (Hemenway 2009). Similarly, public playgrounds can be built and maintained in ways that either reduce or promote injury. In the early 2000s, for example, the Toronto School District Board improved the playground equipment in close to 400 elementary schools, immediately saving some 500 children from injury (Howard et al. 2005).

As a large buyer in the market, government purchases can affect the safety of many private products (e.g., aircraft, automobiles). For example, the purchases of the General Services Administration (GSA) were crucial in providing financial incentives to automobile manufacturers to equip vehicles with both energy-absorbing steering columns and with airbags (Hemenway 2009). In the 1960s, GSA began requiring improved steering assemblies in government-purchased vehicles, and in 1985 GSA demanded that the 5,000 vehicles it planned to purchase have airbags. The exemplary performance of these new safety features provided real-world evidence for the value of government mandating these products for the general public.

Activities such as basic data collection and basic research are also considered to be public goods. In the USA, the Centers for Disease Control and Prevention assembles and makes data available on fatal and nonfatal injuries (e.g., Vital Statistics, National Violent Death Reporting System) and risk factors for injury (e.g., Youth Risk Behavior Surveillance System). It also funds research into injury prevention – knowledge which can then be used free of charge by individuals, institutions, and communities. You can use the data from Vital Statistics without impinging on my use, and I can benefit from the knowledge gained from injury prevention research without in any way limiting your ability to use those results to make yourself or your community safer. Private companies have insufficient incentive to invest in fundamental research (e.g., research in basic mathematics) because the investing company would pay all the costs and reap only a small part of the total reward (i.e., there are positive externalities).

Three Aspects of Regulation

The government does many things which can affect safety, from activities designed to provide accurate injury statistics (e.g., Web-based Injury Statistics Query and Reporting System), to policies requiring the recall of defective products. A common method by which government tries to increase safety is through mandatory regulations. The three main aspects of regulations are (a) the rules themselves, (b) monitoring for compliance (e.g., inspection), and (c) enforcement and penalties for noncompliance.

The discussion that follows concerns a single piece of legislation, the US Brady Law. That law provides for oversight of federally licensed firearms dealers by the US Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF). The primary purpose of the law is to limit the availability of guns to criminals. Virtually every gun sold in the USA is first made by a licensed gun manufacturer and then sold by a licensed gun dealer. The supply of guns to US criminals, unlike the supply of narcotics,

rarely begins in clandestine factories or illegal smuggling. The Brady Law is discussed here both to illustrate the three levels of regulatory intervention and to highlight problems that can result when there are weaknesses at any of the three levels.

The Rules

Before the Brady law, it was already illegal for felons to buy guns. However, the law was difficult to enforce, because felons would simply lie about their criminal histories. The Brady law was designed to eliminate this common “lie and buy” practice.

The Brady law is aimed directly at the licensed dealer, rather than the gun user. The law requires licensed dealers to check the background of each prospective purchaser and not sell firearms to anyone who does not pass legal muster. Those prohibited from buying guns include, among others, convicted felons, illegal aliens, people addicted to controlled substances, people with domestic violence convictions or restraining orders, and people who have been involuntarily committed to a mental institution or adjudicated mentally defective. Individuals on the terrorist watch list are not prohibited from buying guns legally from licensed dealers, nor are individuals convicted of most violent misdemeanors. However, the largest loophole in the law is that no background check is required for guns sold by someone other than a licensed dealer. These “secondary gun transfers” include sales by private individuals (whether at gun shows, home, flea markets, on the internet, or in want ads) and sales by federally licensed dealers from their private collections.

Monitoring

A problem with implementation of the Brady law is that many states either do not supply certain types of data to the National Instant Criminal Background Check System (e.g., involuntary commitments, restraining orders) or do a poor job updating the conviction information (Henigan 2009). Thus many people with serious mental health problems and other disqualifying histories who should be prohibited from obtaining a firearm can readily obtain guns from licensed firearm dealers.

A problem with the oversight of dealers is that the ATF does not have enough resources to effectively monitor the tens of thousands of licensed dealers throughout the nation. For example, the Justice Department’s inspector general reported in 2004 that “most (licensed dealers) are inspected infrequently or not at all” and that with ATF’s limited manpower, “it would take more than 22 years to inspect all (the licensed dealers)” (U.S. Department of Justice 2004). The ATF is also limited by law to a single unannounced inspection every 12 months, and ATF agents are prohibited from posing as felons in undercover operations (a tactic commonly used by drug agents). Using convicted felons in such stings makes prosecution difficult because juries are reluctant to believe the testimony of a convicted felon.

Enforcement

The penalties for violation of the law are minimal. Serious violations by gun dealers are typically misdemeanors rather than felonies, the prosecution has to prove “willful” violation of the law, and the end result even of successful prosecution is often that the dealer will only lose his dealer’s license. Then he often simply transfers the license to his spouse, father, or brother.

Unfortunately, self-regulation has little chance of working in this area. For example, firearm manufacturers, who could be expected to sanction “bad apple” dealers, have abrogated any and all responsibility. Indeed, the manufacturers lobbied successfully for a national law that ensures that they cannot be held liable for lack of oversight concerning the sales of their lethal product. And the one manufacturer that tried to write a code of ethics for its dealers was severely punished by the rest of the industry, forcing its CEO to resign (Henigan 2009).

The end result of the loopholes in the law, and the constraints on ATF monitoring and enforcement, is that what should be a substantial wall separating the legitimate sale of firearms and their diversion to criminal misuse is actually very porous and felons have little difficulty obtaining firearms. ATF gun-trafficking investigations find that corrupt dealers are the largest source of guns trafficked to criminals; the second largest source is gun shows. Sting operations conducted by city and state agencies demonstrate the continuing ease with which felons can obtain firearms directly from licensed dealers (Hemenway 2006).

The Public Health Approach to Policy

The public health policy approach to injury prevention differs substantially from the approach taken by the fields of medicine and criminal justice, two of the other fields with a strong stake in public safety. First, the vast majority of resources for both medicine and law enforcement are spent case by case. Medicine’s principal focus is curing the individual patient, one patient at a time. Similarly, law enforcement seeks to apprehend and punish those committing crimes, one perpetrator at a time. By contrast, the primary concern of public health is with the health and safety of populations rather than individuals. This focus on the health of society is summed up in the motto of the Johns Hopkins School of Public Health – “Protecting Health, Saving Lives – *Millions at a Time.*”

Second, while medicine cares about *preventing* disease and criminal justice hopes to deter crime, almost all the resources are spent, and the activity takes place, after the fact, after someone is sick or injured (medicine), or after someone has broken the law (criminal justice). By contrast, the full emphasis of public health is on prevention, preventing bad things from happening in the first place.

Two other aspects of the public health approach deserve mention. One aspect is that, unlike criminal justice, public health has no interest in finding fault, assigning blame, or punishing people – except to the extent that this may increase prevention. A second aspect is that most empirical investigations have found that the most cost-effective way of preventing unintentional injury and violence is a systems approach – trying to create a culture and an environment that promotes safety and good behavior, rather than trying directly to change the behavior of each individual, or to blame them when they make mistakes or behave inappropriately.

The decisions and “policies” of many groups and institutions affect safety, including foundations, standards writing organizations, product certifiers, trade associations, the faith community, and the media. Public health success has often been due to its ability to attract and mobilize the efforts of many of these groups, such as physicians, women’s and youth organization, civil rights groups, and consumer organizations. The focus of this chapter is on governmental policy, rather than the policies of these myriad other organizations. The public health policy approach is illustrated below with respect to the problems of suicide.

Suicide

When making presentations, or just talking to groups of psychiatrists, I often ask why they think there is so much more suicide in Arizona than in Massachusetts. Their response is usually not an

answer but something along the lines of “that’s interesting, we didn’t know that.” Nor should they. Their interest is in helping each of their patients, one at a time. By contrast, public health is not very interested in why any particular person is suicidal, but why some groups have such high suicide rates, and how to reduce the suicide rates everywhere.

The first step in the public health approach is to create a “surveillance” system that provides rich, contextual information consistently and comparably across areas, and over time. In recent years, public health professionals have been partially successful in developing a National Violent Death Reporting System which, for the first time, provides detailed surveillance data on suicide (Hemenway et al. 2009). The system assembles information from four existing sources: death certificates, police reports, medical examiner/coroner reports, and crime labs. Data are now available, across 18 states, and over time, on issues such as (a) where teens who commit gun suicide obtain their firearms (Johnson et al. 2010), the type of gun used, and whether alcohol was involved; (b) whether female veterans are at higher risk for suicide than nonveterans (Kaplan et al. 2009); and (c) whether accidental gun deaths typically occur at home or away from home, inside or outside, with long guns or handguns, and whether the wound was self-inflicted or other-inflicted (Hemenway et al. 2010).

Other steps in the public health approach include risk identification, development and implementation of evidence-based interventions, and evaluation of the effectiveness of these interventions. Interventions focus on three areas: the human user, the agent of injury, and the environment. Successes have been achieved in all three areas. Focusing on the human and the environment, the US Air Force suicide prevention program reduced the stigma associated with seeking help for social and psychological problems, leading to more personnel accessing mental health and social support services, and a reduction in suicides of 33%. (Knox et al. 2003).

Many suicide prevention successes have focused on the agent of injury. Britain effectively, if inadvertently, eliminated its most common means of suicide (putting one’s head in the oven) when it eliminated carbon monoxide from its gas, with little change in other methods of suicide (Hawton 2005); restrictions on the import and sale of the most lethal pesticides led to marked reduction in suicide in Sri Lanka (Gunnell et al. 2007); and the Israeli Defense Force changed their policies regarding soldiers’ access to firearms, and reduced the young adult suicide rate by 40% (Lubin et al. 2010).

Finally, successes have come from changing the environment. The serious anti-alcohol campaign of Perestroika cut alcohol consumption in half in Russia and helped reduce Russian male suicides by 44%. Perestroika has been called “history’s most effective suicide prevention program for men” (Wasserman 2001, 254).

Conclusion

Free markets can provide strong incentives for safety. Consider airline safety. Air crashes are newsworthy. Airlines, and aircraft manufacturers, will lose customers if their safety comes into question, insurance companies will raise premiums, and many of the families affected will bring tort liability suits. Evidence also shows that airline stock prices fall after major catastrophes. The question becomes whether these market incentives are sufficient to achieve optimal levels of safety. The decision to have safety regulated by an independent agency (the Federal Aviation Administration) suggests that many believed it was not.

Wealth provides many safety benefits. Richer individuals, and richer communities and societies, typically have goods of higher quality and levels of safety – safer cars, safer roads, safer homes, and safer workplaces. A major reason that injuries have fallen over the last century in the USA is that we have become richer. A permanent rise in the standard of living is one of the most effective ways to reduce injury. A concern about overly restrictive governmental regulations, even safety regulations, is that if they reduce real incomes, they may extract a toll in terms of an increase in injuries.

The rationale for government involvement is “market failure” – that one or more of the assumptions of the perfectly competitive model are invalid. For example, final goods consumers rarely have excellent information about the products they buy, they have difficulty behaving completely rationally, and their choices often affect third parties. Government intervention may help improve outcomes, such as by reducing injury and violence.

So much of what the government does affects safety, and the government does so much. For example, government creates data systems, subsidizes research, sets patent and liability rules, forces the recall of defective products, and builds many public environments, from roads to government buildings. It can also create mandatory regulations, such as requiring that all new cars have passive restraint systems. Three aspects of such regulations are the rules themselves, the monitoring for compliance with the rules, and the sanctions for noncompliance. All three aspects affect the overall level of safety actually achieved.

Finally, public health has a policy approach to injury and violence prevention, very different from either the medical or criminal justice approach. One of the important lessons public health has learned from attempting to prevent injuries is the importance of a systems approach – trying to change the agent of injury and the environment, rather than focusing exclusively on the individuals directly involved in the injury. The key to most successful injury and violence prevention policies is to create a system that (a) makes it difficult to make mistakes or behave inappropriately, so that fewer people will do so; and when a few people continue to make mistakes or to behave inappropriately, the system helps ensure that (b) no one is seriously hurt or injured.

References

- Durrant, J. E. (1996). The Swedish ban on corporal punishment: its history and effects. In D. Frehsee, W. Horn, & K.-D. Bussman (Eds.), *Family violence against children: a challenge for society*. New York: Walter de Gruyter.
- Erdmann, T. C., Feldman, K. W., Rivara, F. P., Heimbach, D. M., & Wall, H. A. (1991). Tap water burn prevention: the effect of legislation. *Pediatrics*, *88*, 572–577.
- Gunnell, D., Fernando, R., Hewagama, M., Priyangika, W. D., Konradsen, F., & Eddleston, M. (2007). The impact of pesticide regulations on suicide in Sri Lanka. *International Journal of Epidemiology*, *36*, 1235–1242.
- Hawton, K. (2005). Restriction of access to methods of suicide as a means of suicide prevention. In K. Hawton (Ed.), *Prevention and Treatment of Suicidal Behavior*. New York: Oxford University Press.
- Hemenway, D. (2006). *Private guns public health*. Ann Arbor, MI: University of Michigan Press.
- Hemenway, D. (2009). *While we were sleeping: success stories in injury and violence prevention*. Berkeley, CA: University of California Press.
- Hemenway, D., Barber, C. W., Gallagher, S. S., & Azrael, D. R. (2009). Creating a National Violent Death Reporting System: a successful beginning. *American Journal of Preventive Medicine*, *37*, 68–71.
- Hemenway, D., Barber, C., & Miller, M. (2010). Unintentional firearm deaths: a comparison of other-inflicted and self-inflicted shootings. *Accident; Analysis and Prevention*, *42*, 1184–1188.
- Henigan, D. A. (2009). *Lethal logic*. Dulles, VA: Potomac Books.
- Howard, A. W., MacArthur, C., Willan, A., Rothman, L., Moses-McKeag, A., & MacPherson, A. K. (2005). The effect of safer play equipment on playground injury rates among school children. *Canadian Medical Association Journal*, *172*, 1443–1446.
- Johnson, R. J., & Ettingler, C. F. (1982). Alpine ski injuries: changes through the years. *Clinics in Sports Medicine*, *1*, 181–197.
- Johnson, R. M., Barber, C., Azrael, D., Clark, D. E., & Hemenway, D. (2010). Who are the owners of firearms used in adolescent suicides. *Suicide & Life-Threatening Behavior*, *40*, 609–611.
- Kahane, C. J. (1981). *An evaluation of federal motor vehicle safety standards for passenger car steering assemblies*. Washington, DC: National Traffic Safety Administration.
- Kalbfleisch, J., & Rivara, F. (1989). Principles in injury control: lessons to be learned from child safety seats. *Pediatric Emergency Care*, *5*, 131–134.
- Kaplan, M. S., McFarland, B. H., & Huguet, N. (2009). Firearm suicide among veterans in the general population: findings from the national violent death reporting system. *The Journal of Trauma*, *67*, 503–507.

- Knox, K. L., Litts, D. A., Talcott, G. W., Feig, J. C., & Caine, E. D. (2003). Risk of suicide and related adverse outcomes after exposure to a suicide prevention program in the U.S. Air Force: cohort study. *BMJ*, *327*, 1376–1381.
- Lubin, G., Werbeloff, N., Halperin, D., Shmushkevitch M., Weiser, M., & Knobler, H. Y. (2010). Decrease in suicide rates after a change of policy reducing access to firearms in adolescents: a naturalistic epidemiological study. *Suicide and Life Threatening Behavior*, *40*, 421–424.
- Matter, P., Ziegler, W. J., & Holzach, P. (1987). Skiing injuries in the past 15 years. *Journal of Sports Sciences*, *5*, 319–326.
- Myers, M. L., & McGlothlin, J. D. (1996). Matchmakers' 'phossy jaw' eradicated. *American Industrial Hygiene Association Journal*, *57*, 330–333.
- Nagata, T., Setoguchi, S., Hemenway, D., & Perry, M. J. (2008). Effectiveness of a law to reduce alcohol-impaired driving in Japan. *Injury Prevention*, *14*, 19–23.
- National Fire Protection Association. (2005). *NFPA Journal*. Columns May/June 2005. <http://www.nfpa.org/public-Column.asp?categoryID=992&itemID=24265&src=NFPAJournal>. Accessed January 2011.
- Shope, J. T. (2007). Graduated driver licensing: review of evaluation results since 2002. *Journal of Safety Research*, *38*, 165–175.
- Stephoe, A., Wardle, J., Fuller, R., Davidsdottir, S., Davou, B., & Justo, J. (2002). Seatbelt use, attitudes, and changes in legislation: an international study. *American Journal of Preventive Medicine*, *23*, 254–259.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: improving decisions about health, wealth and happiness*. New Haven, CT: Yale University Press.
- U.S. Department of Justice, Office of the Inspector General. (2004). *Inspections of Firearms Dealers by the Bureau of Alcohol, Tobacco, Firearms and Explosives*, Washington, DC.
- U.S. Environmental Protection Agency. (2000). *America's children and the environment: a first view of available measures*. Office of Children's Health Protection, Washington, DC. [http://yosemite.epa.gov/ee/epa/erm.nsf/vwAN/EE-0438-01.pdf/\\$file/EE-0438-01.pdf](http://yosemite.epa.gov/ee/epa/erm.nsf/vwAN/EE-0438-01.pdf/$file/EE-0438-01.pdf). Accessed 12 Jan 2011.
- Wasserman, D. (Ed.). (2001). *Suicide: an unnecessary death*. London: Martin Dunitz.
- Ziegert, K. A. (1983). The Swedish prohibition of corporal punishment: a preliminary report. *Journal of Marriage and the Family*, *45*, 917–926.

Chapter 29

Environmental Approach

Leon S. Robertson

Introduction

One of the most publicized injuries in history led to Princess Diana's death when her chauffeur-driven limo struck a concrete pillar in the Pont d'Alma underpass in Paris. Most news coverage placed the blame on photographers chasing the vehicle, the speed, the condition of the driver, and the alleged nonuse of seat belts by the princess. But an expert on injury prevention saw more: "On the day of the crash, from network TV coverage, it was immediately clear to me that the cement support pillars in the tunnel are a glaring major hazard. They are very close to the edge of the left lane. They rest on a raised median, approximately three feet in width, bounded by a low sloping curb immediately adjacent to the traffic flow. The curb offers no protection at all from a vehicle coming in at an acute angle. Looking at the crash site from different network camera angles shows that there are other pillars nearby that clearly bear the marks of other direct impacts, and there is also what appears to be old debris lying at the base of at least one of them. The pillars themselves are massive, solid cement structures parallel to the road, and positioned close together. Any vehicle losing control to the left will be at severe risk of hurtling the low sloping curb and smashing directly into the corner of one of the pillars, as Diana's limo did. ... Subsequent investigation indicates that 13 people had been killed in the Pont d'Alma tunnel during the preceding decade. Even in a high traffic area that is excessive. ... On further review of pictures, you can see small sections of the tunnel that have guardrail protection. Apparently the hazard of the columns was recognized at these locations, but for some reason, they were not applied to all the pillars" (Short 2002).

Injuries result from interactions of the human anatomy with environmental energy in excess of the resilience of persons exposed. Injury usually occurs from too much energy but also occurs from too little energy, for example, from lack of oxidation in the case of drowning and lack of warmth in the case of hypothermia. In a collision such as Princess Diana's limo, if the vehicle or environmental structure impacted is rigid, the energy of the moving vehicle is transferred to the occupants. Unfortunately, the term "energy" is widely used as a pseudoscientific New Age reference to claims of psychic energy (unmeasured) affecting physical phenomena. Here, the word refers to actual, measurable physical energy.

Mechanical energy is the most common cause of fatal and severe injury, mainly from motor vehicle collisions, falls, and gunshots. Fatal, disabling, and disfiguring burns result from fires

L.S. Robertson, PhD (✉)
Yale University, New Haven, CT, USA

Green Valley, AZ, USA
e-mail: nanlee252000@yahoo.com

(thermal and chemical energy) and explosions. A variety of chemical energy exchanges with human tissue occur in poisonings. Electrical and ionizing radiation energy accounts for far less injury than other forms of energy because it is usually shielded from human exposure, a prime example of environmental modification to prevent injury.

A slogan popular among injury researchers says, "Injuries are not accidents." While incidents of injury have myriad causes, characteristics of the physical environment, including product design, can be altered to greatly reduce injury severity and, in some cases, incidence. Since alterations of environments can be costly, identification of higher risk sites and product characteristics is essential to efficient injury prevention.

Usually, injuries are not random but occur disproportionately by who, when, where, or how people are injured. Assaults, homicides, and suicides are more or less intentional but, nevertheless, cluster in particular populations, geographic areas, and time periods, as do unintentional injuries. Efficient injury prevention requires the practical use of epidemiology to identify patterns or clusters to target environmental changes where they are most needed.

Causation Versus Prevention

Some epidemiologists have engaged in an unproductive debate about causation. Diseases and injuries are said to be the result of complex causal webs. Too often, that truth is taken to mean that attempts at amelioration are doomed to failure because of the complexity. The presence of many causes does not preclude prevention, often by simple methods. Causes can be classified as necessary, sufficient, or probabilistic conditions. Most "risk factors," a euphemism epidemiologists use in place of cause, are neither necessary nor sufficient to produce injury but increase the probability to a greater or lesser extent. Some, such as age and gender are not modifiable to reduce the risk. The key to prevention is to identify changeable necessary conditions for harmful outcomes. These may be some characteristic of the energy involved, the means by which the energy is conveyed to the injured, characteristics of the injured or other persons in the vicinity, or characteristics of the environment, such as lighting and the presence of guardrails.

A classic example of prevention based on attention to a necessary condition is the reduction in children's falls from windows in high-rise buildings in New York City. Health department epidemiologists examined who, when, where, and how children were falling from such buildings. They found that two-thirds of all fatal falls among children up to 5 years occurred when children fell through open windows. A necessary condition for a fall through a window opening is that the opening is large enough for a mobile child to breach. The Health department initiated a campaign to have residents and owners of high-rise buildings install a barrier with openings too narrow to allow children to go out the windows. Eventually, such barriers were required, and fatal falls to children in high-rise buildings declined from 30 to 50 per year to fewer than five (Bijur and Spiegel 1996).

What if the epidemiologists had emphasized the causal web of risk factors for child falls – parents who were drunk, on the phone, disciplining or diapering another child, or dozens of other possible distractions, or characteristics of the children (hyperactive, autistic)? Would a campaign to address these multiple factors have been as successful as window barriers?

Descriptive Epidemiology

Using relatively simple practical epidemiological methods, injuries can be studied in a manner similar to classic studies of infectious diseases. One of the earliest published accounts of such use to identify a source of disease occurred in London in 1854. John Snow, a physician, questioned the

prevailing hypothesis that cholera and other diseases were caused by miasma, an unwholesome atmosphere. During a cholera epidemic, he used a map of the city streets and marked the number of deaths at given addresses. The cases clustered near and on Broad Street where one of the pumps that provided water to the populace was located. No such clusters were seen around other pumps. The city council was persuaded to take the handle off the Broad Street pump and the epidemic abated. It was decades later that another necessary condition, the cholera bacillus, was discovered, often dwelling in water contaminated by fecal material (Evans 1993).

Environmental health officers of the US Indian Health Service and members of Native American communities have used pin maps of injury locations and other injury data to develop countermeasures. In some cases, the identified injuries were virtually eliminated when government authorities were persuaded to take action based on the data (Smith and Robertson 2000). Examples will suffice to illustrate the power of this approach.

Route 666 north of Gallup, New Mexico, the road to the Navajo Nation, was once described in a newspaper as the most dangerous road in America. As part of her work on an Indian Health Service injury control fellowship program, Nancy Bill obtained police reports of the deaths. She used a pin map to specify where the deaths occurred. Her findings showed that a substantial number of deaths on the road occurred to pedestrians in a 4-mile stretch of the road, an average of about three per year. The deaths occurred at night. Aware of research that showed lighting road sections reduced pedestrian deaths (e.g., Schwab et al. 1982), Ms. Bill persuaded the state to install overhead lights along the roadside. In the 5 years following the installation, there were no pedestrian deaths along that section of road.

In Humboldt County, California, David Short, the emergency medical coordinator for the Hoopa Nation, collected data on the number of deaths that occurred on three highways in the county. The deaths were found mainly when vehicles left the road and hurtled down steep embankments for up to 300 m. Following the installation of guardrail on selected sections of the roads, Mr. Short continued to collect data for 10 years. A comparison of the road section where guardrail was installed with the sections where no guardrail was installed showed that about two deaths per year were prevented by guardrail. In the 7 years before guardrail installation, about two deaths per year occurred at both the installation and noninstallation road sections. In the subsequent 10 years, there were no deaths in the installation road sections but nearly two per year at noninstallation sites. The study accounted for other factors such as changes in the state's belt use law and average daily traffic (Short and Robertson 1998).

Application of Known Principles to Environmental Design

Identification of injury clusters and identification of environmental modifications to prevent them is an after-the-fact, body-count approach. As new vehicles, commercial buildings, houses, swimming pools, roads, and other facilities are designed and built, principles gained from analytic research and knowledge of injury patterns in relation to hazards can be applied to prevent such clusters from occurring before the fact (Fig. 29.1).

Haddon explicated ten strategies for controlling hazards, strategies that are applicable to all hazards, not just those that contribute to injury. See the accompanying box for a list of the strategies and an example of the application of each. In the design process, if the designers were to acquaint themselves with the epidemiological data regarding types of injuries that occur in the environments they are designing and reviewed Haddon's strategies to think of options to reduce potential hazards, fewer severe injuries and deaths would occur in those newly designed environments.

In road environments, separation of slow-moving people and vehicles from motorized vehicles either by providing separate paths, by energy-absorbing barriers, or by time of day are alternative means of reducing the carnage (Berger and Mohan 1996). Timing of lights at intersections has a

1. Prevent the creation of the hazard in the first place. Do not install windows that open in high-rise buildings.
2. Reduce the amount of the hazard brought into being. Reduce the flammability of building materials and the release of toxic gases when they burn.
3. Prevent the release of the hazard that already exists. Include automatic sprinkler systems in building designs.
4. Modify the rate or spatial distribution of release of the hazard from its source. Sensors in dams and levees can be used to release water at a controlled rate to avoid dangerous buildup.
5. Separate, in time or space, the hazard and that which is to be protected. Use over passes at rail and road intersections.
6. Separate the hazard and that which is to be protected by interposition of a material barrier. Install barriers in road medians that redirect errant vehicles into the proper lane and energy absorbing guardrail or other materials to prevent vehicles from leaving roads at higher risk sites.
7. Modify relevant basic qualities of the hazard. Use energy absorbing materials on play ground surfaces.
8. Make what is to be protected more resistant to damage from the hazard. (Increasing human resistance is irrelevant to physical environments.)
9. Begin to counter the damage already done by the environmental hazard. Plan location of hospitals and other emergency services relative to areas of high severe injury frequency to minimize time of transportation of the injured.
10. Stabilize, repair, and rehabilitate the object of the damage. Plan rehabilitation in high disabling injury areas.

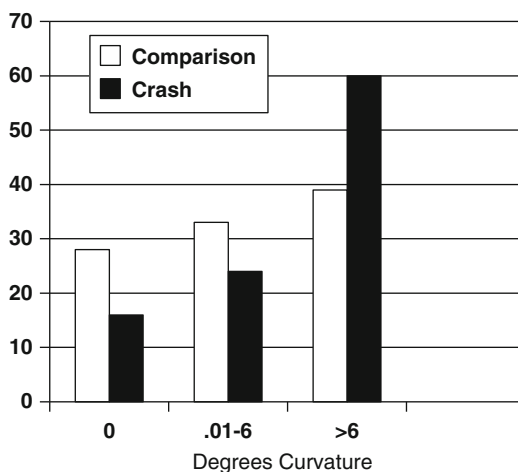
Fig. 29.1 Haddon's injury prevention strategies (Haddon 1970) with examples

substantial effect on collisions. Increasing the amber phase to conform to Institute of Transportation Engineers proposed standards reduced pedestrian injuries 37% (Retting et al. 2002). Various techniques have been employed to control vehicle speed, called "traffic calming," including exponentially spaced road stripes, rumble strips, speed bumps, and roundabouts (Retting et al. 2003; Elvik and Vaa 2004). Collisions of road vehicles and trains are fewer at sites that have automatically lowering gates and are nonexistent at sites where overpasses are used (Meeks and Robertson 1993). The choice of application of any specific site depends on the types of vehicles, road environments, and persons injured in a given cluster.

While it would be economically unfeasible to line every road with energy absorbing guardrail, barrels of sand or water, and the like, the pattern of injury in relation to road characteristics indicates road sections that are most likely to experience off-road excursions that result in severe injury. Research comparing road characteristics where fatal collisions occurred after a vehicle left the road with the same characteristics of control sites a mile back in the direction from whence the vehicle came found that curvature and grade can be used to identify higher risk sites (Wright and Robertson 1979). The logic is straightforward: the vehicle traversed the control site without incident but went off the road at the crash site. The driver and vehicle are the same at both sites, so what is different between the sites? The researchers measured curvature, grade, super elevation, and number of objects along the road at intervals approaching and leaving the crash and comparison sites (Fig. 29.2).

The figure shows the differences in curvature between case and comparison sites. More than half of the fatal crashes occurred when the vehicle was in proximity to curves of 6° or more. Less than

Fig. 29.2 Percent of sites by maximum degrees of roadway curvature within 152 m of comparison and crash sites



40% of comparison sites were near curves of 6° or more. When the characteristics were combined, 6° or greater curves on downhill grades of -2% or more were the most disproportionately involved in fatal crashes with fixed objects.

Using the same design, other researchers have found similar differences in the same road characteristics substantially distinguish the sites where drowning occurred to occupants of vehicles that ran off-road into water (Wintemute et al. 1990) or died when a vehicle rolled over after leaving the road (Zador et al. 1987). Clearly, road sections with 6° or greater curves are high-priority sites for energy absorbing guardrails or other materials to stop vehicle from exiting roads inadvertently.

Motor vehicle features that vary widely among specific makes and models are strongly correlated with death rates per number of vehicles in use. In a study of 27,615 deaths to occupants of cars, utility vehicles, and vans, or other road users (pedestrians, bicyclists, etc.) struck by them on US roads during 2000–2005, I calculated the reduction in deaths achievable by changing a given vehicle characteristic as other characteristics remained the same. If all vehicles were equipped with electronic stability control, the estimated death reduction would have been 11,098, about 42% of the total. If all of the vehicles had received the highest rating by the Insurance Institute for Highway Safety on their offset frontal crash tests, there would have been approximately 2,211 fewer deaths, 8.6% of the total. If the vehicles that had a poor showing on the US government's side crash tests were improved to the average, 4,950 (19.4%) deaths would have been prevented. Static stability (track width divided by half the center of gravity height) of 1.2 or higher among vehicles with lower stability would have prevented 2,737 deaths, 10.7% of the total deaths studied (Robertson 2007a).

The study of vehicle characteristics used the Fatality Analysis Reporting System (FARS) that includes virtually all of the fatal crashes in the USA. To specify vehicle and environmental conditions that can be changed to reduce injury, it is necessary that data specifying vehicles and environments in which injuries occur be included in data systems. FARS contains much such data. While crash test results and vehicle equipment are not included, the inclusion of make and model of vehicle gives researchers the information needed to augment the data from crash test and vehicle characteristic files.

In a project for the Indian Health Service, I developed a set of data forms that environmental health officers use to collect data on specific types of injury – one form for each incident (Robertson 2007b). In addition to information on who, where, and how people are injured, the forms include a

list of possible countermeasures that would have prevented the injury or reduced severity. If used systematically in local areas, it can be used to set priorities for environmental modifications.

Using the form for fall injury, Gina Locklear observed that severe falls to elderly Apaches in the vicinity of Fort Apache were often on porch steps. She compared porch steps where people fell and experienced severe injury to a random sample of houses with residents matched by age and gender. The steps where residents fell had significantly different dimensions as well as being less level and less well lit than comparison stairs (Locklear 1990). Such information can be used to set standards for new construction and rehabilitation of existing stairs. Also, the study design can be used to study other types of falls.

Drowning of young children in home swimming pools is a major problem in warmer and more affluent areas. A comparison of swimming pool drowning in Honolulu, Hawaii and Brisbane, Australia suggests that Honolulu's laws requiring fences around pools with gates that cannot be opened by young children reduces the problem by about 65%. The two cities were similar in size and climate and had similar pool-to-household ratios (Pearn et al. 1979).

Cigarettes left to burn or dropped on flammable surfaces are the primary cause of fatal house fires. While the problem has been reduced in jurisdictions with strong antismoking laws, an effective approach for the remainder of smokers is to allow only cigarettes that self-extinguish (Technical Study Group 1987). Another technology to reduce smoke- and fire-related injury is the installation of automatic-sprinkling systems that activate when fire is detected (http://www.usfa.dhs.gov/citizens/all_citizens/home_fire_prev/sprinklers/). These add about 1% to the cost of a new home, and that cost is offset substantially in time by insurance discounts. In addition to reducing risk to the most vulnerable (children, the elderly, and disabled), automatic sprinklers reduce damage from fire-fighting (chopped holes in walls, high-velocity water hoses).

Even violence thought or known to be intentional can be reduced by attention to weapon and environmental design. Spurred by legislation in New Jersey requiring technology that will not allow anyone but an owner to fire a gun when technology built into the gun identifies the owner, inventors accelerated work on such technology (Allbusiness.com 2003). The technology has the potential to reduce gun use by children or when stolen or by suicidal teenagers, angry spouses, or other house-mates of gunowners.

Violence is often concentrated in certain neighborhoods, kinds of buildings, and businesses. Numerous proposals for design of environments to discourage crime and associated violence have been proposed and some studied (Jeffery 1977; Crowe 2000; Mair and Mair 2003). Bulletproof windows for cashiers and interior visibility from the outside for convenience stores, gas stations, banks, and other high-risk buildings are examples.

Using global positioning technology, it is possible to specify location of injury. Mapping of home injury using the technology indicates geographic areas where such injuries are more common (Chong and Mitchell 2009). Adding such information to data sets of home and other types of injury could be useful in locating areas where resources should be concentrated.

Costs, Unintended Consequences, and Trade-Offs

Opposition to legal and regulatory approaches that require environmental changes for injury prevention centers on costs relative to benefits, unintended consequences such as behavioral adaptations to reduced risk, and trade-offs, such as relative risks from various forms of construction, manufacturing, and transportation. Most people would agree that the most economically efficient

means of injury reduction should be employed and that greater risks should not be substituted for those that exist. In such analyses, it is important to be sure that competing approaches are addressing the same injuries.

The argument that people have the equivalent of a risk thermostat in their heads and will behave more riskily if their protection is increased has been largely discredited. There may be instances in which people have good knowledge of risks and decide to forego protection for competitive advantage, but studies of vehicle safety standards produce no credible evidence for such behavior in the general population. Studies claiming that drivers in safer vehicles drove more riskily and killed more pedestrians depended on aggregated data that, when disaggregated, showed no such effect (Robertson 2007b). Contrary to the theory, drivers using seat belts do not speed or run red lights or turn left in front of oncoming traffic more often than those that do not. Such behavior of belted and unbelted drivers was observed in a jurisdiction where belt use changed from 17% to 77% due to enactment of a belt use law and in a jurisdiction where the law did not change (Lund and Zador 1984).

There are trade-offs that should be considered in planning. For example, almost every hospital with an emergency service has or wants a helicopter to transport patients from injury sites. Crashes of such helicopters kill not only those already injured but emergency personnel and pilots as well, particularly at night and during adverse weather conditions (Baker et al. 2006). Of course, patients and others die in crashes of land vehicles as well. To assess the possible trade-off, data on fatalities in the two forms of transportation under similar conditions is needed.

In many instances of cost-effectiveness comparisons, such as comparisons of light rail versus auto transportation, the benefits are stated entirely in terms of time of travel without consideration of relative risk of injury. Operation of motor vehicles at speeds substantially beyond those achievable by walking or animate conveyors (e.g., camels, horses, oxen) requires smooth roads to accommodate them. Countries with limited road systems have the choice of building alternatives such as tracks for mass transit of people and goods among population, farming, and business hubs, with local roads only for taxis, rental vehicles, busses, and trucks to distribute passengers and goods to and from local points. Countries with highly developed road systems could use the roadbeds to lay tracks for a similar result. Once the car culture is predominant, however, dismantling it will likely occur only when the cost of operating motor vehicles is out of reach of a vast majority of the population. The ecological and economic consequences of reliance on motor vehicles go far beyond their impact on injury rates (Broome 2008).

In the United States, road construction has a huge lobby, called the “road gang” by its opponents (<http://www.highbeam.com/doc/1G1-54754434.html>, accessed June, 2010). Corporations that produce gravel, asphalt, and concrete for roadbeds join with steel manufacturers who build bridge superstructure, construction companies, vehicle enthusiast groups, and vehicle manufacturers to maximize the allocation of public tax monies to road construction and maintenance (Kelley 1971). Other forms of ground transportation (e.g., rail and bicycles) have their corporate interests and enthusiasts, but they are poorly financed and receive a pittance, compared to the road interests, in subsidies by federal, state, and local governing bodies compared to the road interests.

Costs and benefits of various changes in transportation infrastructure are often measured only in travel times with little or no consideration of effects on safety, air pollution, and depletion of fuel supplies. The latter is relevant to the cost side as well because increasing gasoline prices as the world’s oil supply diminishes are not factored into the equation. Given these limitations, it is not surprising that light rail appears in these studies as less cost-beneficial than other approaches such as special road lanes for cars and light trucks based on vehicle occupancy or purchased licenses to drive in the fast lane (e.g., http://www.azdot.gov/TPD/ATRC/Publications/project_reports/PDF/AZ582.pdf, accessed June, 2010).

Conclusion

It is obvious from this discussion that injury reductions depend on interdisciplinary work by a variety of academic, professional, and executive participants. The orientations of different professionals such as engineers, architects, lawyers, and epidemiologists are sometimes in conflict. One of the strengths of university-based injury prevention centers is the development of interdisciplinary perspectives on injury reduction (Winston et al. 1996).

Action by government (local, state or provincial, and national) and businesses is necessary to implement many of the environmental changes identified by professionals as efficacious to reduce injury severity. The goal of corporations to maximize profits is often in conflict with the goal of injury prevention if the environmental modifications add to costs of doing business. Governmental regulatory regimes and even the rules of governance of corporations must be under ongoing scrutiny to resolve such conflicts (Wiist 2010).

References

- Allbusiness.com. (2003). <http://www.allbusiness.com/crime-law/crime-prevention-gun-control/5752432-1.html>. Accessed June 2010.
- Baker, S. P., Grabowski, J. G., Dodd, R. S., Shanahan, D. F., Lamb, M. W., & Li, G. H. (2006). EMS helicopter crashes: what influences fatal outcome? *Annals of Emergency Medicine*, *47*, 351–356.
- Berger, L. R., & Mohan, D. (1996). *Injury control: a global view*. Oxford: Oxford University Press.
- Bijur, P. E., & Spiegel, C. (1996). Window fall prevention and fire safety: 20 years of experience in New York City. *Pediatric Research*, *39*, 102A.
- Broome, J. (2008). The ethics of climate change. *Scientific American*, *298*, 96–102.
- Chong, S., & Mitchell, R. (2009). The use of mapping to identify priority areas for the prevention of home injuries. *International Journal of Injury Control & Safety Promotion*, *16*, 35–40.
- Crowe, T. D. (2000). *Crime prevention through environmental design*. Louisville, KY: National Crime Prevention Institute.
- Elvik, R., & Vaa, T. (2004). *The handbook of road safety measures*. Amsterdam: Elsevier.
- Evans, A. S. (1993). *Causation and disease: a chronological journey*. New York: Plenum.
- Haddon, W., Jr. (1970). On the escape of tigers. *Technology Review*, *72*, 44.
- Jeffery, C. R. (1977). *Crime prevention through environmental design*. Beverly Hills, CA: Sage.
- Kelley, A. B. (1971). *The pavers and the paved*. New York: Donald W. Brown.
- Locklear, G. (1990). A retrospective case-control study of porch step falls on the Fort Apache Indian Reservation. <http://www.injprev.ihs.gov/documents/GLocklear.pdf>. Accessed June 2010.
- Lund, A. K., & Zador, P. F. (1984). Mandatory belt use and driver risk taking. *Risk Analysis*, *4*, 41–53.
- Mair, J. S., & Mair, M. (2003). Violence prevention and control through environmental modifications. *Annual Review of Public Health*, *24*, 209–225.
- Meeks, K. D., & Robertson, L. S. (1993). Study of road-rail crashes in Claremore, OK and allocation of resources for preventive measures. *Public Health Reports*, *108*, 248–251.
- Pearn, J. H., Wong, R. K., Brown, J., Bart, R., & Hammar, S. (1979). Drowning and near drowning involving children: a five-year total population study from the city and county of Honolulu. *American Journal of Public Health*, *69*, 450–454.
- Retting, R. A., Chapline, J. F., & Williams, A. F. (2002). Changes in crash risk following retiming of traffic signal change intervals. *Accident Analysis and Prevention*, *34*, 215–220.
- Retting, R. A., Ferguson, S. A., & McCart, A. T. (2003). A review of evidence traffic engineering measures designed to reduce pedestrian-motor vehicle crashes. *American Journal of Public Health*, *93*, 1456–1463.
- Robertson, L. S. (2007a). Prevention of motor-vehicle deaths by changing vehicle factors. *Injury Prevention*, *13*, 307–310.
- Robertson, L. S. (2007b). *Injury epidemiology*. New York: Oxford University Press.
- Schwab, R. N., Walton, N. E., Mounce, J. M., & Rosenbaum, M. J. (1982). Highway lighting. In *Synthesis of safety research related to traffic control and roadway elements* (Vol. 2) (Report No. FHWA-TS-82-233). Washington, DC: Federal Highway Administration.

- Short, D. (2002). Quickguide to effective injury prevention: saving lives with proactive emergency medical services. <http://www.nanlee.net/ems/quickguide.htm>. Accessed June 2010.
- Short, D., & Robertson, L. S. (1998). Motor vehicle death reductions from guardrail installations. *Journal of Transportation Engineering*, *124*, 501–502.
- Smith, R. J., & Robertson, L. S. (2000). Unintentional injuries and trauma. In E. R. Rhoades (Ed.), *American Indian Health*. Baltimore, MD: The Johns Hopkins University Press.
- Technical Study Group. (1987). *Toward a less fire-prone cigarette*. Washington, DC: US Consumer Product Safety Commission.
- Wiist, W. H. (2010). *The bottom line or public health: tactics corporations use to influence health and health policy, and what we can do to counter them*. New York: Oxford University Press.
- Winston, F. K., Schwarz, D. F., & Baker, S. P. (1996). Biomechanical epidemiology: a new approach to injury control research. *The Journal of Trauma: Injury, Infection, and Critical Care*, *40*, 820–824.
- Wintemute, G. J., Kraus, J. F., Teret, S. P., & Wright, M. (1990). Death resulting from motor vehicle immersions. *American Journal of Public Health*, *80*, 1068–1070.
- Wright, P. H., & Robertson, L. S. (1979). Amelioration of roadside obstacle crashes. *Transportation Engineering Journal*, *105*(TE6), 609–662.
- Zador, P. L., Stein, H. S., Hall, J. W., & Wright, P. H. (1987). Relationships between vertical and horizontal roadway alignments and the incidence of fatal rollover crashes in New Mexico and Georgia. *Transportation Research Record*, *1111*, 27–42.

Chapter 30

Technological Approach

Flaura K. Winston, Kristy B. Arbogast, and Joseph Kianianthra

Introduction

For the first 50 years of motor vehicle traffic safety, major emphasis was placed on trying to protect occupants in crashes through improvements in braking, lighting, tires, and other components along with limited attempts to educate drivers on the need for changing individual driver behavior (Bonnie et al. 1999). In the 1940s, pioneering work in aviation safety by De Haven (1942) demonstrated that serious injuries could be prevented by managing how the crash was experienced by the human, in particular by managing deceleration. His discoveries led to energy-absorbing compartments and safety belts in aircraft that provided dramatic safety benefits. In 1950, De Haven developed the concept of “packaging” car occupants to prevent them from being tossed against injurious surfaces in a crash.

In the 1950s, Colonel John Stapp demonstrated that humans could withstand rapid deceleration if properly restrained by safety belts. Under well-controlled conditions, Stapp (1955, 1957) was strapped to a rocket sled, subjected himself to deceleration up to 40 G (40 times the force of gravity), and suffered no permanent injury. Stapp’s studies and others involving “well-packaged” human volunteers subjected to rapid deceleration of translational (linear) motion were important in the development of the vehicle safety belt.

In the 1970s, Dr. William Haddon, Jr., the first administrator of the National Highway Safety Bureau, the predecessor agency of the National Highway Traffic Safety Administration (NHTSA), introduced a conceptual model for expanding traffic injury prevention beyond the role of the driver

F.K. Winston, MD, PhD (✉)

Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania,
11th floor, 3535 Market Street, Philadelphia, PA 19104, USA

Center for Injury Research and Prevention, The Children’s Hospital of Philadelphia, Philadelphia, PA, USA
e-mail: flaura@mail.med.upenn.edu

K.B. Arbogast, PhD

Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
e-mail: ARBOGAST@email.chop.edu

J. Kianianthra, PhD

Active Safety Engineering, LLC, 20763 Crescent Pointe Place, Ashburn, VA 20147, USA
e-mail: jkianianthra@asafetyeng.com



Fig. 30.1 The Haddon Matrix: a discrete approach for examining the components of a crash and identifying areas of potential intervention for fatality and injury reduction (*source: NHTSA, 2003*)

to encompass the increasing role of vehicle safety engineering and policy. Patterned after the public health approach to prevention, the components of the Haddon Matrix (Fig. 30.1) deconstructed a crash into discrete causal steps in a sequence and allowed exploration of the contributions of the human, vehicle, and environment (Haddon 1972, 1980). Since its introduction, the Haddon Matrix has served as a guide for safety research and countermeasures and has paved the way for not only educational programs, policies, and laws but also technological advances in product development, testing, and regulations.

Today, engineers work in concert with behavioral scientists and epidemiologists to create a comprehensive view of injury and its mitigation (Winston et al. 1996). Epidemiologists define the magnitude of the hazard, identify risk factors for injury, and evaluate the effectiveness of interventions. Behavioral scientists examine the contribution of the human and social contexts to the injury. Biomechanical engineers apply the laws of physics and other engineering principles to systematically determine injury causation and define technology's role in mitigation to allow for new advances (e.g., products, safety standards, and test procedures). Recognizing that even the most diligent and careful humans can be involved in situations that result in motor vehicle crashes, engineers attempt to reduce the chance of crashes and their impact on the human body through thoughtfully designed environments, vehicles, and occupant protection systems.

Traditional Approach to Safety Engineering

The traditional safety engineering approach is systematic and iterative, involving design, modeling, experimentation, evaluation, and revision in order to find an optimal safety solution. Steps in this approach typically include (1) crash or event investigation, (2) physical or computational modeling

to replicate the injury mechanism and thus validate these models, (3) development of safety countermeasures, and (4) testing and revision of the countermeasures using the models until an optimal safety solution is achieved. To design and test solutions, the engineer does not work in isolation but rather partners with other experts, such as human factors and behavioral scientists, who focus on the acceptance and usability of technology by humans, and epidemiologists and statisticians who measure the effectiveness and safety of the countermeasures in the real world.

Crash or Event Investigation

At the core of the engineering approach is failure analysis: understanding (or envisioning) how a crash and the resultant injuries occur. Starting in the real world, a team of engineers collect detailed information from the scene of the crash, the vehicles, the treating emergency medical and hospital personnel, the family, and medical and other records. Interdisciplinary teams interpret these data, bringing to bear knowledge of traffic engineering, human factors, injury biomechanics, structural mechanics, and medicine, in order to piece together the modifiable causes of the crash and the resultant injuries. This approach allows the team to better understand how injuries are caused, forming the basis for innovations and recommendations to reduce the risk of injury to other drivers and occupants in similar situations. Regular joint review of cases by researchers, vehicle manufacturers, restraint suppliers, and regulators helps to set evidence-based priorities for future research, identifies necessary safety design modifications and innovations, and informs the development of anthropomorphic test devices (ATDs) and test procedures.

The majority of traffic safety engineering and regulatory efforts in the USA rely on the National Automotive Sampling System (NASS) and the Fatality Analysis Reporting System (FARS). These data sources have been augmented by interdisciplinary crash investigation review as exemplified by NHTSA's Crash Injury Research and Engineering Network (CIREN). In an effort to integrate crash data collected by engineers and crash reconstructionists with detailed injury and radiological information collected by clinical teams at hospitals, NHTSA began funding hospital-related studies in the 1980s and initiated the Highway Traffic Injuries Studies, the forerunner of CIREN, in 1991. Over the next several years, research projects were funded at four Level I Trauma Centers to collect detailed injury information on motor vehicle occupants involved in crashes. In 1996, with funding from a settlement agreement with General Motors (NHTSA Notes 1996), CIREN was created to integrate these efforts into a uniform centralized data system and three additional Level I Trauma Hospitals were added.

Today, CIREN is a sponsor-led multicenter research program involving a collaboration of clinicians and engineers in academia, industry, and government pursuing in-depth studies of crashes, injuries, and treatments to improve processes and outcomes. Its mission is to improve the prevention, treatment, and rehabilitation of motor vehicle crash injuries and to reduce deaths, disabilities, and human and economic costs. Hospital-based surveillance identifies injured drivers and their occupants. Once a subject meets the criteria for enrollment and consents to participate in the study, an in-depth investigation is initiated to collect detailed crash and injury information. Traumatology experts review the occupant's radiology and clinical data for the location and type of the injury. Crash investigators investigate the involved vehicles and the crash scene to determine the severity of the impact and the physical evidence of occupants' interaction within the crash environment. Mechanical and biomechanical engineers experienced in the field of crash testing and biomechanics research evaluate each case to determine the role of the vehicle's design and the level of interaction with the occupant. Together, these multidisciplinary teams of engineers and clinicians review the cases to assess injury causation scenarios and injury mechanisms.

This approach has delineated the mechanism of many types of crash-related injuries. For example, teams comprised of the described specialists were the first to investigate child air bag-related fatalities. Frontal air bags were designed as a restraint system to supplement the vehicle seat belt and reduce the risk of head, face, and brain injuries. In 1995, soon after the introduction of frontal air bags into the automobile fleet, a 20-day-old child arrived fatally injured in The Children's Hospital of Philadelphia Emergency Department. A team of investigators from NHTSA and the National Transportation Safety Board were dispatched to the crash scene, and an interdisciplinary investigation revealed that the infant, seated in a rear-facing child restraint system in the right front seat, was the first reported child air bag-associated fatality (Hollands et al. 1996a). Over the next several years, over 100 more deaths and injuries to children and small-statured women in motor vehicle crashes would be attributed to exposure to deploying passenger air bags (Winston and Reed 1996).

In November 1995, the Morbidity and Mortality Weekly Report issued by the US Centers for Disease Control and Prevention described eight deaths of child occupants involving air bag deployment that were of special concern because they involved low-speed crashes in which the children otherwise might have survived (CDC 1995). The risk to small occupants from a deploying air bag had been a concern for the automotive industry for several years (Kent et al. 2005; Mertz 1988). As passenger air bags diffused into the market, numerous case reports began appearing in the medical literature describing brain and skull injuries sustained by children in rear-facing child restraints and brain and cervical spine injuries sustained by older children, often unrestrained or restrained in seat belts inappropriate for their age (CDC 1996; Giguere et al. 1998; Hollands et al. 1996a; Huff et al. 1998; Marshall et al. 1998; Willis et al. 1996). Several researchers implemented the crash investigation approach to crash safety engineering to elucidate the mechanisms of injury (Augenstein et al. 1997; Huelke 1997; Kleinberger and Simmons 1997; McKay and Jolly 1999; Quinones-Hinojosa et al. 2005; Shkrum et al. 2002).

For children killed in a rear-facing child restraint system, the air bag typically deployed into the rear surface of the child restraint, often fracturing the plastic shell of the restraint near the child's head and causing fatal skull and brain injuries. Older children who were either unrestrained or restrained in seat belts inappropriate for their age were placed in proximity to the deploying air bag due to preimpact braking. In one typical scenario, the air bag deployment forces the neck into combined tension and hyperextension loading, resulting in a spectrum of injuries to the brain and cervical spine. These include atlantooccipital fracture, brain stem injuries, and diffuse axonal injury of the brain. The largest case series was from NHTSA's Special Crash Investigation program and is summarized in Winston et al. (1996) and Kleinberger et al. (1997). These analyses led to changes in the federal motor vehicle regulations that govern air bags and therefore changes to the actual design of air bags in order to mitigate these injuries.

Physical or Computational Modeling

Once an injury mechanism is identified, engineers use a variety of models to replicate the mechanism in the laboratory to study how to prevent it. Approaches include the use of human volunteers, cadavers, animals, and physical and mathematical models (Stapp 1949a, b; Ewing et al. 1968; Mertz and Patrick 1971; Wismans et al. 1987; King 2000, 2001; McElhaney et al. 1976; Nahum and Melvin 1985). These studies allow the application of carefully controlled engineering inputs and provide a means to clearly define the injury-producing response. Each approach has its value.

Human volunteer experiments have a long-established history in crash safety engineering. Early researchers used themselves as test specimens as described above (Stapp 1949a, b) or enrolled adult human volunteers to define the dynamic response of the human body to trauma (Ewing et al. 1968; Mertz and Patrick 1971; Wismans et al. 1987). In the current age, ethical reasons prevent the utiliza-

tion of volunteer subjects for injurious loading, and as a result, research studying the effects of subinjurious loading has become more common (Arbogast et al. 2009). Since the human body is rate sensitive in its response to trauma (i.e., injuries are related to both the magnitude of the deceleration and the length of time over which it occurs), extrapolation of noninjurious human volunteer data to the injurious situation is challenging. Study of injury mechanisms using postmortem human subjects (PMHSs) or cadavers allows for the exploration of the injury mechanism at potentially injurious energy levels. PMHS experiments are advantageous in that they are tests on human tissue and are adequate for predicting fracture, yet because the tissue is dead, injuries that are due to a physiological cascade such as brain injuries cannot be assessed. Animal experiments are advantageous because the biomechanical and physiologic effects of injury can be assessed, yet geometrically, animals are dissimilar from human, and some scaling is required to transfer tolerance values developed on an animal to a human.

The limitations associated with these biological surrogates lead engineers to develop physical or mathematical models to evaluate the injury-producing event. An example of a physical model is the ATD or crash test dummy. ATDs are one of the primary tools by which occupant protection in motor vehicles is achieved. The key body regions of the ATD are equipped with sensors that measure engineering parameters such as acceleration or force. During testing, the measurements from the sensors are compared against body region-specific injury criteria to evaluate the likelihood that a particular scenario would result in injury to an occupant.

In order to maximize the data obtained from ATDs, there are two critical aspects to their design and interpretation. First, their movement must be biofidelic or humanlike. They need to respond to injury-producing events in the same way a human would. This validation step is often achieved by comparing the gross kinematics or movement of the ATD to that of a PMHS (Forman et al. 2006; Lopez-Valdes et al. 2009, 2010; Tornvall et al. 2005) or more recently human volunteers at subinjurious levels (Arbogast et al. 2009; Wismans et al. 1987). Second, the injury criteria must relate to real injuries in the human body, both in the specific metric quantified (e.g., rotational head acceleration versus linear head acceleration) and the threshold value of that metric that relates to injury. If these biofidelity criteria are not met, the actual injury risk to a human exposed to a similar collision environment may be misrepresented by the ATD, thus providing poor guidance for countermeasure design and testing.

Traditionally, improvements in ATD biofidelity are achieved through rigorous evaluation of PMHS impact testing. Although this approach is an accepted method for obtaining adult ATD design specifications, child PMHS data is limited, and thus, current pediatric ATDs are based on adult biomechanical test data scaled to account for geometric and, to the extent such data are available, material differences between adults and children. However, during the human developmental process, local and regional anatomical structures change in ways that are not quantitatively considered in the scaling processes. In addition to biofidelity, instrumentation limitations can prevent assessment of important occupant injuries. For example, as of 2011, pediatric ATDs do not have accepted measurement capability in the abdomen. Efforts are underway to incorporate this capability into ATD design (Kent et al. 2006, 2008; Arbogast et al. 2005).

Advances in computing capability have led to an increased reliance on mathematical models. Computational versions of the ATDs as well as models of the human, called human body models, offer better efficiency in the design and testing of countermeasures. Such models allow many different parameters to be evaluated, often simultaneously, in a time- and cost-efficient manner. Parametric exploration is achieved via strategies such as the “design of experiments” approach which systematically evaluates multiple combinations of key parameters in an automated, full-factorial manner. Often, new countermeasures are initially designed and evaluated using computer simulations, thus fine-tuning the actual design specifications before prototype production. Such models allow for the evaluation of several “what if” scenarios that can facilitate choices among designs for a particular countermeasure before its production and evaluation in the laboratory.

Computational models are limited by the extent to which they have been validated. The computational model must show comparable responses to the item it is modeling under a similar set of test conditions. As an example, a computational model of a vehicle to be used in crash testing must be compared to an actual vehicle that was crashed in a similar manner. Overall geometry as well as parameters such as speed just prior to impact and deformation of key structures must match between the model and the actual test. Once validated, the model should only be used or “exercised” within the range of test conditions for which it has been validated. For example, it may be inappropriate to use a human body model that has been validated against a PMHS in a frontal crash, in other crash modes such as rollover or side impact. There is no evidence that the model adequately mimics the PMHS movement in these other crash directions.

Key requirements of a rigorous computational human body model are that the “tissues” in the model adequately represent human tissues and that the engineering metrics extracted from the model can relate directly to the likelihood of injury. To do so requires quantification of the mechanical properties of the specific tissues being modeled. Tissue-level material testing of human tissues such as bone (Schreiber et al. 1997; Nyquist et al. 1985; Morgan et al. 1990; Kuppa et al. 2001) or the abdomen (Kent et al. 2006, 2008; Hardy et al. 2001; Miller 1989; Viano et al. 1989) provides characteristics of that tissue that can be directly inserted into the computational model. Again, it is critical that this assessment of material properties is done under similar conditions in which the material will be loaded in the actual model. For example, quantifying the response of brain tissue to small deformations at slow rates is not necessarily useful for modeling brain tissue in impact situations where it undergoes high-rate large deformations. In summary, computational models can be powerful tools that increase efficiency and facilitate the creative design of countermeasures; however, use of an unvalidated model results in data that cannot be trusted.

Development of Safety Countermeasures

Stapp’s landmark research evolved safety engineering from a simplest view of crash prevention to one that also encompassed management of crash energy experienced by the occupant in order to prevent injury. From the point of view of the occupant of a motor vehicle, crashes can be viewed as a series of three collisions: primary collision – i.e., vehicle–vehicle or vehicle–object collision; secondary collision – i.e., occupant collision with the vehicle interior or restraints or other humans/objects in the vehicle; and tertiary collision – i.e., collision within the occupant (e.g., between the body’s internal organs and the bony structures enclosing them).

Energy management for the primary vehicle–vehicle or vehicle–object collision involves managing potential crash incompatibility through optimized vehicle crashworthiness. Crash incompatibility results from impacts between objects of different masses (e.g., crashes between small and large vehicles will likely result in a greater burden for the small vehicle to manage crash energy). The goal is to ensure that the vehicle experiences a relatively low change in velocity through strategies such as controlled vehicle crush and increases in vehicle mass. Properly designed vehicle structures dissipate energy during the crash as they deform, thus being able to preserve the survival space in the passenger compartment while also limiting the velocity change experienced by the occupant during the crash. The effectiveness of energy-absorbing interior components of the vehicle and the restraint system performance can be improved through dissipation of a larger portion of the crash energy in the “crush zones” in the vehicle outside of the passenger compartment. Today’s vehicles have short front ends and lightweight architecture designed for fuel efficiency and require strategies that balance stiffness with low mass while meeting requirements for crashes from all directions.

The vehicle restraint system is the key strategy to manage the energy of the secondary collision between the occupant and the vehicle interior or other occupant. The severity of the occupant's injury is directly related to the deceleration experienced by the occupant (with or without direct impact) as well as localized deformation of key structures of the human body. Restraint systems (e.g., safety belts and air bags) allow the occupant to experience a more gradual deceleration (also known as ride-down) by extending the time over which he or she comes to a stop. Technical advances in safety belts, including load limiters and pretensioning devices, address several of the challenges of the original safety belt: quickly reducing the belt slack in a crash situation will better couple the occupant to the vehicle; permitting the belt to slightly stretch during the crash will reduce the deceleration and consequently the belt forces; and limiting the force exerted by the belt on the occupant will reduce injuries caused because of excessive belt loads.

Air bags were designed as a supplement to the vehicle safety belt for frontal crash protection with the primary aim of preventing head impact. Designed for average-sized adult males, early air bags resulted in serious injuries and, in some cases, fatalities among unbelted occupants, small adults, and children who were in the path of the deploying air bag. In response to these fatalities, industry and the government worked together to facilitate the introduction of "smart" air bags that were designed to prevent injuries to vulnerable users. Specifically, NHTSA revised Federal Motor Vehicle Safety Standard 208, which resulted in the redesign of frontal air bags to reduce the force with which they deploy. By late 1990s, the incidence of air bag deployment-related fatalities virtually disappeared (Chidester and Roston 2001).

Air bag designs continue to evolve to further optimize the protection provided to occupants. There are air bags not only for frontal crash protection but also for side-impact protection, upper interior head protection, and ejection mitigation. In addition, air bag sensors that detect the occupant's position, size, and weight provide input to algorithms that decide the appropriate timing, volume, and rate of deployment to match the demands of the crash event and the occupant. In an integrated safety approach, this information is shared between the air bag and the primary restraint system, the seat belt, to properly design an advanced belt system. Such systems are under development and will soon find their way into the vehicle fleet of the future.

The most challenging collision to manage, the tertiary collision, results in injuries that occur when the energy experienced by the occupant results in differential movement among the body's organs, for example, between the body's internal organs and their anatomical tethers to the bony structures enclosing them. The primary strategy for reducing the tertiary collisions is by minimizing the load applied directly to certain body regions (i.e., the chest or thorax) or reducing differential movement among body parts (i.e., relative movement between the brain and the skull). For example, while the vehicle safety belt restrains the occupant's torso and hips, other supplemental protection systems, such as padding of the vehicle interior or air bags, absorb impact energy or, in the case of side-impact air bags, provide a layer of protection between the body and an intruding vehicle or other structure.

Crash safety engineering often requires a trade-off of protection to different body regions. For example, reducing the forces of the seat belt on the chest may result in fewer thoracic injuries but will lead to increased head excursion as the torso is allowed to flex farther forward in a frontal crash. Detailed understanding of the acute and long-term impact of certain injuries on an occupant's health helps in assessing this trade-off. As an example, rib fractures, which heal quickly in a young adult, may be life threatening for an elderly occupant.

A particularly challenging injury for which to consider technology solutions for prevention is concussion or mild traumatic brain injury. Many of these "mild" injuries lead to poor neurological outcomes that adversely affect the person's quality of life – i.e., second impact syndrome, postconcussion syndrome, and long-lasting neurocognitive deficits such as learning disabilities, memory problems, and emotional or behavioral changes. A National Institutes of Health consensus panel described the "societal burden" of concussion due to the sheer number of patients affected and the

potential for enduring neurologic sequelae (National Institutes of Health 1999). Substantial questions remain regarding the mechanisms of concussion, and thus, technological solutions for prevention are not clear. Determination of how these injuries occur and translating that information into countermeasures should be a priority for injury scientists.

Testing of the Countermeasures

The performance of vehicle occupant protection is evaluated through crash testing and recording injury performance measures using representative test devices (ATDs), thereby assessing human injury risks from those measures. For example, for frontal crash protection assessments, full frontal rigid barrier tests, offset barrier tests, angled barrier tests, vehicle-to-vehicle tests and moving barrier tests, pole (fixed object) tests, and other such tests representing real-world crash occurrences are typically used. Thus, the traditional approach is to evaluate occupant protection performance of vehicle designs through crash testing and mathematical simulations as necessary.

No single test or simulation can evaluate vehicle performance across different occupant types under the wide variety of crash conditions that occur in the real world. Each test evaluates a slightly different aspect of crash performance. For example, consider a full frontal rigid barrier test and an offset barrier test, both conducted at identical crash severities. In the full frontal rigid barrier test, the crash forces and the impact energy are distributed over the full width of the vehicle. In the offset barrier test, where only a portion of the front plane of the vehicle interacts with the crash barrier, the crash energy has to be absorbed by a smaller portion of the front end structure and, as a result, the vehicle experiences much more deformation than in the case of full frontal barrier test. This will naturally lead to greater likelihood of intrusion into the survival space while at the same time leading to a lower mean deceleration experienced in the passenger compartment. In contrast, in the full frontal barrier test, the distributed absorption of crash energy by the structural components of the vehicle such as the engine mounts results in higher mean deceleration while the passenger compartment remains intact. Therefore, it must be understood that while neither test described above fully represents all real-world crash occurrences, the two tests together evaluate different key aspects of crash protection. The full frontal barrier test evaluates the restraint system in the vehicle (seat belts and air bags), and the offset barrier test is a good evaluation tool for monitoring intrusion into the occupant's survival space. It is important to realize that each of these two test methods will likely lead to different and perhaps competing design solutions. For example, to limit the intrusion in an offset test, a vehicle will have structures that are stiffer than the one subjected to the full frontal barrier test.

In addition to full vehicle testing, a common method of evaluation is a sled test. In this method, a crash sled reproduces an acceleration pulse that is similar to what is observed in a full-scale vehicle crash. A sled buck is securely fastened to the sled and serves as a surrogate for the vehicle seat. The buck can feature a full vehicle seat or a simplified test bench. Numerous sled tests can be conducted for the cost – in both time and money – of a full vehicle crash test. While sled tests cannot easily account for the effects of intrusion and deformation of the vehicles' interior, they can provide a reliable assessment of the kinematics or movement of the occupants and their interaction with the restraint system. Sled tests are often used to evaluate different restraint designs on the kinematics and injury metrics as measured by an ATD. It is critical that the sled deceleration pulse and the test buck adequately replicate the full vehicle environment in (1) geometry, (2) component-level response such as the stiffness of the vehicle seat, and (3) system-level response or how the components work together as a whole.

Federal motor vehicle safety regulations incorporate combinations of full vehicle crash tests and sled tests. The portfolio of tests aims to ensure a common standard for safety. Many manufacturers,

however, conduct many more tests using varied approaches. This testing, called due care testing, provides the necessary checks to ensure that unregulated innovations are safe.

The Future of Crash Prevention Engineering: Moving to an Integrated Approach

While the Haddon Matrix has served as an essential guide for traffic safety for the past 40 years, advances in safety technology for the future require an expanded, more integrated approach. The current engineering understanding of crash causation recognizes the considerable overlap between the human and the vehicle and the vehicle and the environment from the time prior to the crash (pre-crash) to the postcrash situation. A shift is required from the current approach to countermeasures, which is often fragmented and compartmentalized based on the literal view of the Haddon Matrix, to a “total safety” solution that capitalizes on the crash causation continuum to identify opportunities to enhance safety. This new approach is possible because of advances in sensors, computers, and decision-making algorithms that were not available when Haddon created his matrix but are now at the disposal of safety engineers to develop the most advanced safety vehicles.

Consider advances that have resulted in dramatic reductions in crash injury based on the Haddon Matrix approach and the additional advances that are needed in crash injury prevention for the future. For most trips today, driving is a benign and uneventful activity: laws and their enforcement have been created to improve traffic conditions and safety, many roadways have been designed to pose minimal risk, vehicles are designed to be reliable and safe, driver controls are designed so that drivers with a range of ability can take advantage of the vehicle’s safety devices and manage everyday hazards, and vehicle crashworthiness and occupant protection systems reduce the severity of injuries should a crash occur.

Crashes occur when, despite the available safety countermeasures, the driver does not recognize a hazard with sufficient time to respond, or the demands of the crash situation placed on the driver exceed his ability to respond to the hazard and avoid or minimize the severity of the crash (through sudden braking, changes in acceleration, steering or any other combination of actions). The events involved in a crash (or its avoidance) occur on a timeline on the order of milliseconds to minutes, as shown in Fig. 30.2 below, and involve complex interactions between the driver, vehicle, and environment. To address many of today’s crash risks, technological advances need to manage three critical time periods: the few critical seconds before time zero (the instant of the crash occurrence), the few milliseconds during the crash, and the few minutes after the crash.

Technological advances make it possible to incorporate vehicle safety technologies into every phase along the crash timeline, even far before time zero. Unfortunately, due to the potential for added vehicle cost and the current lack of regulations governing these devices, the incorporation of many innovative safety features is often driven by consumer demand. If not designed to address the most important and safety-critical aspects of the precrash, crash, and postcrash events, features attractive to consumers may not necessarily improve safety or, as was the case with early air bags, may cause harm (CDC 1995, 1996; Hollands et al. 1996b; Huff et al. 1998; Marshall et al. 1998; Winston and Reed 1996). Simultaneously, with assessing the attractiveness and acceptability of safety features to consumers, manufacturers and regulators need to examine where on the crash timeline the safety technology falls so that it may adequately address the safety needs at that time. The earlier in the crash timeline, the greater the margin of error must be allowed.

Hazard detection technologies, employed at times well in advance of time zero (when the crash occurs), need to ensure accurate hazard detection and must assist rather than distract the driver. Technologies acting immediately prior to time zero to assist the driver in braking or steering require more precision, ensuring they act at a rate that is necessary to move the vehicle away from its

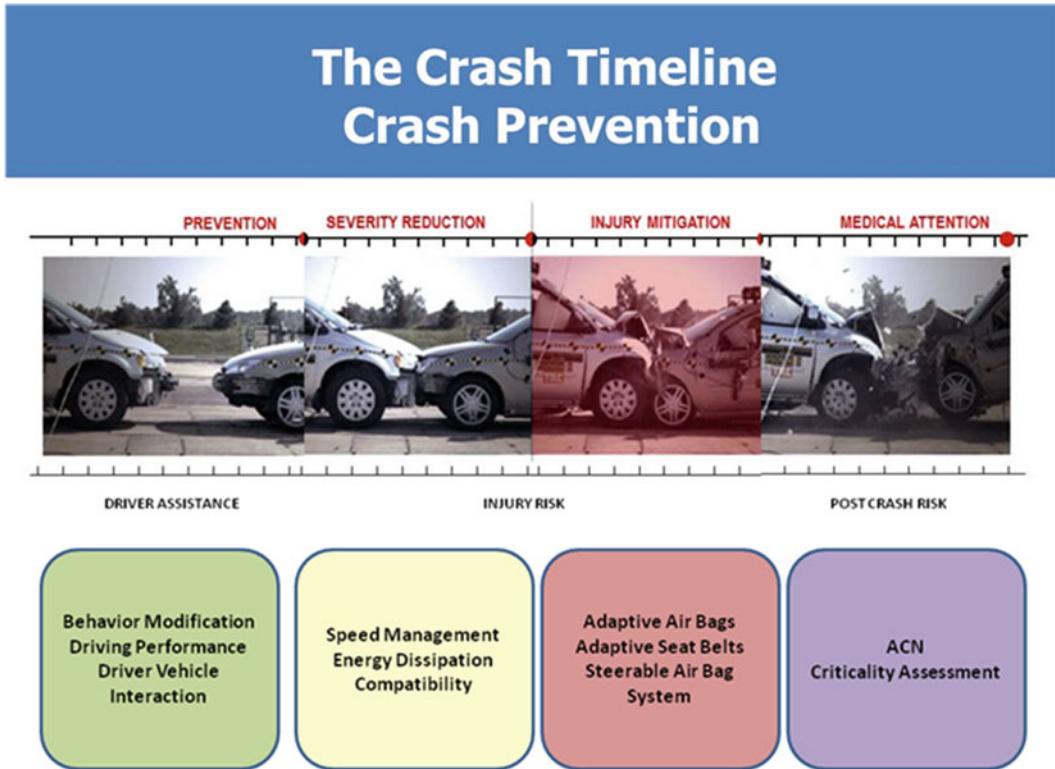


Fig. 30. 2 The crash timeline. In the integrated approach, the steps in the crash timeline are considered simultaneously and optimized for injury reduction, recognizing that how energy is managed early in the crash sequence can affect later injury mitigation and that the more time available to act, the greater the degree of severity reduction that can occur. For example, automatic collision notification (ACN) should begin as soon as sensors detect the severity of the impact

dangerous path. Advanced occupant protection that anticipates a crash and aims to improve occupant safety requires further precision as it must not only accurately detect the hazard but also apply technologies in a manner that reduces rather than increases the risk of injury.

Many such advanced vehicle technologies are becoming available today to warn drivers of imminent crash situations and assist them in taking corrective actions. These include avoidance technologies for rear-end, side, and intersection-type collisions, rollovers, and road departures. Others reduce the crash severity or minimize potential injuries when impacts are unavoidable. Postcrash, vehicles can automatically call emergency services when air bags deploy. Advanced occupant protection technologies include improvements to air bag deployment and other performance characteristics. New and innovative approaches to enhanced safety belt performance, seat performance, and other areas are candidates that could be part of the integrated approach to enhance safety.

Comprehensive technologies that require the most sophistication assess the driving situation and the driver's state (e.g., a drowsy driver), intervene, and take over vehicle control to prevent a crash. These solutions rely on intelligent vehicle technologies that are capable of sensing situations correctly and making decisions at the appropriate time as necessary to assist the driver or intervene when needed. The effectiveness of such technologies will depend upon the sophistication and accuracy of the sensor technologies and the decision-making algorithms, as well as the drivers' willingness to accept the technologies' intervention. If such crash prevention systems are passive and drivers are largely unaware of their intervention, then such technologies may be readily accepted. On the other

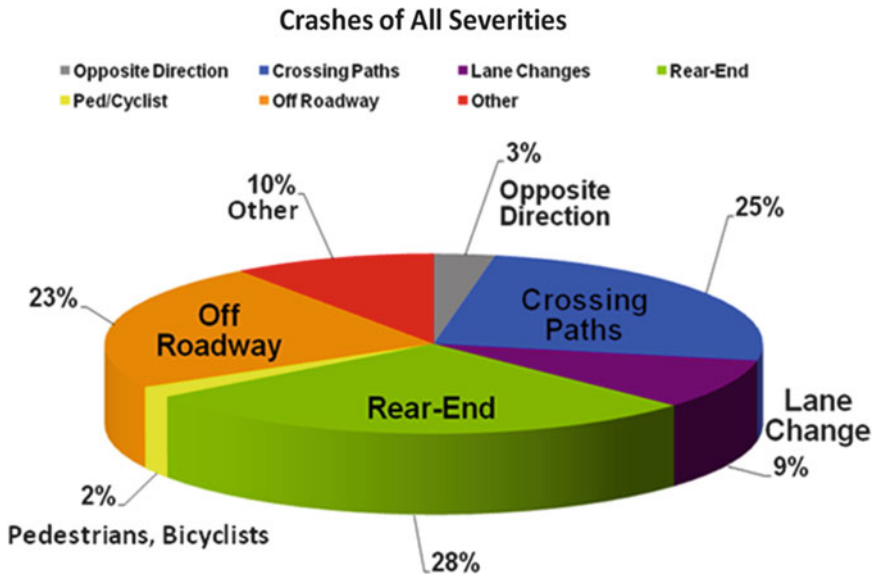


Fig. 30.3 Distribution of crash types (*source: Wassim et al. (2003), Report DOT HS 809 573*)

hand, if drivers see that the vehicle is taking control away or if they have too many false positives and/or false negatives, they may not be immediately willing to use the device.

Close collaboration between injury epidemiologists, behavioral scientists, and engineers is needed to prioritize requirements for the future development of comprehensive safety solutions and technology development. Designs must address the distribution of (1) crash severity, (2) crash types, (3) crash causal factors, and (4) driver and occupant factors. It is important to note that prioritization of safety solutions should be dictated by country- or region-specific needs, but this makes it challenging to design safe vehicles for the global marketplace. Engineers must consider trade-offs on safety benefits to meet the needs of disparate crash scenarios and a range of drivers and occupants. For purposes of illustration of the integrated approach, the following applies to how the actual crash experience in the USA could guide redesign of vehicle safety systems. In the most comprehensive approach, a systems approach to reducing the incidence and severity of crashes and injuries would incorporate roadway, signage, and other environmental factors.

Distribution of Crash Types in the USA

The approximate distribution of crash types in 2009 in the USA based on previous NHTSA analysis (Wassim et al. 2003) is shown in Fig. 30.3, with the majority (nearly three quarters) of crashes encompassing the following crash types:

- Road departure crashes (off-roadway): a nonintersection crash in which a vehicle crosses an edge line or centerline, or leaves the traveled roadway, including intersections at interchange areas (e.g., nonintersection crashes in which the first event for at least one of the involved vehicles was ran-off-road to the right or left, crossed the centerline or median, went airborne, or hit a fixed object)
- Rear-end crashes: a crash in which one vehicle runs into the rear of another vehicle
- Crossing path crashes: a crash that does not occur at a highway interchange but occurs at an intersection or is intersection-related

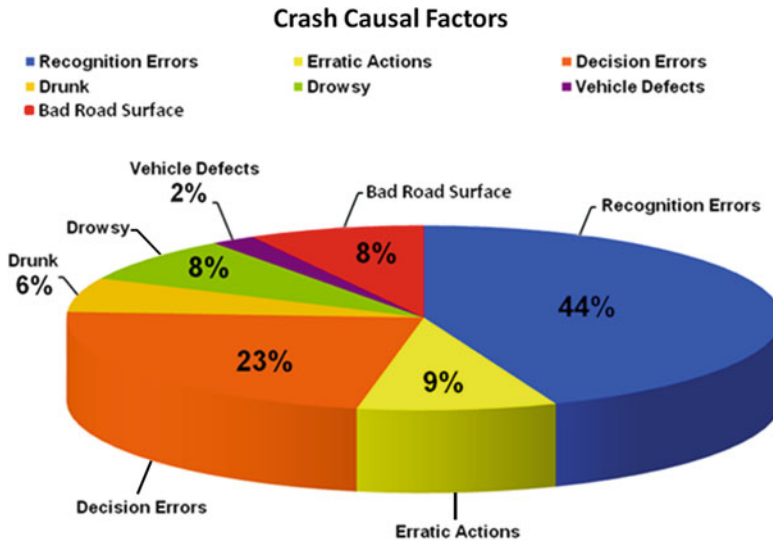


Fig. 30.4 Crash causal factors (*source*: Wassim et al. (1995), Report DOT HS 808 263)

An additional 9% of crashes are of the lane change/merge type and occur when drivers attempt to shift lanes without making sure other vehicles are not encroaching on their intended path. Other crash types, such as vehicle–pedestrian crashes, are less prevalent in the USA than elsewhere, but their prevention is particularly important in emerging economies with mixed road users. While these crash statistics are guides for prioritizing future safety technology development, they are conservative estimates of the true risk in that they do not include crashes that are not reported to the police (and are, therefore, not part of NASS-GES) and the many more near misses that may or may not be realized by the drivers.

Crash Causal Factors for US Crashes

Based on the analysis of real-world crashes (Wassim et al. 1995), the five leading causes of US crashes from the perspective of the vehicle are driver recognition errors, driver decision errors, driver physiological impairment, driver erratic actions, road surface conditions, and vehicle defects (Fig. 30.4). Other causes of crashes not related to the vehicle involve those attributed to the road design or environment.

Driver recognition errors are the most common contributing factor in crash causation (44% of crashes). Examples of recognition errors include driver inattention, driver demonstrated improper scanning, driver looked but did not recognize the hazard, internal and external driver distraction, driver vision obstructed by intervening vehicles, roadway geometry, and roadway appurtenances.

Driver decision errors (23%) constitute the second leading cause of crashes and include crashes in which drivers accelerate to avoid a traffic signal or to overtake another car, misjudging gap/velocity of approaching vehicles, tailgating a lead vehicle, or driving at excessive speeds for the road conditions. This particular category is often seen in combination with other factors such as driver inattention in rear-end crashes, excessive speed and alcohol in single-vehicle road departure crashes, inappropriate speed and improper lookout in backing crashes, and excessive speed and bad pavement conditions in lane change crashes.

Driver physiological impairment (14%) includes driving while intoxicated, drugged, drowsy, or ill and is the third leading cause of crashes. This category is followed by crashes that involve erratic

actions (9%) which mostly involve unlawful driving, unsafe driving acts, and evasive maneuvers. Unlawful drivers are those who deliberately violate signals/signs. Road surface conditions and vehicle defects play a much smaller role in crash causation than other factors related to the driver.

Driver and Occupant Factors

Comprehensive safety systems now allow for the inclusion of “smart” safety strategies that protect occupants based on occupant age, gender, and location in the vehicle for given crash severity, types, and causes. In estimating the benefits and risks of a countermeasure strategy, an analysis of the safety problems and target populations at each stage of the crash event is important. For example, if the technologies of interest are those that are available for preventing crashes, a detailed synthesis of the various critical events leading up to the crash, identification of relevant technologies that could be used in countermeasures, and assessment of effectiveness form the basis for estimating potential benefits (Kanianthra 2006).

Before new safety systems are put into the fleet, iterative testing is needed, including test track testing, simulation studies, and, in some cases, field operational tests that provide substantiating data. An important element in this process is the development of suitable test procedures that are objective and are related to real-world circumstances. Results from these test procedures are key to estimating the anticipated benefits in addressing specific safety problems. Once technologies are introduced into the fleet, postmarket surveillance is needed to ensure safety and effectiveness. However, required time needed to recognize, report, and confirm safety issues; the limited effectiveness of recalls (U.S. Consumer Product Safety Commission 2003); and the long vehicle product cycle limits ability to quickly correct safety problems once they are recognized and places greater pressure on premarket testing.

The Promise of Advanced and Active Safety Systems

There is no question that technologies are bringing new opportunities for enhancing safety in vehicles that never existed before. For example, many new technologies are already on the road or are on the verge of introduction in vehicles. These include products that are extensions of the antilock brake systems such as electronic stability and traction controls, adaptive cruise controls, road departure warning systems, night vision systems, and advanced restraint and occupant protection systems. Many of these advanced safety technologies are expected to produce benefits, but it is too early to say if indeed, as designed, they result in improved safety. There is limited real-world crash experience with such vehicles as most of these new technologies are only introduced in high-end, limited production vehicles. Evaluations require a considerable amount of time, research, analyses, and testing and consume extensive resources but remain an essential step in vehicle safety evolution.

These new technologies often operate in synergy from precrash to crash to postcrash, further blurring clearly demarcated distinctions of the Haddon Matrix. In addition, it is no longer helpful to classify safety technology along the traditional lines of active versus passive safety. Traditionally, active safety countermeasures have been categorized as those that involve human action to avoid crashes or injury (e.g., warning systems for drivers and adult safety belts for occupants) while passive systems spring into action automatically without the driver actively participating in triggering the system (e.g., air bags). Advanced safety systems are all capable of preventing or attempting to prevent crashes but might involve variable levels of human response.

While the fatalities and serious injuries in the USA have shown significant decline since 2007, it is not clear whether that decline will continue in future years, particularly after the economy

improves. It is possible the casualties may continue at an unacceptable level of 30,000 fatalities per year, i.e., at the rate of over 80 people dying every day in motor vehicle crashes. The safety countermeasures that have been added to vehicles to meet consumer demand, the current government safety standards, and programs intended to increase consumer knowledge (e.g., from the Insurance Institute of Highway Safety) over the last 40 years have all focused on crash protection to protect the occupants, in the belief that many crashes are inevitable.

Currently, however, that mind-set is changing to lean toward crash prevention. For years, the main safety standards that work toward crash prevention have been the lights and brakes. New efforts are beginning to forge regulations for advanced safety systems, such as electronic stability control. There are many other technologies and a demand for their introduction into the fleet, but their efficacy in enhancing safety cannot currently be assured. Manufacturers and regulators will be challenged with determining their proper introduction with the availability of limited validated testing methods and regulations. Their potential for enhancing safety is enormous. Even if they are not completely successful in preventing crashes, they will likely reduce the severity of the crash and the likelihood for injuries and fatalities. These safety advancements can supplement the more traditional approaches related to crash avoidance, crash severity, the protection of occupants, post-crash safety, and even the improvement of structural integrity and preservation of survival space in motor vehicles.

Active safety solutions have great promise, and many of the systems will find their way into the fleet as “cooperative systems,” where a combination of the driver and the vehicle technology has to occur to derive potential benefits. For example, a forward collision warning system that warns the driver still requires the driver to heed the warning and take the appropriate action in time. Ideally, systems would be established to intervene autonomously if the driver fails to respond or takes inappropriate action. As a result, human factor research and consideration for the wide range of drivers and occupants will need to be considered. Validated methods are needed that will assess how active safety technologies affect the ability of drivers to process information, improve driving capacity, or improve driver alertness. The ability of drivers to cope with the information that is continuously being provided needs to be investigated, particularly among at-risk driver populations such as novice teenage drivers and elderly drivers. Similarly, remembering the deaths to short women and children from exposure to air bags designed for the average-sized adult male, designers of passive safety systems must consider how they will perform with the range of drivers and occupants and the real-world ways in which they might interact with the systems. Also, the integration of multiple technologies and data fusion needs to be assured as many of the products are developed by separate suppliers, each wanting to preserve intellectual property rights. It is not enough to thoroughly research the design and development of safety components. An integrated system approach, exploring issues of both active and passive safety, is needed.

The Challenge Ahead

Advances in safety technology have contributed to the dramatic drop in motor vehicle fatalities and injuries in the USA over the past 40 years with considerable reductions seen between 2007 and 2009. Despite this, in 2009, according to NHTSA, nearly 34,000 people died in motor vehicle crashes (Traffic Safety Facts, DOT HS 811 363 2010), and new approaches are needed.

While some credit safety advances for the recent crash fatality reductions, others point to similar reductions experienced with other economic downturns. Would a drop in fatalities such as that seen in the last 2 years be sustainable in future years, should there be an economic boom? An early estimate of traffic fatalities by NHTSA (Hedlund 2010) for the first three quarters of 2010 shows a further decline in fatalities from the same period in 2009 by 4.5%. However, in the third quarter

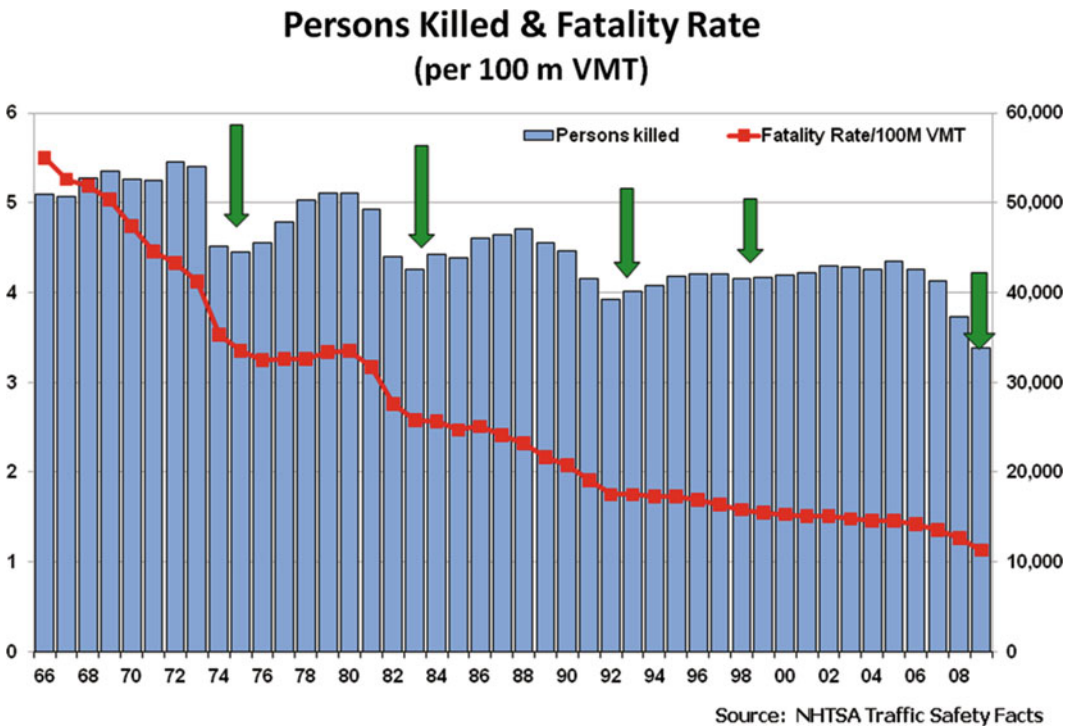


Fig. 30.5 Motor vehicle crash fatalities 1966–2009 (source: Traffic Safety Facts, DOT HS 811 363 2010)

alone, there was an increase of 226 fatalities compared to the third quarter in 2009. Therefore, it is estimated that improved economy and the growth in vehicle miles traveled alone could add another 400 fatalities on average per year, once the vehicle miles traveled and the economy reach normal levels (Kanianthra 2011). This means that by 2020, there could be an increase of 4,000 fatalities from the level in 2009, at the average rate of 400 additional fatalities per year. In order to keep the fatalities at the 2009 level or below, there is a need to take an aggressive approach beyond what has been attempted up until now (Fig. 30.5).

There is the potential for significant safety improvements with the emerging vehicle technological advancements that use an integrated safety strategy – i.e., attempting crash prevention first, thus at least reducing crash severity and mitigating injuries through improved protection countermeasures that use advanced technologies (Kanianthra and Mertig 1997). However, it should be stressed that in order to close the anticipated gap in safety, it does not seem possible at this point to achieve sufficient penetration of advanced technologies in the fleet in the near future without a concerted effort by the automobile industry, the government, and the safety community. It is also impossible to completely close the gap through active safety technologies alone. Therefore, it is important to find a strategy for ensuring the effectiveness and safety for advanced safety systems and accelerating their penetration into the fleet. Otherwise, without ready demand for such systems, there will be no incentive for the suppliers to continue efforts in developing future technologies.

One strategy for promoting adoption of active and advanced safety technologies is through consumer information and demand and safety rating programs such as NHTSA’s New Car Assessment Programs (NCAP) and the Insurance Institute for Highway Safety. Regulations will be more challenging in the future as they will require not only full vehicle system-based standards but also standards for assessing driver response to the systems. Other advanced safety technologies could be offered through market incentives initially, until consumers become aware of their safety potential

and demand further penetration into the fleet. However, after the technologies have been in the fleet for a few years, close surveillance will be required to evaluate their effectiveness and identify safety hazards. Therefore, in order to advance safety, it is important that new approaches other than safety regulations be pursued to accelerate the penetration of advanced safety technologies into the fleet.

This chapter focused on advances in the technological approach to driver and occupant protection. It is important to note that in order to address the global burden of traffic injury, a more comprehensive and challenging view is needed. The vast majority of fatalities occur to vulnerable road users (e.g., pedestrians and bicyclists) particularly on roads with mixed user types. In addition, countries bearing the large proportion of road injuries and fatalities are also those with the least well-developed roads, older technology in vehicles, and new vehicles that are manufactured to weak or no safety regulations, and the population lacks the resources to afford the advanced safety technologies. This scenario mixed with the increasing demand for low-cost vehicles portends a continued and growing global health crisis in traffic injuries, as highlighted by the World Health Organization-led Decade of Action for Road Safety (Peden 2010). Innovative, low-cost, easily integrated solutions must be part of future planning.

This approach incorporating the principles of biomechanics and engineering can be applied to prevention strategies for injury causes other than motor vehicle crashes, including but not limited to falls in the elderly, sports injuries, natural disasters, and fires. Once the mechanism of injury is known, technology can be designed to mitigate the transmission of this energy exposure to the human. Sensors can detect a potential excess energy exposure which can trigger warnings and automated preventive and emergency response actions that can reduce the incidence or severity of injury. For example, wireless building sensors attached to buildings can monitor building sway from earthquakes and send messages to trigger algorithms that control building sway through automated dampers. Regardless of the injury targeted for prevention, the following key principles are critical to an integrated, advanced safety approach:

- Ground all activities in evidence and evaluation. Data driven systems should be used to set priorities.
- Work in multidisciplinary teams of engineers, behavioral scientists, and others.
- Integrate countermeasures across the safety timeline. All phases from prevention, severity reduction, injury mitigation, and medical attention remain important and should be emphasized in prevention activities.
- Adapt proven strategies to local conditions and injury prevention needs.

Acknowledgments The authors acknowledge Mari Allison and Caitlin Locey from The Center for Injury Research and Prevention at The Children's Hospital of Philadelphia (CHOP) for their help in formatting the manuscript. The authors would like to acknowledge the National Science Foundation (NSF) for sponsoring the writing of this chapter. The views presented are those of the authors and not necessarily the views of CHOP and the NSF.

References

- Arbogast, K. B., Mong, D. A., Mari-Gowda, S., Kent, R., Stacey S, Mattice, J., et al. (2005). *Evaluating pediatric abdominal injuries*. Paper presented at the 19th International Enhanced Safety of Vehicles (ESV) Conference, Washington, DC.
- Arbogast, K. B., Balasubramanian, S., Seacrist, T., Maltese, M. R., Garcia-Espana, J. F., Hopely, T., et al. (2009). Comparison of kinematic responses of the head and spine for children and adults in low-speed frontal sled tests. *The Stapp Car Crash Journal*, 53, 329–372.
- Augenstein, J., Perdeck, E., Williamson, J., Stratton, J., Digges, K., & Lombardo, L. (1997). *Air bag induced injury mechanisms for infants in rear facing child restraints*. Paper presented at the 41st Stapp Car Crash Conference, Lake Buena Vista, FL.

- Bonnie, R. J., Fulco, C., Liverman, C. T., & Institute of Medicine (U.S.). Committee on Injury Prevention and Control. (1999). *Reducing the burden of injury: advancing prevention and treatment*. Washington, DC: National Academy Press.
- Centers for Disease Control and Prevention (CDC). (1995). Air-bag associated fatal injuries to infants and children riding in front passenger seats – United States. *MMWR*, vol. 44, 845–847.
- Centers for Disease Control and Prevention (CDC). (1996). Update: Fatal air bag-related injuries to children – United States, 1993–1996. *MMWR. Morbidity and Mortality Weekly Report*, 45, 1073–1076.
- Chidester, A. B., & Roston, T. A. (2001). *Air bag crash investigations*. Paper presented at the 17th International Enhanced Safety of Vehicles (ESV) Conference, Amsterdam, The Netherlands.
- De Haven, H. (1942). Mechanical analysis of survival in falls from heights of fifty to one hundred and fifty feet. *War Medicine*, 2, 586–596.
- Ewing, C. L., Thomas, D. J., Beeler, G. W., Jr., Patrick, L. M., & Gillis, D. B. (1968) *Dynamic response of the head and neck of the living human to -Gx impact acceleration*. Paper presented at the 12th Stapp Car Crash Conference, Warrendale, PA.
- Forman, J., Lessley, D., Shaw, C. G., Evans, J., Kent, R., Rouhana, S. W., et al. (2006). Thoracic response of belted PMHS, the Hybrid III, and the THOR-NT mid-sized male surrogates in low speed, frontal crashes. *The Stapp Car Crash Journal*, 50, 191–215.
- Giguere, J. F., St-Vil, D., Turmel, A., Di Lorenzo, M., Pothel, C., Manseau, S., et al. (1998). Air bags and children: a spectrum of C-spine injuries. *Journal of Pediatric Surgery*, 33(6), 811–816.
- Haddon, W., Jr. (1972). A logical framework for categorizing highway safety phenomena and activity. *The Journal of Trauma*, 12(3), 193–207.
- Haddon, W., Jr. (1980). Options for the prevention of motor vehicle crash injury. *Israel Journal of Medical Sciences*, 16(1), 45–65.
- Hardy, W. N., Schneider, L. W., & Rouhana, S. W. (2001). Abdominal impact response to rigid-bar, seatbelt, and air bag loading. *The Stapp Car Crash Journal*, 45, 1–32.
- Hedlund, J. (2010). *Pedestrian traffic fatalities by state, DOT HS 811 431*. Washington, DC: National Highway Traffic Safety Administration.
- Hollands, C. M., Winston, F. K., Stafford, P. W., & Lau, H. T. (1996a). Lethal air bag injury in an infant. *Pediatric Emergency Care*, 12(3), 201–202.
- Hollands, C. M., Winston, F. K., Stafford, P. W., & Shochat, S. J. (1996b). Severe head injury caused by air bag deployment. *The Journal of Trauma*, 41(5), 920–922.
- Huelke, D. F. (1997). *Children as front seat passengers exposed to air bag deployments*. Paper presented at the 41st Stapp Car Crash Conference, Lake Buena Vista, FL.
- Huff, G. F., Bagwell, S. P., & Bachman, D. (1998). Air bag injuries in infants and children: a case report and review of the literature. *Pediatrics*, 102(1), e2.
- Kanianthra, J. (2006). *Re-inventing safety: do technologies offer opportunities for meeting future safety needs?* Paper presented at the SAE Convergence Conference, Detroit, MI.
- Kanianthra, J. (2011, April) *Advancing safety in the future: the role of technologies, the government, and the industry*. Paper published by SAE in the Progress in Technology Series book titled Active Safety and the Mobility Industry, PT-147, SAE International, Detroit, MI.
- Kanianthra, J & Mertig, A. (1997). Opportunities for collision countermeasures using intelligent technologies. In *Proceedings of the International Symposium on Real World Crash Injuries*. Leicestershire, England.
- Kent, R., Viano, D. C., & Crandall, J. (2005). The field performance of frontal air bags: a review of the literature. *Traffic Injury Prevention*, 6(1), 1–23.
- Kent, R., Stacey, S., Kindig, M., Forman, J., Woods, W., Rouhana, S. W., et al. (2006). Biomechanical response of the pediatric abdomen, part 1: development of an experimental model and quantification of structural response to dynamic belt loading. *The Stapp Car Crash Journal*, 50, 1–26.
- Kent, R., Stacey, S., Kindig, M., Woods, W., Evans, J., Rouhana, S. W., et al. (2008). Biomechanical response of the pediatric abdomen, Part 2: injuries and their correlation with engineering parameters. *The Stapp Car Crash Journal*, 52, 135–166.
- King, A. I. (2000). Fundamentals of impact biomechanics: Part I – Biomechanics of the head, neck, and thorax. *Annual Review of Biomedical Engineering*, 2, 55–81.
- King, A. I. (2001). Fundamentals of impact biomechanics: Part 2 – Biomechanics of the abdomen, pelvis, and lower extremities. *Annual Review of Biomedical Engineering*, 3, 27–55.
- Kleinberger, M., & Simmons, L. (1997, November 10–11). Mechanisms of injuries for adults and children resulting from air bag interaction. In: *41st Annual Proceedings of the Association for the Advancement of Automotive Medicine*, Orlando, FL.
- Kuppa, S., Wang, J., Haffner, M., & Eppinger, R. (2001). *Lower extremity injuries and associated injury criteria*. Paper presented at the 17th International Enhanced Safety of Vehicles (ESV) Conference, Amsterdam, The Netherlands.

- Lopez-Valdes, F. J., Forman, J., Kent, R., Bostrom, O., & Segui-Gomez, M. (2009). A comparison between a child-size PMHS and the Hybrid III 6 YO in a sled frontal impact. *Annals of Advances in Automotive Medicine*, 53, 237–246.
- Lopez-Valdes, F. J., Lau, A., Lamp, J., Riley, P., Lessley, D. J., Damon, A., et al. (2010). Analysis of spinal motion and loads during frontal impacts. Comparison between PMHS and ATD. *Annals of Advances in Automotive Medicine*, 54, 61–78.
- Marshall, K. W., Koch, B. L., & Egelhoff, J. C. (1998). Air bag-related deaths and serious injuries in children: injury patterns and imaging findings. *AJNR. American Journal of Neuroradiology*, 19(9), 1599–1607.
- McElhaney, J. H., Roberts, V. L., Hilyard, J. F., & Kenkyu jo, N. J. (1976). Properties of human tissues and components: nervous tissues. In J. H. McElhaney, V. L. Roberts, & J. F. Hilyard (Eds.), *Handbook of human tolerance*. Tokyo: Japan Automobile Research Institute (JARI).
- McKay, M. P., & Jolly, B. T. (1999). A retrospective review of air bag deaths. *Academic Emergency Medicine*, 6(7), 708–714.
- Mertz, H. (1988). *Restraint performance of the 1973–76 GM air cushion restraint system*. Warrendale, PA: Society of Automotive Engineers. SAE Paper No. 973296.
- Mertz, H. J., & Patrick, L. M. (1971). *Strength and response of the human neck*. Paper presented at the 15th Stapp Car Crash Conference, Coronado, CA.
- Miller, M. A. (1989). The biomechanical response of the lower abdomen to belt restraint loading. *The Journal of Trauma*, 29(11), 1571–1584.
- Morgan, R. M., Eppinger, R., Marcus, J., & Nichols, H. (1990). Human cadaver and Hybrid III responses to axial impacts of the femur. In *International Research Council on the Biomechanics of Impact (IRCOBI)*, Bron, France.
- Nahum, A. M., & Melvin, J. (1985). *The biomechanics of trauma*. Norwalk, CT: Appleton-Century-Crofts.
- National Highway Traffic Safety Administration (NHTSA). (2003). *Analysis of light vehicle crashes and pre-crash scenarios based on the 2000 General Estimates System*. Washington, DC: U.S. Department of Transportation.
- National Highway Traffic Safety Administration (NHTSA) Notes. (1996). *Annals of Emergency Medicine*, 28(4), 450–452.
- National Institutes of Health. (1999). Consensus conference. Rehabilitation of persons with traumatic brain injury. NIH Consensus Development Panel on Rehabilitation of Persons With Traumatic Brain Injury. *JAMA*, 282(10), 974–983.
- Nyquist, G. W., Cheng, R., El-Bohy, A. A. R., & King, A. I. (1985). *Tibia bending: strength and response*. Paper presented at the Society of Automotive Engineers, Warrendale, PA.
- Peden, M. (2010). UN General Assembly calls for decade of action for road safety. *Injury Prevention*, 16(3), 213.
- Quinones-Hinojosa, A., Jun, P., Manley, G. T., Knudson, M. M., & Gupta, N. (2005). Air bag deployment and improperly restrained children: a lethal combination. *The Journal of Trauma*, 59(3), 729–733.
- Schreiber, P., Crandall, J. R., Micek, T., Hurwitz, S., & Nusholtz, G. (1997, September). Static and dynamic bending strength of the leg. In *International Research Council on the Biomechanics of Impact (IRCOBI)*, Hannover, Germany.
- Shkrum, M. J., McClafferty, K. J., Nowak, E. S., & German, A. (2002). Driver and front seat passenger fatalities associated with air bag deployment. Part 2: A review of injury patterns and investigative issues. *Journal of Forensic Sciences*, 47(5), 1035–1040.
- Stapp, J. P. (1949a). *Human exposures to linear acceleration; I. Preliminary survey of aft-facing seating position* (Air Force Technical Report No. 5915). Dayton, IL: Wright Air Development Center.
- Stapp, J. P. (1949b). *Human exposures to linear deceleration; II. The forward facing position and the development of a crash harness* (Air Force Technical Report No. 5915). Dayton, IL: Wright Air Development Center.
- Stapp, J. P. (1955). Effects of mechanical force on living tissues. I. Abrupt deceleration and windblast. *Journal of Aviation Medicine*, 26(4), 268–288.
- Stapp, J. P. (1957). Human tolerance to deceleration. *American Journal of Surgery*, 93(4), 734–740.
- Tornvall, F. V., Svensson, M. Y., Davidsson, J., Flogard, A., Kallieris, D., & Haland, Y. (2005). Frontal impact dummy kinematics in oblique frontal collisions: evaluation against post mortem human subject test data. *Traffic Injury Prevention*, 6(4), 340–350.
- Traffic Safety Facts, DOT HS 811 363. (2010). Washington, DC.
- U.S. Consumer Product Safety Commission. (2003). Recall effectiveness research: a review and summary of the literature on consumer motivation and behavior. CPSC Order No. CPSC-F-02-1391.
- Viano, D. C., Lau, I. V., Asbury, C., King, A. I., & Begeman, P. (1989). Biomechanics of the human chest, abdomen, and pelvis in lateral impact. *Accident; Analysis and Prevention*, 21(6), 553–574.
- Wassim, N., Mironer, M., Koziol, J., & Wang, J. (1995). *Synthesis report: examination of target vehicular crashes and potential ITS countermeasures* (Report No. DOT HS 808 263). Washington, DC: National Highway Traffic Safety Administration.

- Wassim, N., Basav, S., Smith, J. D., & Campbell, B. N. (2003). *Analysis of light vehicle crashes and pre-crash scenarios based on the 2000 General Estimate System* (Report No. DOT HS 809 573). Washington, DC: National Highway Traffic Safety Administration.
- Willis, B. K., Smith, J. L., Falkner, L. D., Vernon, D. D., & Walker, M. L. (1996). Fatal air bag mediated craniocervical trauma in a child. *Pediatric Neurosurgery*, 24(6), 323–327.
- Winston, F. K., & Reed, R. (1996) *Air bags and children: results of a National Highway Traffic Safety Administration special investigation into actual crashes*. Paper presented at the 40th Stapp Car Crash Conference, Warrendale, PA.
- Winston, F. K., Schwarz, D. F., & Baker, S. P. (1996). Biomechanical epidemiology: a new approach to injury control research. *The Journal of Trauma*, 40(5), 820–824.
- Wismans, J., Philippens, M., van Oorschot, E., Kallieris, D., & Mattern, R. (1987). *Comparison of human volunteer and cadaver head-neck response in frontal flexion*. Paper presented at the 30th Stapp Car Crash Conference, Warrendale, PA.

Chapter 31

Behavioral Approach

Andrea Carlson Gielen, Eileen M. McDonald, and Lara B. McKenzie

Introduction

Reducing injuries requires the expertise of many disciplines as well as the social and political will to make individual and environmental changes that are often initially unpopular. There are numerous examples of successes that attest to the need for engaging multiple stakeholders to find solutions, such as the reductions in motor vehicle-related and occupational injuries that occurred in the USA in the twentieth century (MMWR 1999a, b). What was the role for behavioral approaches in such successes, and what can behavioral sciences contribute to addressing contemporary injury problems as well as future hazards that are yet to emerge? That is the focus of this chapter. Over the years, a wealth of empirical and theoretical work has advanced the behavioral sciences, making it possible to more fully explore the role of behavior change across the spectrum of individuals whose actions determine the public's injury risk (Gielen et al. 2006; Gielen and Girasek 2001).

The overall goal of the chapter is to acquaint readers with behavior change opportunities and applications to injury reduction from the perspectives of both the well-known epidemiological framework of host, vector, and environment and the ecological framework commonly used in health promotion. This chapter (1) describes the roles of behavior change in reducing injury, highlighting the need for comprehensive approaches that address multiple levels of the ecological model and (2) provides examples from the literature of changes in individual behavior, products, and environments to illustrate the value in considering a variety of audiences and goals for behavior change.

A.C. Gielen, ScD, ScM (✉)

Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health,
624 N. Broadway Room 554, Baltimore, MD 21205, USA
e-mail: agielen@jhsph.edu

E.M. McDonald, MS

Center for Injury Research and Policy, Johns Hopkins Bloomberg
School of Public Health, 624 N. Broadway Room 731, Baltimore, MD 21205, USA
e-mail: emcdonal@jhsph.edu

L.B. McKenzie, PhD, MA

Center for Injury Research and Policy, The Research Institute at Nationwide Children's Hospital,
The Ohio State University, 700 Children's Drive, Columbus, OH 43205, USA
e-mail: lara.mckenzie@nationwidechildrens.org

Roles for Behavioral Approaches

When the field of injury prevention and control was in its infancy, one of the many important contributions of William Haddon and Susan Baker was to change the focus from *accident* prevention directed at individual behavior modification to *injury* prevention through environmental modification (Haddon 1970; Haddon 1980; Baker 1973). Their work identified ways to reduce both the chance that an injury-producing event would occur and the risk that an injury would occur if the event happened. This paradigm shift emphasized the epidemiological triad of host, agent, and environment and was responsible for generating many of the effective injury countermeasures we have today. Over time, it has become clear that this problem-solving paradigm for reducing injury should not be misinterpreted as a dichotomy between “changing the individual or changing the environment.”

Fishbein noted that “even more than most behavioral scientists, injury prevention scientists have taken an ecological perspective that has led them to pay important attention to the interaction between people and their environments” (Fishbein 2006, p. x). It is this interaction that makes obvious the need for comprehensive approaches. For instance, effective policy solutions to injury problems, such as seat belt laws and building codes, work when there is awareness and enforcement. With notable exceptions (e.g., shatter-resistant windshields, vehicle roll-over protection), even passive protection options have critically important “active” components. For instance, antilock brakes must be used properly, air bag protection requires passengers to buckle their seat belts and parents to place young children in the backseat, and smoke alarm batteries must be changed. As noted previously, “although passive, all of these intervention strategies require some human interaction to achieve their full safety potential” (Gielen and Sleet 2006, p. 6).

Another View of the 3 Es

The most widely used mnemonic in injury prevention, “the 3 Es” (National Committee for Injury Prevention and Control 1989), may have contributed to the earlier conceptual divide between behavior change and environmental approaches. The 3 Es of injury prevention explain that engineering addresses physical environmental risk factors, enforcement addresses social and political risk factors, and education addresses behavioral risk factors. Grossman noted that with the maturing of the field of injury prevention, it was time to expand the paradigm beyond the 3 Es model of prevention into one that better addresses the behavioral components of injury problems (Grossman 2006).

The problem is that the 3 Es approach ignores the reality that engineering and enforcement are powerful tools to change behavior. For example, seat belt laws – an enforcement strategy – helped boost use rates from 11% in 1980 when seat belt use was not mandated to 82% in 2007 (Nichols 1994; Glassbrenner and Ye 2007). Installing traffic-calming devices – an engineering strategy – reduces motor vehicle and pedestrian injuries by changing individual driving behavior (Ewing 1999). Thus, a more accurate problem-solving framework would clarify that the 3 Es can reduce the risk of injury either directly (e.g., as in the case of removing a hazard from the environment entirely) or indirectly through changing individual behavior (e.g., as in the case of mandating a safer behavior) (see Fig. 31.1). Moreover, as described above, success in reducing injury often relies on the use of a combination of these strategies.

Reducing fall injuries at home among children illustrates the individual and interactive effects of the 3 Es. Balconies, decks, and porches pose fall risks to young children, and vertical railings with a maximum of 4 in. spacing between the balusters are preventive (Stauton et al. 2007). This environmental modification is now required in building codes for new construction in much of

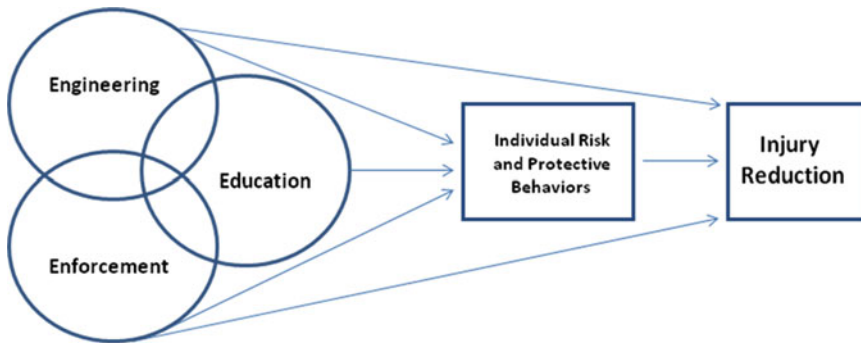


Fig. 31.1 The 3 Es conceptual framework for reducing injury

the USA. This enforcement strategy, while effective for new construction, does nothing to protect children living in homes constructed prior to the code. Education can be an effective strategy to communicate this potentially lifesaving information to families who have yet to benefit from improved building codes. The first two approaches in this example (e.g., environmental modification and building codes) reduce injury risk directly, avoiding having to change the child or parent’s behavior. Education in this example requires parents to take some action to protect their child (e.g., modify the railings themselves, increase supervision). Thus, all 3 Es come into play to provide a comprehensive approach to reducing children’s injury risks.

Education and Behavior Change

Education is defined as “the imparting and acquiring of knowledge through teaching and learning” (Encarta Dictionary). Education, in the truest sense of the word, is critical for enhancing or creating necessary precursors of behavior change such as understanding, attitudes, motivation, skills, and confidence. An axiom in health education, derived from decades of research and practice, is that knowledge is necessary but insufficient for behavior change (Green and Kreuter 2005). Sustained behavior change at the population level typically requires environmental supports that reduce barriers and provide incentives. For instance, reviews of the evidence for reducing motor vehicle occupant injuries among children consistently find that the most effective approach is strong, enforced legislation combined with education and car seat distribution programs (US Preventive Services Task Force 2007). Similarly, the Cochrane Collaboration has found that child home safety is enhanced by efforts that both provide education and make products available (Turner et al. 2011).

Education and behavior change have also been linked in the injury literature in terms of who needs to be educated and whose behavior needs to change to reduce injury risk. As early as 1991, Wilson, Baker, and colleagues applied a wide lens to the utility of behavior change in their book *Saving Children* (1991) when they highlighted the role of decision makers. For children, the injury prevention knowledge and behaviors of multiple stakeholders are critically important. Wilson et al. (1991) include schools, child care centers, legislators and regulators, law enforcement, voluntary organizations, designers, architects, builders and engineers, business, industry, and mass media as key actors in reducing childhood injury. This same approach – i.e., changing the knowledge and behavior of key influencers in addition to the individuals at risk – can and should be utilized when addressing all injury problems. This is the basic premise from which this chapter draws its inspiration.

Ecological Models

An organizing framework that helps to systematically think about comprehensive approaches and the behavior of multiple key stakeholders is an ecological model. The concept of an ecological model has a long history in public health and health promotion (for more in-depth reviews, see Sallis et al. 2008; Allegrante et al. 2006; Stokols 1992). In brief, ecological models focus on the environment and the ways in which individuals interact with their physical and social environments, making them a natural expansion of Haddon's original tripartite epidemiological model. One widely used ecological framework in health promotion (McLeory et al. 1988; Green and Kreuter 2005; Hanson et al. 2005; Allegrante et al. 2006) consists of five levels of influence:

- Intrapersonal – individual characteristics including knowledge, beliefs, attitudes, skills, and other relevant cognitive or affective factors
- Interpersonal – social networks, including family, friends, peers, colleagues, and others who are important to an individual
- Organizational – mediating structures such as associations, social institutions, workplaces, and others that have rules and regulations and pursue specific objectives that can influence an individual
- Community – structurally, a geographic or political jurisdiction and functionally, a group that shares identity, values, norms, and other factors that can influence an individual
- Societal – larger systems with political boundaries that have the power to distribute resources and control over communities and individuals

According to Allegrante et al. (2006, p. 111), this “socioecological paradigm emphasizes the dynamic interface among the three dimensions – the individual, the physical environment, and the social environment – acting at five levels.” The levels “provide the ecological context in which the individual acts.” These authors go on to note that the higher or deeper levels are more difficult to change, but when changed, are more likely to achieve sustained change. For example, changing laws (i.e., the societal level) is more difficult than changing what a group of individuals know about an injury problem (i.e., the intrapersonal level), but when laws are in place and enforced, sustained behavior change occurs over time. In fact, societal-level changes such as the passage of safety laws exert their influence by changing community norms, organizational structures, and social networks. Exerting such influence throughout the chain of influencing levels is what supports and reinforces long-term sustained change in the way individuals behave.

To illustrate the application of ecological thinking to injury prevention, consider that the USA achieved such significant reductions in motor vehicle crash death rates that the CDC called it one of the ten greatest public health achievements of the twentieth century (MMWR 1999a). Since the 1920s, the number of vehicle miles traveled in the USA has multiplied ten times, while the annual death rate per billion vehicle miles has decreased 90% (MMWR 1999a). While the reduction is due in part to the fact that, on average, there are fewer people in each vehicle, and therefore fewer at risk in each vehicle mile, most of this success is the result of changing the physical and social environment and human behavior: modifications to roadways and vehicles, increased use of seat belts and child safety seats, and decreased rates of drunk driving (MMWR 1999a; Nichols 1994; Graham 1993; Waller 2001; Rivara and MacKenzie 1999; Zwerling and Jones 1999). Gielen and Sleet (2006) provide examples of the multilevel impacts of ecological and comprehensive approaches in their discussion of the motor vehicle safety success story:

... the policy environment ... changed dramatically with the introduction of laws requiring changes in the behaviors of restraint use and drunk driving. The organizational environments of workplaces changed with requirements for using seat belts when driving for the job. The inter- and intra-personal influences on these behaviors have also changed. Individuals are aware of the need to buckle up and not drink and drive through a variety of influences: the media, pediatric counseling, school programs, and social norms, for example. Taken together, these influences have changed the behaviors of millions of people and dramatically influenced the prospects of improved motor vehicle safety. (pp. 10–11)

Other examples of effective, ecologically focused, and comprehensive injury prevention initiatives include Treno and Holder's (1997) community mobilization program to reduce alcohol-related injury, which focused on changing the social and structural contexts of alcohol use to effectively modify individual behavior. Comparing intervention communities with control communities, this study found significant reductions in alcohol consumption, alcohol-related crashes, and alcohol-related assault injuries. The approach of the Injury Free Coalition for Kids is also ecologically oriented, and its first successful community-based initiative modified the physical environment (e.g., safe play areas) and the social environment (e.g., window guard legislation for high-risk apartments) to achieve behavior change and injury reduction and demonstrated significant declines in pedestrian injury, playground injuries, and overall injuries (InjuryFree 2011; Davidson et al. 1994). A final example is the WHO Safe Communities model (WHO 2011; Svanstrom 2000), which uses a community empowerment model to foster changes in the physical and social environment and individual behavior. Despite methodological limitations in the evidence, the Cochrane Collaboration systematic review concluded that the model was effective for reducing injuries in whole populations (Spinks et al. 2005).

Thus, an ecological approach "provides a complex web of causation and creates a rich context for multiple avenues of intervention" (Allegrante et al. 2006, p. 111). The examples described above illustrate *which* interventions can be effective at multiple levels and in changing physical and social environments. What has not been elucidated is *how* these interventions occur. How do we change laws and regulations? How do we get products and environments changed? How do we reach individuals with effective messages and programs? Who are the key influencers, and how can they be reached and encouraged to reduce injury risk? That is the focus of the next section, with an emphasis on how behavioral sciences can contribute to finding the answers.

Applying Ecological Thinking to the Host, Vector, and Environmental Factors Influencing Injury: Selected Examples of Changing Behavior

The first steps in any behavior change effort are to understand your target audience and specify your behavior change goals. In injury prevention, the vast majority of behavior change research has focused on the public at large or high-risk groups as the target audience and individual risk and protective behaviors as the behavior change goals. Changing individual behavior is rarely possible without the influence of other forces in the individual's interpersonal, organizational, community, and social contexts. Thus, it is important to utilize ecological thinking when attempting to influence individual behavior. This requires an appreciation for and understanding of the many other individuals whose decisions and behaviors influence the injury risks of populations. These are the "influencers" at the various levels of the ecological model. In injury prevention, these include law makers and enforcers at the societal level, industry leaders and manufacturers at the organizational level, advocates at the community level, and so forth. As noted previously, the decisions of these influencers can affect our injury risk directly (e.g., mandating air bags in cars) or indirectly through behavior change (e.g., requiring the use of seat belts).

There has not been a systematic assessment of how to "influence the influencers," despite the well-established importance of comprehensive approaches and the need to change environments and vectors. In this section, we begin to address this gap by using the traditional injury prevention epidemiological framework as a guide. Table 31.1 lists target audiences and behavior change goals for the host, vector, and environmental factors that affect injury risk (Gielen and Girasek 2001). Selected examples are provided to illustrate what can be done to educate and change these target audiences in support of injury reduction. For each example, we provide a brief synopsis of the injury problem and behavior change issues, followed by a description of the literature related to understanding and changing that behavior, highlighting illustrations of constructs from the ecological model and behavioral sciences.

Table 31.1 Using the epidemiological framework to identify behavior change audiences and goals for injury reduction^a

Influencing factors	Possible audiences for behavior change	Possible behavior change goals	Selected examples
(A) Human/host	<ul style="list-style-type: none"> • At-risk individuals • Public at large 	<ul style="list-style-type: none"> • Modify individual personal behavior • Advocate for change in products, environments, and laws 	<ul style="list-style-type: none"> • Fire escape behaviors • Drinking and driving
(B) Vector/ vehicle	<ul style="list-style-type: none"> • Manufacturers • Product designers and engineers 	<ul style="list-style-type: none"> • Make safer products • Make products that are easier to use safely 	<ul style="list-style-type: none"> • Water heaters and scald burn risk • Child-resistant packaging and unintentional poisoning risk
(C) Physical and social environment	<ul style="list-style-type: none"> • Policy makers • Public safety officers • Architects and engineers • Business leaders • Authority figures (coaches, teachers, clinicians) • Media 	<ul style="list-style-type: none"> • Support, create, and enforce laws and regulations promoting safety • Design and create safer environments in schools, workplaces, homes, communities • Make safety products more accessible • Communicate information about injury prevention widely and effectively 	<ul style="list-style-type: none"> • Traffic-calming and pedestrian injuries • Interventions in well child care and pediatric injuries • Media coverage and house fires

^a Adapted from (Gielen and Girasek 2001)

Host

The host factor consists of the individuals at risk for an injury, which can be the public at large or members of specific high-risk groups. Here, we provide examples of behavioral approaches to changing risk and protective behaviors of individuals that illustrate how behavior is related to injury risk and how it has or could be changed to maximize protection of the public. For a more in-depth examination of the behavioral theories behind individual- and community-level changes in injury prevention, readers are referred to Gielen et al. (2006); Gielen and Girasek (2001) and Trifiletti et al. (2005). As shown in Table 31.1, behavior change addressing host factors can be targeted to at-risk individuals or the public at large. Appropriate behavior change goals can include modifying the individual's personal behavior or advocating for change in products, the environment, or laws.

Fire Escape Behaviors

Although the number of fatalities and injuries caused by residential fires has declined gradually over the past several decades, home fires and burns are the third leading cause of fatal home injury in the USA (Runyan et al. 2005). Most victims of fires die from smoke or toxic gases and not from burns (Hall 2001). Groups most at risk of fire-related injuries and deaths include children younger than 5 years, older adults ages 65 and older, African Americans and Native Americans, the poorest Americans, and persons living in rural areas (Istre et al. 2001; Ahrens 2003; Flynn 2010). Fire deaths occur mostly in homes without smoke alarms (Ahrens 2009) and occur during the winter months (Flynn 2010).

There are many resident behaviors that contribute to fire-related injuries. For instance, risky behaviors like drinking alcohol and smoking in bed and use of candles and space heaters contribute to the likelihood that a fire will occur and could be targets for primary prevention efforts. Once a fire occurs, there are behaviors associated with the likelihood that an injury will result, such as whether residents have working smoke alarms and can respond appropriately to the alarm. Very little public health research has been focused on fire response, although some recent work has shown the greater efficacy of parent voice-activated smoke alarms for waking children (Smith et al. 2006) and fire escape planning is receiving increased attention in the public media (e.g., National Fire Protection Association Fire Prevention Week).

The recommendation for fire escape plans is that members of a household make a fire escape plan and practice it at least twice a year (National Fire Protection Association, NFA). The escape plan should incorporate plans for assisting those who are unable to get out without help, such as young children and the frail or disabled. The plan should also identify two different routes of leaving every room in the house and should designate a meeting place outside of the house. According to the National Fire Protection Association (NFPA), 71% of Americans report having an escape plan in case of a fire, but only 45% of these families report having practiced their escape plan (National Fire Protection Association, NFA).

Some efforts to improve the rates of formulating, practicing, and implementing fire escape plans have involved computer modeling to predict fire escape behaviors and to plan egress solutions. "Little has been done in health education, however, to develop appropriate theoretical constructs that would guide promising interventions for escape in the event of a fire" (Thompson et al. 2004). Their literature review identified only two studies of escape planning behavior guided by behavioral theory (Jones et al. 1981; Kronenfeld et al. 1991).

Jones et al. (1981) focused on improving the escape skills of 8- and 9-year-old children using concepts from Social Cognitive Theory. Four scenarios were created that required children to make decisions about completing specific fire escape steps. Five children were evaluated on their performance, given a questionnaire to complete and given training on escaping a fire in a simulated bedroom setting. To enhance the mastery of the skills, a variety of theory-based methods were used, including modeling, role-playing, corrective feedback, social reinforcement, and actual practice. Fifteen days later, a follow-up assessment was performed, and the children maintained their training levels and improved their knowledge.

Kronenfeld et al. (1991) used the cognitive construct of perceived risk from the Theory of Reasoned Action and the Health Belief Model to examine escape planning. Using cross-sectional data from a multiyear, random-digit-dial telephone survey, they hypothesized that perceived risk and parental deficits associated with stress and coping would influence the likelihood of having a fire escape plan. They found the higher the coping skills, the greater the likelihood of having a fire escape plan, while perceived risk and stress were not associated with having a plan. Sociodemographic variables (age of the mother, income, race, number of children) were also significant correlates of having a fire escape plan, suggesting that interventions designed to increase coping among mothers of low-income and minority races may be an effective strategy.

Based on their literature review and work on smoke alarm maintenance, Thompson et al. (2004) created a model for fire escape planning that includes three target behaviors: (1) formulating (or reformulating) a plan, (2) practicing the plan, and (3) implementing the plan. Five behavior change constructs are thought to influence each of the three behaviors: information, attitudes, perceived skill, actual skill, and reinforcement. This example, at the intra- and interpersonal levels of the ecological model, illustrates an injury prevention behavior where there is untapped potential for including behavior change theory in public education such as the annual fire prevention week and other information, education, and communication efforts being conducted by fire departments, health departments, and others.

Drinking and Driving

Alcohol impairment is based on blood alcohol concentration (BAC) of 0.08 g per deciliter (g/dL) or higher at the time of a crash as reported by the Fatality Analysis Reporting System (FARS) or imputed when BAC values are not reported to FARS (National Highway Traffic Safety Administration 2009). According to the National Highway Traffic Safety Administration (2009), there were 10,830 people killed in alcohol-impaired-driving crashes in 2009, which is a dramatic reduction from the 26,173 deaths in 1982 (Hingson and Sleet 2006). Alcohol-related traffic deaths have decreased 48% per 100,000 population and 63% per 100 million vehicle miles of travel, and the greatest proportional decline has been among those aged 15–20 years (Hingson and Sleet 2006). Risk factors for an alcohol-related traffic fatality are aged 21–45 years, being Native American, having previously had a driver's license suspended, speeding, and not wearing a seat belt (Hingson and Sleet 2006).

There are many factors contributing to the success in the USA in reducing the toll of drunk driving. In a clear application of ecological thinking, Hingson and Sleet (2006, p. 238) point out that:

The behavior of driving while intoxicated or under the influence of alcohol is not only shaped by individual choice and motivation, but also strongly associated with organizational, economic, environmental, and social factors. Approaches that use one approach alone to bring about change in alcohol impaired driving are likely to have limited success. Each preventive intervention builds on the strength of every other one.

For the purpose of this example, we examine the role of grassroots organizing as a strategy that focuses not so much on the individual drinking driver, but rather on the public at large to spur changes in the social and political context in which this individual behavior occurs.

Mothers Against Drunk Driving (MADD) has been described as playing “a pivotal role in motivating change” (Hingson and Sleet 2006, p. 249). Candy Lightner started MADD in 1980 following the tragic drunk-driving crash that killed her daughter (Isaacs and Schroeder 2001). MADD used mass media and science to inform their agenda. Stories about victims of drunk drivers and their families were widely covered in all major media, and local MADD chapters spread all over the country. The Institute of Medicine's report on injury prevention (Bonnie et al. 1999) notes that MADD was successful in pressuring policy makers by elevating the visibility of the families of victims and thus influencing the public agenda. Between 1981 and 1985, state legislatures passed 478 laws to deter drunk driving, leading Isaacs and Schroeder (2001) to call the effect of this movement on public policy “stunning.” In 1984, Congress required states to pass the minimum legal drinking age of 21 (MLDA) or risk losing a portion of their federal highway funds, and the MLDA has been cited as particularly effective in reducing drunk driving (Hingson and Sleet 2006). As Graham (1993) describes it, “...changes in social norms, in part spurred by such citizen activist groups as MADD, have apparently achieved what many traffic safety professionals believed was virtually impossible: a meaningful change in driver attitudes and behaviors resulting in a reduction of traffic fatalities” (p. 524).

From a behavioral science perspective, the change in social norms – i.e., the acceptability of the drinking and driving behavior – can be attributed to reframing the behavior from a matter of personal risk to one of imposing risks on others. This is an example of working at the most distal level of the ecological model to influence all of the contexts in which individual drinking behavior is either facilitated or hindered. Social norms and social influences are important concepts in multiple behavior change theories (Simons and Nansel 2006). Understanding both their importance and how to change them is particularly important in injury prevention, where generating public support for policy intervention is often key.

Vector/Vehicle

The vectors for injury (or vehicles, the inanimate vectors) describe how the injury-producing agent reaches the host. These are most often products in the environment, and it is the human–product interaction that allows the transfer of energy in injurious amounts. This concept is well illustrated by motor vehicle crashes and falls from ladders, both of which are significant sources of injury. Behavior change can be focused on the driver or the home do-it-yourselfer (the “host”), but we should also consider the behavior of those who design and manufacture such products. These individuals have the ability to design safer products and to make the safe use of products easier (see Table 31.1). So that generalizable lessons can be learned and applied to other injury problems, here we provide examples of product design changes that addressed these goals and attempt to understand the process by which manufacturers’ behavior was changed.

Water Heaters and Scald Burn Risk

Hot tap water causes about one quarter of all childhood scald burns (most occur in the bathroom), and the damage of hot tap water scalds tends to be more severe than other types of scalds (SAFE KIDS Campaign 2004). A study of burns treated in US emergency departments from 1990 to 2006 reported that there were an estimated 522,988 scald burns, representing 26% of all burns in patients <21 years (D’Souza et al. 2009).

With very hot water, burns can occur quickly. The thermostat setting on hot water heaters in the past was preset at factories to 140 or 150°F. For example, water at 140°F causes a burn within 3 s of exposure (Moritz and Henriques 1947). However, water at 120°F takes approximately 10 min to cause significant thermal injury to the skin. Hence, hot water heaters are ideally preset with this as the maximum temperature to give people time to react and escape the damaging effects. Thus, the solution to scald burns from hot tap water may seem technologically very straightforward – have manufacturers preset the temperature of the water heaters to a safe temperature at the factory.

In Washington State, the number of domestic hot water scalds was reduced by combining an educational campaign with a law that required preset water heater temperatures of 120°F (Erdmann et al. 1991). Five years after the 1983 law went into effect, Erdmann et al. (1991) compared hot water temperatures in homes with the new preset thermostats (“cases”) to homes without them (“controls”). They found that 84% of the cases compared to 70% of the controls had safe hot water temperatures, leading the authors to suggest that increased public awareness about the law as a result of the educational campaign contributed to the overall higher proportions of homes in both groups with safe hot water temperatures relative to earlier prelaw assessments in 1977. The impact of the educational component in this study was not separately evaluated, nor do we know if it was based on any theories of behavior change or principles of adult learning.

An earlier educational campaign in the mid-1980s in Wisconsin distributed thermometers along with utility bills, resulting in lowering the temperature of an estimated 20,000 water heaters (Katcher 1987). However, another study in Philadelphia found that 12 months after community workers tested and adjusted residents’ hot water temperatures, the temperatures were significantly higher in the intervention area relative to a comparison area (Schwarz et al. 1993). This result leads the authors to hypothesize that the intervention families had been educated in how to make the adjustment and chose to turn the temperatures back up to the higher, unsafe levels. Neither of these studies provided information on the theoretical underpinning of the educational messages that were part of their interventions.

Erdmann et al. (1991) provide some clues to changing manufacturer's behavior. The US Consumer Product Safety Commission (CPSC) did not initially support a federal regulation on preset temperatures on water heaters; however, gas water heater manufacturers voluntarily set their heaters to 130°F, and electric water heater manufacturers voluntarily set theirs to 140°F, and both industries placed warning labels on their products. Subsequently, state level advocacy efforts by pediatricians led to some successes and failures, and ultimately, the manufacturers chose to adopt the 120°F standard voluntarily rather than face the prospect of having to meet different requirements in different states.

In parallel efforts to reduce bathtub and shower scald injuries, the CPSC published a report that prompted the American Society of Testing Materials (ASTM) to develop performance requirements related to protection of showers from scalding (George 2011). By 1987, most of the model codes covering the USA included protection against thermal shock and scalding with references to ASSE 1016 with a maximum temperature limit stop set to 120°F. Today, almost all codes in the USA and Canada have thermal shock and scalding protection, but these address running water, not the preset temperature on the water heater. Although the Washington State experience with mandating that residential water heaters be set at 120°F was positive, and both the American Academy of Pediatrics and the American Public Health Association recommend presetting the temperature of water heaters to a safe level, many areas of the USA do not regulate presetting the water heater temperatures (Stauton et al. 2007). Moreover, some in the plumbing engineering community argue that it is preferable to use thermostatic valves to deliver water at the safer 120°F, noting that even when the thermostat is set at a lower safe level, the actual temperature of water stored in the heater can be up to 30°F above the thermostat setting (George 2011). Antiscald devices that residents can install in their existing faucets are available, but the extent to which they are widely accessible or in use, especially among low-income, high-risk families, is not known.

This example illustrates working at a community and organizational level of the ecological model. It also demonstrates the inadequacy of our experience with educational approaches and the complexity of what initially seems like a simple manufacturing solution. First, there have been few educational programs, and none that we could find that were theory-based or derived from an assessment of how people actually test the temperature of their hot water or their knowledge about scald burns. Second, there are multiple engineering solutions and manufacturer standards that come into play to reduce scald burns. It seems that the best path forward involves having the injury prevention public health community and the plumbing engineering community work together to find solutions. Although behavioral scientists could contribute to developing a better understanding of what might change manufacturer behavior (i.e., voluntarily adopt a safer practice), a higher priority, given the lack of consensus on what the technological fix should be, is developing a better understanding of the public's scald prevention concerns and current practices, and more sophisticated efforts to help them protect themselves while product design solutions are improved and made more widely available.

Child-Resistant Packaging and Unintentional Poisoning Risk

Although the mortality rate from childhood poisonings was gradually decreasing from the 1950s, there were still an estimated 2 million unintentional poisoning ingestions annually among young children prior to enactment of the Poison Prevention Packaging Act (PPPA) in 1970 (Walton 1982). Poisonings continued to occur despite the establishment of two new strategies, namely, local poison control centers and the reformulation and repackaging of children's aspirin. In the late 1960s, the number of children's aspirin tablets per bottle was reduced to a sublethal dose (36 per package) for most 2-year olds, and the strength of each tablet was decreased. In 1969, two major manufacturers of baby aspirin voluntarily improved their packaging by producing containers with safety closures. The use of these safety closures was recognized as vitally important to the decline in the number of children poisoned by ingesting baby aspirin that followed this change on the part of manufacturers.

Based on this success, the PPPA established special packaging requirements for an additional 21 products of toxic household substances and medications (Schieber et al. 2000). Unfortunately, the legislative process can be slow. The first congressional hearings on this topic took place 4 years before the law was enacted, 6 years before the first drug was required to be stored in child-resistant containers, 8 years before all prescription drugs were regulated in this manner, and 14 years before the last nonprescription drug (acetaminophen) was so regulated. Nevertheless, in the years that followed, the benefits of this law and its regulations were realized in a big way. Unintentional poisoning due to children's aspirin declined by almost 50% (Walton 1982). Similar declines were also found in other products with revised packaging.

These packaging changes were a significant improvement over previous attempts that relied exclusively on parents and caregivers to lock up medications to protect children. While successful, even the passive approach of using child-resistant containers requires parents to take the behavioral action of replacing lids correctly after each use. Solutions for safe medication storage today are still very limited. Lock boxes and cabinet locks and latches are readily available; however, in studies that have observed medication and household product storage, these are rarely stored safely and often left out in the open where young children can reach them (Gielen et al. 2001; Gielen et al. 2002).

This example demonstrates the power of modifying consumer products to reduce injury risk as well as the need for patience and continued vigilance to make products safer. We do not know the story of what happened "behind the scenes" to influence manufacturers to voluntarily improve their packaging. Nor do we know what ultimately influenced legislators to pass the PPPA. The availability of such information, for example, through in-depth case studies of successes and failures, can only help to advance injury prevention efforts in practice. A better understanding of these issues could also help advance the development of behavior change theories and strategies applied to diverse audiences at these more distal levels of the ecological model.

Physical and Social Environment

The physical and social environment is a critical area of consideration for injury prevention initiatives. It is obvious that injury can occur when individuals interact with elements in their physical environment. The social environment (laws, norms, common practices, etc.) also contributes to injury risk by creating a culture that hinders or facilitates safer behaviors. As displayed in Table 31.1, there is a long list of individuals whose decisions affect the injury risk of populations (e.g., policy makers, authority figures, and media), and behavior change goals for this group include a wide range of opportunities to make the environment safer (e.g., designing safer public spaces, communicating effectively). Here, we provide examples of changing the pedestrian environment, the health care environment for delivering injury prevention anticipatory guidance, and media coverage of injury.

Traffic Calming and Pedestrian Injuries

On average in the USA, 12 pedestrians are killed by motor vehicles every 24 h, an average of one death every 2 h. Another pedestrian is injured every 9 min. This translates into approximately 59,000 pedestrian injuries and 4,092 deaths in 2009 alone (National Highway Traffic Safety Administration 2010). As with other injury areas, children are particularly vulnerable to pedestrian injuries. The good news is that between the years 2000 and 2009, the USA has experienced a 14% decrease in pedestrian injuries (National Highway Traffic Safety Administration 2010). However, some would suggest that such reductions have more to do with fewer pedestrians, especially among children, than to effective pedestrian interventions (Bergman et al. 2002).

As with many injury events, multiple factors contribute to pedestrian injuries and include unsafe pedestrian behaviors, factors related to motor vehicles and driver behaviors, poorly designed or unsafe roadways, sidewalks and intersections, and other special conditions that contribute to pedestrian injuries such as weather conditions and physical limitations of pedestrians. For more than four decades, NHTSA has taken a comprehensive approach to pedestrian safety, being concerned with the interaction between the person and the roadway environment (NHTSA 2008). The coordinated approach has included not only the need to educate pedestrians especially young children since they are overrepresented in the burden of the problem, but also a strategic focus on changing the physical environment, which in turn makes the safer behavior the easier choice (or even default) for both drivers and pedestrians. Here, we focus on an example of modifying the environment – specifically traffic-calming techniques – a classic example of passive countermeasures that protect multiple users through no action of their own.

Traffic calming has been defined as “the combination of mainly physical measures that reduce the negative effects of motor vehicle use, alter driver behavior and improve conditions for non-motorized street users” (Lockwood 1997). Describing the variety of traffic-calming options available is beyond the scope of this chapter. (For more details about traffic calming, readers are directed to Ewing 1999 and Burden 2000). Among the most commonly used techniques are roundabouts, roadway narrowing, partial street closures, speed humps, diverters, and median barriers. Each strategy is designed to address a unique driver or pedestrian issue or hazard. For instance, partial street closures reduce the volume of traffic while roundabouts and speed humps help reduce traffic speed. Medians provide safe refuge for pedestrians in the middle of a roadway.

Two recent Cochrane Collaborations have reviewed different solutions to the pedestrian injury problem. One on safety education to reduce pedestrian injuries (Duperrex et al. 2009) found that although education (especially targeted to children) can enhance their safety knowledge and pedestrian behavior, the extent to which these gains can lead to reductions in deaths and injuries is unknown. The authors also noted the poor quality of the limited number of research studies they found. The other, on traffic calming (Bunn et al. 2009), covered a much wider array of interventions than what has been mentioned here and includes strategies such as enforcement, signals, and financial incentives. Regardless, the review did conclude that traffic-calming approaches “appear to be a promising intervention for reducing traffic injuries and deaths” (p. 12).

Seattle, Washington is often recognized as an early adopter of environmental change to reduce pedestrian injuries. In the early 1970s, their first demonstration project involved a 12-block area known as the Stevens Neighborhood and was designed to reduce “cut-through” traffic through the use of various traffic-calming techniques, mostly diverters, partial road closures, and traffic circles. This early work yielded some important lessons, including the use of temporary environmental modifications to test both their effectiveness and acceptability prior to making permanent changes. Similarly, the project highlighted the need to assess and solicit public support for the changes. Finally, “opting for the most conservative design that will do the job” (Ewing 1999, p. 15) was an important lesson from Seattle’s first experience. Evaluation of the project tracked not only accidents (which declined from 12 to 0 in the 2 years of the project) but also public satisfaction with it, which was generally high.

For more than a decade, the Harborview Injury Prevention and Research Center embarked on a comprehensive approach to pedestrian safety that included the 3 Es, education, enforcement, and environmental change but realized only “modest” improvements in pedestrian injuries (Bergman et al. 2002). Passage of stricter laws and enforcement had little impact on driver behavior (Britt et al. 1995); implementation of a pedestrian curriculum for elementary-age children provided no improvements in half its participants (Rivara et al. 1991). The injury control team (Bergman et al. 2002) decided to move “upstream” to try to influence city planners, engineers, and city politicians to design safer roadway environments. They embarked on a community action campaign that involved broadening the constituency who supported the issue, defining the local pedestrian problem through the

compilation of statistics, and publicizing the personal stories of recent pedestrian victims or their families. Strategically, the team also “redefined” success, not as reducing pedestrian injury deaths (which are relatively rare events) but rather counting the number of their municipalities that would seek available state-level funds for traffic-calming projects and the number that would install them. At the end of their experience, all ten municipalities submitted grant applications and received funds, and seven completed or initiated their plans during the study period. This example clearly illustrates how changing pedestrians’ individual level risk is significantly influenced by the ecological context and how working at the community, organizational, and societal level is necessary.

Interventions in Well Child Care and Pediatric Injuries

The burden of pediatric injuries has long been recognized as a public health problem. An estimated one in four children experiences a medically attended injury each year (Scheidt et al. 1995). Beyond motor vehicle injuries, injury events that occur in and around the home result in significant morbidity and mortality for young children. For each childhood injury death that occurs in the home, another 1,500 children suffer nonfatal injuries; the most common nonfatal injury event is a fall (Casteel and Runyan 2004).

Many of these injuries could be averted or mitigated through the promotion of a constellation of behaviors commonly referred to as “childproofing” practices. These environmental modifications (e.g., use of stair gates, cabinet locks, carbon monoxide alarms, and smoke alarms) reduce injury risk and give parents additional time to intervene when an infant or toddler gets into a potentially dangerous situation. Although parents can and do receive information about childproofing and other safety measures from numerous sources, their child’s pediatrician is a highly respected source, and pediatricians have frequent interactions with families, especially in the child’s early years. The American Academy of Pediatrics presents age- and developmentally appropriate anticipatory guidance that includes recommendations to use safety devices, such as smoke and CO alarms, stair gates, cabinet latches, and locks. The extent to which pediatricians provide effective anticipatory guidance is variable (Gielen et al. 2001; Gielen et al. 2002), and there are many opportunities to enhance the way injury prevention is incorporated into the delivery of pediatric health care (Frame et al. 1997; McDonald et al. 2005).

The SAFE Home Study, a project guided by the Precede–Proceed conceptual framework, is an example of using elements of both organizational change theory to increase pediatricians’ counseling behaviors and the delivery of other injury prevention services, as well as individually oriented theoretical constructs such as risk perceptions and self-efficacy to promote parent childproofing behaviors among individual parents (Gielen and McDonald 2002). The study took place in a large urban, teaching hospital and the focus was on working with pediatric residents to enhance the effectiveness of the injury prevention services they delivered to their patients.

Two important organizational changes were the development of an injury prevention anticipatory guidance training program for pediatric residents and the creation of a children’s safety resource center that provided reinforcement for the pediatrician’s counseling and access to low-cost safety products in the clinical setting. The counseling training program incorporated adult-learning and behavior change theories and resulted in significantly more injury prevention counseling by pediatricians and more satisfied parents relative to parents who received usual care from nontrained physicians (Gielen et al. 2001). Parents who received the enhanced counseling and used the children’s safety center significantly increased their childproofing behaviors (Gielen et al. 2002).

The program was a collaboration between the Johns Hopkins Center for Injury Research and Policy in the Bloomberg School of Public Health and the Johns Hopkins Department of Pediatrics. Implementing the program required substantial buy-in from the leadership of the pediatrics department of the hospital as well as the faculty responsible for the pediatric residency training program.

There were needs for new space, new training time, and a new anticipatory guidance component to routine practice. Successful organizational change was facilitated by everyone's shared goal of preventing injury to children, extensive involvement of the director of the pediatric clinic and residency training program in decision making, and a planning process that was as inclusive and flexible as possible.

This model for delivering injury prevention services in the context of pediatric health care is continuing in the study hospital after more than a decade, and safety centers can now be found in many children's hospitals. Injury topics typically extend beyond home injury to include motor vehicle safety, given its primacy as the leading cause of injury death for children over the age of one (CDC 2007). The extent to which pediatricians universally receive systematic training in effective counseling as part of their residency training is unknown. However, this example suggests that working with the leadership of the health care setting can lead to long-term sustainable organizational change that benefits future clinicians as well as individual patients and families (McDonald et al. 2003).

Media Coverage and House Fires

House fires caused more than \$7 billion worth of property damage, killed more than 2,500, and injured in excess of 13,000 people in 2009 (Karter 2010). Residential fires are a leading source of injury and death, yet there is concern that the general public does not appreciate them as a public health problem. Such lack of awareness may contribute not only to low rates of adoption of individually oriented home safety practices, but also to the lack of support for more policy-oriented interventions, such as funding for fire departments or support for residential sprinklers in new construction. While there are many ways to address this issue, the role of the mass media is critically important for shaping policy agendas and communicating powerfully with the public. As such, mass media is one of the important ways to address the distal elements of the ecological model.

Smith et al. (2007) were the first to examine how print journalists frame the issue of house fires. They monitored four daily newspapers over a 12-month period and examined in detail any article on residential fires. Specifically, they explored issues of location and length of the article, specific content of the article (cause, prevention message, etc.), and whether the article placed residential fires in a public health context. In general, they found that the causes and consequences of residential fires are routinely reported in newspapers, but it is rare that the specific fire incident is placed within a larger public health frame. Prevention messages are rarely included in newspaper coverage of fire incidents. Clearly, current coverage of residential fires is missing an opportunity to educate the public about the public health problem of house fires. Framing residential fires as isolated and individual events is a missed opportunity to engender political will or support for public resources for more innovative fire prevention strategies. As described by Smith et al. (2007): "Successful implementation of upstream policy interventions also requires an informed and supportive public, and the news media is an effective means by which to influence public understanding of policy issues." Smith and coauthors suggest that "public journalism" – combining the objectives of informing the public while promoting the common good – should be applied to the reporting of fire-related incidents by creating media advocacy partnerships.

The Fire Spokesperson's Pocket Media Guide by the Centers for Disease Control and Prevention (2011) is one effort to address this need. When residential fires occur, this can be a valuable opportunity to provide a community with safety messages during a "teachable moment." Public Information Officers (PIOs) play a major role in communicating important fire safety and prevention information to the public and news media. Giving the news media information, they can share with the public about fire prevention and improve the safety of the community. When the public's interest and attention are at a peak because of a recent fire in the community, PIOs can share one or more messages

that encourage viewers/readers to take action that could save a life. The Pocket Media Guide focuses on instructing PIOs what they can do before and during an interview with the news media.

The importance of message framing also extends to focusing on how journalists and other communication professionals position health and social topics in the news. Social and behavioral scientists have realized the need to change reporters' behavior related to communicating public health issues through the news for almost a decade (Chapman 1999). Few examples, however, are available on injury-specific topics (Connor and Wesolowski 2004). Thus, this example highlights the need to build the scholarly and theory-based literature on working with the media to influence the social context and public discourse on injury prevention.

Conclusions

This chapter sought to describe the roles of behavior change in reducing injury, highlight the need for comprehensive approaches that address multiple levels of the ecological model, and provide examples of changes in individual behavior, products, and environments that illustrate the need to consider a variety of audiences and goals for behavior change. Ecological models have been increasingly embraced by public health researchers and practitioners because the problems to be solved are complex and influenced not just by an individual's volitional behavior but by multiple contextual factors that operate at organizational, community, and societal levels. Sometimes, change at these more distal levels directly affect injury risk, as in the banning of a dangerous product (a societal-level intervention), but often change indirectly affects injury risk by influencing human behavior, as in changing drinking and driving behavior by community mobilizing efforts that increase awareness and enforcement (community- and societal-level interventions). Thus, behavior change, which was typically confined to the "host" cell of the Haddon matrix has a much broader challenge – how can the behaviors of those who create the "vectors" and shape the "environments" be influenced to reduce injury risk?

The work presented in this chapter suggests that the answer to this question is equally complex and often not well researched or described in the typical public health/injury prevention literature. While there are numerous examples of successes in reducing injury risk across the influencing factors of the Haddon matrix using interventions at multiple levels of the ecological framework, the processes and theoretical underpinnings of how the changes occurred have not been systematically investigated. Without careful attention to these issues, it is difficult to generalize across situations and benefit from the lessons learned across injury problems.

As the injury field has matured and there are success stories to tell, it is important that we invest in the types of research that will facilitate our understanding of how change occurs at all levels. There are some recent examples of qualitative implementation research (Frattaroli et al. 2010; Frattaroli et al. 2006) that helps to address this need, and more such work should be undertaken in the future to build the theory-based and empirical evidence. Evaluating the complex interventions that operate at multiple levels requires not just new theories, but also new measurement techniques to explain how change in one level affects another level and how comprehensive programs affect outcomes. For example, what are the "active ingredients" in multilevel, community mobilization programs that have demonstrated success in reducing alcohol-related injuries?

Behavioral sciences have demonstrated utility for building programs addressing individuals' risk behaviors. However, behavior change theory and methods offer a largely untapped potential for facilitating change among the people who make laws and design products in ways that can ultimately protect entire populations. Moving forward will require multidisciplinary expertise and new partnerships. Scholars in political science, economics, anthropology, sociology, psychology, and education are but a few examples of partners who could contribute to enhancing our understanding

of the behavior change process across the various audiences. Partnerships have been a mainstay of successful injury prevention efforts in the past and most certainly will be in the future as well. Behavioral scientists can complement the work of epidemiologists and others in the variety of settings where injury prevention research and practice takes place.

References

- Ahrens, M. (2003). *U.S. fire problem overview report: Leading causes and other patterns and trends*. Quincy, MA: National Fire Protection Association.
- Ahrens, M. (2009). *Smoke alarms in U.S. home fires*. Quincy, MA: National Fire Protection Association.
- Allegrente, J. P., Marks, R., & Hanson, D. W. (2006). Ecological models for the prevention and control of unintentional injury. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: Behavioral science theories, methods, and applications* (pp. 105–126). San Francisco, CA: Jossey-Bass.
- Baker, S. P. (1973). Injury control, accident prevention and other approaches to reduction of injury. In P. D. Sartwell (Ed.), *Preventive medicine and public health* (10th ed.). New York: Appleton-Century-Crofts.
- Bergman, A. B., Gray, B., Moffat, J. M., Simpson, E. S., & Rivara, F. P. (2002). Mobilizing for pedestrian safety: An experiment in community action. *Injury Prevent*, 8, 264–267.
- Bonnie, R. J., Fulco, C. E., & Liverman, C. T. (Eds.). (1999). *Reducing the burden of injury, advancing prevention and treatment*. Washington, DC: Institute of Medicine, National Academy Press.
- Britt, J. W., Bergman, A. B., Moffat, L. (1995). *Law enforcement, pedestrian safety, and crosswalk laws: evaluation of a four-year Seattle campaign*. Transportation Research Record No 1485 (pp. 160–167). Washington, DC: National Academy Press.
- Bunn, F., Collier, T., Frost, C., Ker, K., Steinback, R., Roberts, I., Wentz, R. (2009). Area-wide traffic calming for preventing traffic related injuries. 4: CD003110. Available online at: <http://thecochranelibrary.com>.
- Burden, D. (2000). *Streets and sidewalks, people and cars: The citizen's guide to traffic calming*. Sacramento, CA: Local Government Commission Center for Livable Communities.
- Casteel, C., & Runyan, C. W. (2004). Leading causes of unintentional home injury in high-risk groups. In C. W. Runyan & C. Casteel (Eds.), *The state of home safety in America: Facts about unintentional injuries in the home* (2nd ed., pp. 61–68). Washington, DC: Home Safety Council.
- Centers for Disease Control and Prevention. (2011). National Center for Injury Prevention and Control. Fire Spokesperson's Pocket Media Guide; Available at: <http://www.cdc.gov/HomeandRecreationalSafety/pdf/FireSafetyPocketGuide-a.pdf>. Accessed 31 May 2011.
- Centers for Disease Control and Prevention. Web-based Injury Statistics Query and Reporting System (WISQARS). (2010). National Center for Injury Prevention and Control, Centers for Disease Control and Prevention. Available from: URL: www.cdc.gov/ncipc/wisqars.
- Chapman, S. (1999). The news on tobacco control: time to bring the background into the foreground. *Tob Control*, 8, 237–239.
- Connor, S. M., & Wesolowski, K. (2004). Newspaper framing of fatal motor vehicle crashes in four Midwestern cities in the United States, 1999–2000. *Injury Prevention*, 10(3), 149–153.
- D'Souza, A. L., Nelson, N. G., & McKenzie, L. B. (2009). Pediatric burn injuries treated in US emergency departments between 1990 and 2006. *Pediatrics*, 124(5), 1424–1430.
- Davidson, L. L., Durkin, M. S., Kuhn, L., O'Connor, P., Barlow, B., & Heagarty, M. C. (1994). The impact of the safe kids/healthy neighborhoods injury prevention program in Harlem, 1988 through 1991. *American Journal of Public Health*, 84(4), 580–586.
- Duperrex, O. J. M., Roberts, I. G., Bunn, F. (2009). Safety education of pedestrians for injury prevention. The Cochrane Collaboration (pp. 1–13). Available online at: <http://www.thecochranelibrary.com>. Accessed 15 Mar 2011.
- Erdmann, T. C., Feldman, K. W., Rivara, F. P., Heimbach, D. M., & Wall, H. A. (1991). Tap water burn prevention: The effect of legislation. *Pediatrics*, 88, 572–577.
- Ewing, R. (1999). Traffic Calming: State of the Practice. Prepared for the US Department of Transportation, Federal Highway Administration, Office of Safety Research and Development. FHWA-RD-99-135.
- Fishbein, M. (2006). Foreword. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: Behavioral sciences theories, methods, and applications* (p. x). San Francisco, CA: Jossey-Bass.
- Flynn, J. D. (2010). *Characteristics of home fire victims*. Quincy, MA: National Fire Protection Association.
- Frame, P. S., Berg, A. O., & Wolfe, S. (1997). U.S. preventive services task force: Highlights of the 1996 report. *American Family Physician*, 55(2), 567–576. 581–2.

- Frattaroli, S., DeFrancesco, S., Gielen, A. C., Bishai, D. M., & Guyer, B. (2006). Local stakeholders' perspectives on improving the urban environment to reduce child pedestrian injury: Implementing effective public health interventions at the local level. *Journal of Public Health Policy, 27*(4), 376–388.
- Frattaroli, S., Pollack, K. M., Jonsberg, K., Crotwau, G., Rivera, J., & Mendel, J. S. (2010). Streetworkers, youth violence prevention, and peacemaking in Lowell, Massachusetts: Lessons and voice from the community. *Progress in Community Health Partnership: Research, Education and Action, 4*(3), 171–179.
- George, R. L. (2011). Plumbing to prevent domestic hot water scalds. Available online at: <http://www.hgexperts.com/article.asp?id=5135>, Accessed 10 Apr 2011.
- Gielen, A. C., & Girasek, D. (2001). Integrating perspectives on the prevention of unintentional injuries. In N. Schneiderman, M. A. Speers, J. M. Silva, H. Tomes, & J. H. Gentry (Eds.), *Integrating behavioral and social sciences with public health* (pp. 203–227). Washington, DC: American Psychological Association.
- Gielen, A. C., & McDonald, E. M. (2002). Using the PRECEDE-PROCEED planning model to apply health behavior theories. In K. Glanz, B. Rimer, & F. Lewis (Eds.), *Health behavior and health education: Theory, research and practice* (3rd ed., pp. 409–436). San Francisco: Jossey-Bass.
- Gielen, A. C., McDonald, E. M., Wilson, M. E., Hwang, W. T., Serwint, J. R., Andrews, J. S., & Wang, M. C. (2002). Effects of improved access to safety counseling, products and home visits on parents' safety practices: Results of a randomized trial. *Archives of Pediatrics and Adolescent Medicine, 156*(1), 33–40.
- Gielen, A. C., & Sleet, D. A. (2006). Injury prevention and behavior: An evolving field. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: Behavioral science theories, methods, and applications*. San Francisco, CA: Jossey-Bass.
- Gielen, A. C., Sleet, D. A., & DiClemente, R. J. (2006). *Injury and violence prevention: Behavioral science theories, methods, and applications*. San Francisco, CA: Jossey-Bass.
- Gielen, A. C., Wilson, M. E., McDonald, E. M., Serwint, J. R., Andrews, J. S., Hwang, W. T., & Wang, M. C. (2001). Randomized trial of enhanced anticipatory guidance for injury prevention. *Archives of Pediatrics and Adolescent Medicine, 155*(1), 42–49.
- Glassbrenner, D., Ye, T. J. (2007). Seat Belt Use in 2007 – Overall Results. Traffic Safety Facts Research Note. National Center for Statistics and Analysis, National Highway Traffic Safety Administration, DOT HS 810 841.
- Graham, J. D. (1993). Injuries from traffic crashes: Meeting the challenge. *Annual Review of Public Health, 14*, 515–543.
- Green, L. W., & Kreuter, M. W. (2005). *Health program planning: An educational and ecological approach* (4th ed.). New York: McGraw-Hill.
- Grossman, D. C. (2006). Foreword. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: Behavioral science theories, methods, and applications* (p. xiv). San Francisco, CA: Jossey-Bass.
- Haddon, W. (1970). On the escape of tigers: An ecological note. *American Journal of Public Health, 60*, 2229–2235.
- Haddon, W. (1980). Advances in the epidemiology of injuries as a basis for public policy. *Public Health Reports, 95*, 411–421.
- Hall, J. R. (2001). *Burns, toxic gases, and other hazards associated with fires: Deaths and injuries in fire and non-fire situations*. Quincy, MA: National Fire Protection Association, Fire Analysis and Research Division.
- Hanson, D., Hanson, J., Vardon, P., McFarlane, K., Lloyd, J., Muller, R., & Durrheim, D. (2005). The injury iceberg: An ecological approach to planning sustainable community safety interventions. *Health Promotion Journal of Australia, 16*(1), 5–15.
- Hingson, R., & Sleet, D. A. (2006). Modifying alcohol use to reduce motor vehicle injury. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: Behavioral science theories, methods, and applications* (p. 249). San Francisco, CA: Jossey-Bass.
- InjuryFree. (2011). About the injury free coalition for kids. Available online at: http://injuryfree.org/about_history.cfm.
- Isaacs, S. L., & Schroeder, S. A. (2001). Where the public good prevailed. *The American Prospect, 12*(10), 1–10.
- Istre, G. R., McCoy, M. A., Osborn, L., Barnard, J. J., & Bolton, A. (2001). Deaths and injuries from house fires. *New England Journal of Medicine, 344*, 1911–1916.
- Jones, R. T., Kazdin, A. E., et al. (1981). Social validation and training of emergency fire safety skills for potential injury prevention and life saving. *Journal of Applied Behavior Analysis, 14*(3), 249–260.
- Karter, M. J. (2010). *Fire Loss in the United States during 2009*. Quincy, MA: National Fire Protection Association, Fire Analysis and Research Division.
- Katcher, M. L. (1987). Prevention of tap water scald burns: Evaluation of a multi-media injury control program. *American Journal of Public Health, 77*(9), 1195–1197.
- Kronenfeld, J., Gliik, D., & Jackson, K. (1991). Home fire safety and related behaviors among parents of preschoolers. *Children's Environments Quarterly, 1*(8), 31–40.
- Lockwood, I. M. (1997). ITE traffic calming definition. *Internat Traffic Engin Journal, 67*, 22–24.

- McDonald, E. M., Gielen, A. C., Trifiletti, L. B., Andrews, J. S., Serwint, J. R., & Wilson, M. E. (2003). Evaluation activities to strengthen an injury prevention resource center for urban families. *Health Prom Practice, 4*(2), 129–137.
- McDonald, E. M., Solomon, B., Shields, W., Serwint, J. R., Jacobsen, H., Weaver, N. L., Kreuter, M., & Gielen, A. C. (2005). Evaluation of kiosk-based tailoring to promote household safety practices in an urban pediatric primary care practice. *Patient Education and Counseling, 58*, 168–181.
- McLeary, K. R., Bibeau, D., Steckler, A., & Glanz, K. (1988). An ecological perspective on health promotion programs. *Health Education Quarterly, 15*, 351–377.
- MMWR. (1999a). Achievements in public health, 1900–1999: Motor vehicle safety: A 20th century public health achievement. *MMWR, 48*(18), 369–374.
- MMWR. (1999b). Achievements in public health, 1900–1999: Improvements in workplace safety – United States, 1900–1999. *MMWR, 48*(22), 461–469.
- Moritz, A. R., & Henriques, F. C. (1947). Studies of thermal injury: II. The relative importance of time and surface temperature in the causation of cutaneous burns. *The American Journal of Pathology, 23*(5), 695–720.
- National Committee for Injury Prevention and Control. (1989). Injury prevention: Meeting the challenge. *American Journal of Preventive Medicine, 5*(3), 1–303. New York: Oxford University Press.
- National Fire Protection Association (NFPA). No Date Available (NDA). Escape Planning. <http://www.nfpa.org/assets/files/PDF/Public%20Education/EscapePlanningTips.pdf>.
- National Highway Traffic Safety Administration. (2009). Traffic safety facts – 2009 data – impaired driving. Available online at: <http://www-nrd.nhtsa.dot.gov/Pubs/811385.PDF>. Accessed 31 May 2011.
- National Highway Traffic Safety Administration. (2010). Traffic safety facts – 2009 data – pedestrians. Available online at: <http://www-nrd.nhtsa.dot.gov/Pubs/811394.pdf>. Accessed 11 Apr 2011.
- Nichols, J. L. (1994). Changing public behavior for better health: Is education enough? *American Journal of Preventive Medicine, 10*(3S), 19–22.
- Rivara, F. P., Booth, C. L., Bergman, A. B., Rogers, L. W., & Weiss, J. (1991). Prevention of pedestrian injuries in children: Effectiveness of a school training program. *Prevention, 88*(4), 770–775.
- Rivara, F., Thomson, D. C., Beahler, C., & MacKenzie, E. M. (1999). Systematic reviews of strategies to prevent motor vehicle injuries. *American Journal of Preventive Medicine, 16*(Suppl 1), 1–5.
- Runyan, C. W., Casteel, C., Perkis, D., Black, C., Marshall, S. W., Johnson, R. M., Coyne-Beasley, T., Waller, A. E., & Viswanathan, S. (2005). Unintentional injuries in the home in the United States: Part I: mortality. *American Journal of Preventive Medicine, 28*(1), 73–79.
- SAFE KIDS Campaign. (2004). Burn injury fact sheet. Available online at: http://www.preventinjury.org/PDFs/BURN_INJURY.pdf. Accessed 12 Mar 2010.
- Sallis, J. F., Owen, N., & Fisher, E. B. (2008). Ecological models of health behavior. In K. Glanz, B. K. Rimer, & K. Viswanath (Eds.), *Health behavior and health education: Theory, research and practice* (pp. 465–485). San Francisco, CA: Jossey-Bass.
- Scheidt, P. C., Harel, Y., Trumble, A. C., Jones, D. H., Overpeck, M. D., & Bijar, P. E. (1995). The epidemiology of nonfatal injuries among US children and youth. *American Journal of Public Health, 85*(7), 932–938.
- Schieber, R. A., Gilchrist, J., & Sleet, D. A. (2000). Legislative and regulatory strategies to reduce childhood unintentional injuries. *The Future of Children: Unintentional Injuries in Childhood, 10*(1), 111–136.
- Schwarz, D. F., Grisso, J. A., Miles, C., Holmes, J. H., & Sutton, R. L. (1993). An injury prevention program in an urban African-American community. *American Journal of Public Health, 83*, 675–680.
- Simons, B. G., & Nansel, T. (2006). The application of social cognitive theory to injury prevention. In A. C. Gielen, D. A. Sleet, & R. J. DiClemente (Eds.), *Injury and violence prevention: Behavioral science theories, methods, and applications* (pp. 41–64). San Francisco, CA: Jossey-Bass.
- Smith, K. C., Cho, J., Gielen, A. C., & Vernick, J. (2007). Newspaper coverage of residential fires: An opportunity for prevention communication. *Injury Prevention, 13*, 110–114.
- Smith, G. A., Splaingard, M., Hayes, J. R., & Xiang, H. (2006). Comparison of a personalized parent voice smoke alarm with a conventional residential tone smoke alarm for awakening children. *Pediatrics, 118*, 1623–1632.
- Spinks, A., Turner, C., Nixon, J., McClure, R. (2005). The WHO safe communities' model for the prevention of injury in whole populations. *Cochrane Database Systematic Reviews, 18*(2), CD004445.
- Stauton, C. E., Frumpkin, H., & Dannenberg, A. L. (2007). Changing the built environment to prevent injury. In L. S. Doll, S. E. Sandra, P. A. Sleet, & J. A. Mercy (Eds.), *Handbook of injury and violence prevention* (pp. 257–275). New York, NY: Springer.
- Stokols, D. (1992). Establishing and maintaining healthy environments: Toward a social ecology of health promotion. *American Psychologist, 47*(1), 6–22.
- Svanstrom, L. (2000). Evidence-based injury prevention and safety promotion: State of the art. In D. Mohan & G. Tiwari (Eds.), *Injury prevention and control* (pp. 181–198). London: Taylor and Francis.
- Thompson, N. J., Waterman, M. B., & Sleet, D. A. (2004). Using behavioral science to improve fire escape behaviors in response to smoke alarms. *Journal of Burn Care and Rehabilitation, 45*, 179–188.

- Treno, A. J., & Holder, H. D. (1997). Evaluating effort to reduce community-level problems through structural rather than individual change: A multicomponent community trial to prevent alcohol-involved problems. *Evaluation Review, 21*, 133–139.
- Trifiletti, L. B., Gielen, A. C., Sleet, D. A., & Hopkins, K. (2005). Behavioral and social sciences theories and models: Are they used in unintentional injury prevention research? *Health Education Research, 20*(3), 298–307.
- Turner, S., Lyons, A. G., Weightman, A. L., Mann, M. K., Jones, S. J., John, A., Lannon, S. (2011). Modification of the home environment for the reduction of injuries (Review). The Cochrane Collaboration. 16(2): CD003600. Available online: <http://www.thecochranecollaboration.com>.
- US Preventive Services Task Force. (2007). Counseling about proper use of motor vehicle occupant restraints and avoidance of alcohol use while driving: U.S. Preventive Services Task Force recommendations statement. *Annals of Internal Medicine, 147*(3), 187–193.
- Waller, P. (2001). Public health's contribution to motor vehicle injury prevention. *American Journal Preventive Medicine, 21*(4S), 3–4.
- Walton, W. W. (1982). An evaluation of the Poison prevention packaging act. *Pediatrics, 69*(3), 363–370.
- WHO Collaborating Center on Community Safety Promotion. (2011). Safe communities network members. Available online at: http://www.phs.ki.se/csp/who_safe_communities_network_en.htm.
- Wilson, M. H., Baker, S. P., Teret, S. P., Shock, S., & Garbarino, J. (1991). *Saving children*. New York, NY: Oxford University Press.
- Zwerling, C., & Jones, M. P. (1999). Evaluation of the effectiveness of low alcohol blood concentration laws for young drivers. *American Journal of Preventive Medicine, 16*(1), 76–80.

Chapter 32

EMS and Trauma Systems

Lenora M. Olson and Stephen M. Bowman

Introduction

Injury is a leading cause of death and disability in the USA (CDC 2007). As several chapters in this book have shown, the prevention of injuries is the best way to save lives. Unfortunately, we cannot prevent all injuries. As a result, when a person is injured, timely treatment of the injury is needed to mitigate the impact. Trauma systems were developed to provide immediate and coordinated care of the injured patient. A trauma system consists of three major providers and associated components – prehospital, acute care, and rehabilitation organized in a defined geographic area to deliver timely care to an injured patient from time of injury through transport to acute care facility and to rehabilitation (MacKenzie et al. 2003).

In this chapter, we first describe a brief history of trauma systems development in the USA. We define the key characteristics and purpose for trauma systems to lay a foundation for issues related to trauma systems development and how these issues affect available data and past and current research. We briefly describe these issues including how configurations of trauma systems differ by state, regional, and local circumstances. We then introduce and discuss the advantages and disadvantages of national and state trauma datasets and highlight examples of trauma systems research from the 1960s to the present. We conclude with future directions and challenges.

Overview of Trauma Systems Research

Since the National Academy of Sciences first labeled injury as the “neglected disease” in 1966 (National Academy of Sciences 1966), trauma care has evolved from an emphasis on wartime injuries into coordinated regional systems dedicated to reducing injury incidence, disability, and mortality among civilian populations. Trauma systems are designed to serve a defined geographic area by integrating the full range of medical services required by injured persons, from initial care at the

L.M. Olson, PhD (✉)

Intermountain Injury Control Research Center, University of Utah Department of Pediatrics,
Salt Lake City, UT, USA
e-mail: lenora.olson@hsc.utah.edu

S.M. Bowman, PhD

Center for Injury Research and Policy, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
e-mail: smb Bowman@jhsph.edu

scene to timely transport of the seriously injured patient to specialist hospitals (Level 1 trauma centers) to long-term rehabilitation for restoration of functionality. The premise underlying the development of state and regional trauma systems is that an organized system of trauma care ensures that critically injured patients are appropriately triaged and/or transferred to high-quality definitive care without delay (Nathens et al. 2000a). While this sounds like a simple premise, the development of trauma systems and associated research is more complex (see Mullins 1999 for an in depth historical perspective regarding trauma system development and research). Indeed, proof of the effectiveness of trauma systems is an ongoing research issue. In response to the 1966 National Academy report, the American College of Surgeons (ACS) developed criteria for the designation of trauma centers in 1976. The American College of Surgeons Committee on Trauma (ACS-COT) is an organization that maintains a voluntary verification and consultation program to assist states in the development of trauma systems and the verification and/or designation of trauma centers. Progress was slow. Even though positive outcomes for patient survival had been reported since the early 1990s, by 1997, only 22 states had fulfilled the criteria for a state trauma system (Nathens et al. 2000a). Today, 50 states and the District of Columbia have established trauma systems that include the formal designation or verification of trauma centers (ACS 2011).

The goal of trauma systems research is to identify the optimal components of the trauma system that reduce morbidity and mortality, while efficiently using limited resources. As trauma systems and their components have evolved over the last four decades, so have research and interest in examining the efficacy of such systems. In the early 1970s, many state agencies began developing trauma systems. For the next decade, research concentrated on trauma system development and implementation (West et al. 1988; Bazzoli et al. 1995). Not surprisingly, research from this time period concentrated on mortality as the outcome, with many studies comparing survival of injured patients in states with a trauma system to states that did not have a system or comparing outcomes before the implementation of the trauma system in the state to after implementation. Studies from this time period generally found that severely injured patients had improved chances of survival if they were treated within a regionalized system of trauma care (e.g., Cales and Trunkey 1985; Shackford et al. 1986).

In the last two decades, much of the research in trauma centers has continued to focus on patient outcomes, particularly survival, to assess trauma system effectiveness. Mann et al. (1999) conducted a systematic review of the published evidence on trauma system effectiveness using articles published from 1966 to 1998. Articles were categorized: (a) panel review, (b) comparison to national criterion injury registry, or (c) population-based studies. The authors found that the evidence does support that organized systems of trauma care are effective in reducing *deaths*. Panel studies found up to a 50% reduction in preventable deaths with the implementation of trauma centers. Trauma registry data overall showed a 15% reduction when compared with the Major Trauma Outcome Study (MTOS) norms and population-based studies demonstrated a 15–20% reduction in the risk of death after trauma system implementation. As the authors state, all of the studies could be improved by including prehospital and post-discharge trauma deaths, standardizing trauma registry inclusion criteria, and developing a contemporary national reference norm for trauma outcome.

Today, trauma systems contain some or all of the following components: prehospital care, medical direction, transportation, hospital care, communication, training, triage, medical evaluation, public education, prevention, and rehabilitation (ACS-COT 2006). State and regional oversight agencies establish and coordinate the diverse components of a trauma system and evaluate system performance (National Academy of Sciences 1999). The regional or state systems have a variety of administrative structures ranging from organization at the state level, a regional level, or a combination of both. The diverse oversight of trauma systems is both a strength and a weakness. By having local or state control, a trauma system can be more adaptable to a state's needs and geography, but the configuration of trauma systems varies from state to state, thus making comparisons of outcomes

and performance between states problematic. For example, the majority of trauma systems are considered “inclusive” and allow all hospitals and acute care facilities in a designated region to participate in the trauma system, at varying levels of trauma designation (Mann et al. 2005). However, there are a few systems that are “exclusive” in that they are organized around a Level 1 trauma center and mainly treat only major or severe trauma patients in a region (Lansink and Leenen 2007). In addition, each of the separate components that are part of a trauma system (prehospital, acute care, and rehabilitation) has its own system of data collection and data elements – there is not a federally supported national surveillance system for trauma. As a result, most trauma system research is conducted using secondary data sets. For the most part, trauma surveillance systems are managed by state trauma oversight agencies (typically state health departments) and/or individual trauma centers that maintain hospital-based trauma registries. While a core set of standardized variables exist for trauma registries who submit to the National Trauma Data Bank (NTDB), there is still variability in the scope of individual trauma registries, with some being more comprehensive than others.

It is not only implementation of a regional or state wide trauma system but also maturation of the system that may affect patient outcomes as systems are evaluated and refined. States with mature trauma systems may have a greater effect in reducing injury mortality (Durham et al. 2006). A study conducted in Oregon showed a 35% reduction in mortality 2 years after implementation of the trauma system (Mullins et al. 1994). Nathens et al. (2000b), examining the effect of trauma system care on motor vehicle crash deaths, found that mortality decreased by 8% after a state had a system in place for at least 10 years. Finding that deaths are reduced after the implementation of a trauma system is a start to understanding effectiveness, but it is not enough. Bazzoli et al. (1995) point out that while studies show improved patient outcomes after implementation of trauma systems, most of the studies do not pinpoint the specific system characteristics that lead to improved outcomes.

Data Sources for Trauma Systems Research

Interest in the effectiveness of trauma systems continues to receive considerable attention among injury researchers. With significant financial and human capital commitments, hospital administrators and physicians expect evidence-based policy with regard to the structure and organization of trauma systems. The complexity of trauma systems, the regulatory and political constraints, and the need to use secondary data sets are challenging for the injury researcher. As trauma systems encompass a continuum of care from first responders to definitive hospital care, comprehensive data are essential to this research.

To conduct trauma systems research, injury researchers need access to a variety of patient-level data sources. To study the effectiveness and best practices of organized trauma systems, reliable and complete data from each component of care is needed. Table 32.1 illustrates some of the secondary data sources available for trauma systems research, with select examples of published studies.

Emergency Medical Services (EMS) Data Systems

Virtually all EMS providers document their encounters with patients in the prehospital setting. However, the degree to which these data are available for research purposes varies greatly. Some rural volunteer EMS agencies collect paper run sheets and may or may not enter data into an

Table 32.1 Data sources for trauma systems research

Data source	Description	Examples of use for trauma systems research
<i>Emergency medical services data</i>		
<ul style="list-style-type: none"> National EMS Information System (NEMSIS) State EMS Data Systems Local and Regional EMS Data Systems EMS Agency Data 	Patient-level encounter data from EMS providers, including patient and injury characteristics, response type, level of personnel, prehospital treatment(s), transport destination	<ul style="list-style-type: none"> Maryland State EMS data used to study undertriage of elderly trauma patients to state-designated trauma centers (Chang et al. 2008) Seven county EMS data (Southwest Alabama) used to assess mortality in rural vehicular trauma and identify contributing factors (Gonzalez et al. 2006) The Detroit Fire Department EMS data used to describe the epidemiology of pediatric transports and non-transports in an urban emergency medical services system (Kannikeswaran et al. 2007) Fire Department of the City of New York EMS data used to examine utilization of air medical transport in a large urban environment (Asaeda et al. 2001)
<i>Emergency department data</i>		
<ul style="list-style-type: none"> National ED Survey (NHAMCS) State Emergency Department Data (SEDD) State Emergency Department Datasets Hospital/Trauma Center ED Data 	Patient and injury characteristics, admitted and non-admitted cases, treatment performed in the ED, charges, disposition	<ul style="list-style-type: none"> The National Hospital Ambulatory Medical Care Survey used to describe the epidemiology of emergency medical services use by children (Shah et al. 2008) State Emergency Department Data used to characteristic children's utilization of injury-related emergency department care (Owens et al. 2008)
<i>Hospital discharge data</i>		
<ul style="list-style-type: none"> Nationwide Inpatient Sample (NIS) Kid's Inpatient Database (KID) Medicare Fee-for-Service Data State Inpatient Database (SID) State Hospital Discharge Data 	Patient characteristics, external cause of injury codes (E-codes) diagnoses, and procedures, length of stay, charges, payer(s), discharge disposition	<ul style="list-style-type: none"> Nationwide Inpatient Sample (NIS) used for population-based survival assessment of categorizing Level III and IV rural hospitals as trauma centers (Arthur et al. 2009) NIS used to examine trauma designation and its relationship to outcomes among hospitals in rural communities (Bowman et al. 2008) Medicare fee-for-service records used to examine regional variation in mortality for injured Medicare patients (Gorra et al. 2008) Kid's Inpatient Database used to assess hospital characteristics associated with optimal management of children with spleen injuries (Bowman et al. 2005) Florida hospital discharge data used to evaluate a mature trauma system (Durham et al. 2006) and to assess effectiveness in lowering mortality rate (Papa et al. 2006)
<i>Trauma registries</i>		
<ul style="list-style-type: none"> National Trauma Data Bank State Trauma Registries Hospital/Trauma Center Registries 	Patient characteristics, injury description and E-codes, diagnoses and procedures, Abbreviated Injury Scale (AIS), Injury Severity Score, intensive care unit utilization, complications, comorbidities, length of stay, charges, payer(s), discharge disposition	<ul style="list-style-type: none"> National Trauma Data Bank used to examine volume-outcome relationships for Level I trauma centers (Bennett et al. 2010) New York State trauma data used to study direct transport within an organized state trauma system and its association with reduced mortality in patients with severe traumatic brain injury (Härtel et al. 2006) Level I hospital trauma registry used to compare surgeon- and system-based influences on trauma mortality (Haut et al. 2009) Level I hospital trauma registry used to conduct a propensity score analysis of prehospital factors and directness of transport of major trauma patients to a Level I trauma center (Garwe et al. 2010)

(continued)

Table 32.1 (continued)

Data source	Description	Examples of use for trauma systems research
<i>Death certificate/vital records</i>		
<ul style="list-style-type: none"> National Death Index Fatality Analysis Reporting System (FARS) Multiple Cause of Death Data (NCHS) State Vital Records Medical Examiner/Coroner Reports 	Trauma decedent characteristics, location of death, underlying cause of death, deaths occurring after acute care hospital discharge, autopsy status	<ul style="list-style-type: none"> National Death Index with cause of death codes, linked Level I trauma registry data and census data to evaluate risks for late death after injury (Claridge et al. 2010) National Death Index used to assess mortality in chronic spinal cord injury (Garshick et al. 2005) Fatality Analysis Reporting System data used to measure the association between urban sprawl and EMS response time and test the hypothesis that features of urban sprawl development increase the probability of delayed ambulance arrival (Trowbridge et al. 2009) Colorado vital statistics data used for a population-based study of mortality after discharge from acute care hospitalization with traumatic brain injury (Ventura et al. 2010) Miami-Dade County Medical Examiner data used to assess resuscitation-related injuries and outcomes among infants and children (Matshes and Lew 2010)
<i>Rehabilitation records</i>		
<ul style="list-style-type: none"> Uniform Data Set for Medical Rehabilitation (UDSMR) Trauma rehabilitation programs 	Continuum of care, post-acute medical treatment and utilization, rehabilitation services	<ul style="list-style-type: none"> UDSMR data used to compare and evaluate inpatient rehabilitation for older adults with hip fracture (Kortebein et al. 2009) UDSMR data used to examine race/ethnicity and outcomes following inpatient rehabilitation for hip fracture (Graham et al. 2008)

electronic system. Conversely, some high volume urban EMS agencies may use handheld or tablet computers to enter all run sheet information, with subsequent wireless transmission to a central database. In between, many hybrid options are in use. For example, some EMS providers document procedures and demographic data on paper, but rely on computer-aided dispatch systems to track response and transport times.

To standardize EMS reporting and facilitate research efforts, the National Emergency Medical Services Information System (NEMSIS) now exists (NEMSIS 2011). Under NEMSIS, a national data dictionary allows software vendors and EMS providers to use a standardized set of data elements and definitions. Simple and secure electronic data submissions aim to decrease the burden associated with reporting. Ultimately, NEMSIS will support and maintain a National EMS Database that can be used for research on trauma systems. However, participation in NEMSIS is voluntary. Currently, 25 states are submitting data to the NEMSIS program (NEMSIS 2011).

Linking of data systems is often critical to understanding the care provided in the prehospital setting. Multiple EMS providers may provide care to any given patient. For example, a first responder may arrive at the scene of a crash to begin patient assessment and extrication. A paramedic unit may subsequently be dispatched to provide advanced life support. Finally, an air medical transport team may be mobilized to fly the patient to a Level I trauma center. To fully understand the prehospital care provided to this patient, linking of EMS care reports among different providers at the scene is critical. In addition, obtaining and linking outcomes from the hospital is also essential for both quality improvement and research purposes.

Trauma Registries

Trauma registries are a key component of any designated and/or verified trauma center. At the individual trauma center level, registries are used for quality improvement, quality assurance, planning and resource allocation, injury prevention, and compliance documentation for state and local rules and regulations. Trauma registries are *not* population-based data systems. That is, there are explicit inclusion and exclusion criteria that determine which patients are to be included in the registry. State trauma programs typically adopt statewide trauma registry inclusion criteria that specify which cases must be abstracted and submitted to the state trauma registry. For example, some state trauma programs have established a minimum length of hospital stay for cases to be included in the trauma registry (e.g., greater than 2 days). This limits the volume of minor trauma patients who may stay in the hospital for a single night. Some hospitals will voluntarily choose to include all hospitalized patients. Similarly, some state registries may require all injury hospitalizations, regardless of length of stay. For research purposes, it is critically important to understand the data inclusion and exclusion criteria for the specific study.

At the national level, the ACS provides leadership in the area of trauma registries, with both established guidelines for hospital trauma registries, and an ongoing initiative to create and maintain the NTDB. For 2009, the NTDB received 681,990 records from 682 trauma facilities. For the researcher, the NTDB offers several research data sets. The NTDB research data set contains all records sent to the NTDB for each admission year. Alternatively, a National Sample Program (NSP) research data set is available, containing weighted data for up to 100 randomly selected trauma centers. This data set can be used for estimating adult patients seen in Level I and II trauma centers as well as for trend analysis across years. An important limitation to the NTDB is the voluntary participation of trauma centers and the lack of data from non-trauma center hospitals.

Hospital Discharge Data

In the absence of more detailed trauma registry data, hospital discharge data are frequently used for trauma systems research. One advantage of this data source is the population-based nature of hospital discharge data. Typically, all acute care hospitals (with the exception of federally run hospitals) are included, with all discharges captured. The limitation is that the level of detail does not approach trauma registries, with a significant gap in clinically relevant data such as physiologic measures (e.g., systolic blood pressure, respirations, and pulse) that contribute substantially to adequate case mix adjustment. Hospital discharge data do offer the potential to analyze differences between trauma centers and non-trauma centers – the latter of which are not included in trauma registry data sets.

Research on Effectiveness of Trauma Systems

Historically, retrospective studies of trauma systems effectiveness were most frequently conducted and subsequently published in the literature. Retrospective studies, while lacking rigor in design, are relatively simple to conduct, with readily available data from state or regional trauma registries and/or state hospital discharge databases. Researchers often focus on comparing the outcomes and performance in trauma centers with that of similar non-trauma hospitals. In addition, retrospective studies have been used to compare states with formal trauma systems to states without such systems.

As part of the American College of Surgeon's trauma verification program, trauma centers are required to have active trauma research programs. Retrospective studies are a low-cost means of meeting the research requirements. Some recent examples of retrospective trauma studies include:

- Comparing pediatric trauma patient mortality at designated trauma centers and non-designated hospitals in Florida (Pracht et al. 2008).
- Comparing scoop and run to a trauma center versus initial care at a local hospital prior to transfer and the effect of transfer on mortality (Nirula et al. 2010).
- Assessing the effects of trauma center care, admission volume and surgical volume on paralysis after spinal cord injury (Macias et al. 2009).
- Examining survival of seriously injured patients first treated in rural hospitals (Mullins et al. 2002).
- Evaluating undertriage of elderly trauma patients to state designated trauma centers (Chang et al. 2008).
- Evaluating long-term mortality trends from injury in a mature, inclusive Canadian trauma system (Moore et al. 2010), and
- Assessing inclusive trauma systems and whether they improve triage or outcomes of the severely injured patient (Utter et al. 2006).

The common themes of these retrospective studies are (1) trauma systems improve outcomes and (2) trauma patients do better at trauma centers than at non-trauma centers.

In contrast to the extensive historical focus on acute care or in-hospital trauma outcomes, prehospital and post-acute trauma rehabilitation has received relatively little attention in the literature when examining the effectiveness of trauma systems. One area of effectiveness research in the prehospital care arena is the use of air medical service or air ambulance (rotor wing helicopters and fixed-wing aircraft). For more than 25 years, air medical services have been a part of organized trauma systems yet limited data exist on the medical effectiveness of using an air ambulance over a ground ambulance to transport an injured patient from the scene of the trauma. Air medical services are believed to improve outcomes for an injured patient due to reduced transport time to definitive care as well as providing a higher level of care during transport. Branas et al. (2005) estimated that medical helicopters provided access for 81.4 million Americans who otherwise would not be able to reach a trauma center within an hour. Several studies reported a reduced mortality of 20–50% in severely injured trauma patients transported by helicopter (e.g., Brown et al. 2010; Davis et al. 2005). On the other hand, some studies report little to no improvement in outcomes of air ambulance service relative to ground service (e.g., Biewener et al. 2004; Chappell et al. 2002). One reason may be that gains in transport time do not necessarily occur given the time it takes the helicopter crew to launch, find a suitable landing position, and provide care at the scene. In addition, the costs of maintaining an air ambulance (Gearhart et al. 1997) and safety concerns of using aeromedical ambulances (Baker et al. 2006) need to be considered. This is especially true when the distance from definitive care to the scene is short (Asaeda et al. 2001). Despite the controversial and limited data available on the use of air ambulances, many trauma systems continue to use air medical services under certain circumstances with the assumption that it is faster than land and should be employed to minimize the patient's time to definitive care.

Post-acute trauma care and associated long-term mortality outcomes as well as nonfatal trauma are of great importance when assessing the effectiveness of trauma systems. Yet similar to prehospital, information on rehabilitation is not as extensive as acute care outcomes. A number of studies have measured functional outcomes after discharge using scales such as the Medical Outcomes Study Short Form Health Survey (SF-36), Chronic Pain Grade Scale, Quality of Well-Being score, Pediatric Quality of Life Inventory, Sickness Impact Profile, and others (Edwards et al. 2007; Holbrook et al. 1999; Mackenzie et al. 2004; McCarthy et al. 2006; Rivara et al. 2008). These studies

typically describe the functioning levels of trauma patients in the months and years following injury. Trauma systems research includes a study by MacKenzie et al. (2008) that found modest treatment benefit for patients with lower limb injuries who were treated at trauma centers compared with non-trauma centers. Similarly, Cudnik et al. (2009) observed improved functional outcomes at Level I trauma centers compared with Level II trauma centers. Riggs et al. (2010) studied patients with joint replacement or hip fracture and reported that discharge to acute inpatient rehabilitation was associated with decreased risk of hospital readmission. Conversely, a recent Cochrane review found insufficient evidence to recommend practice changes in the area of rehabilitation intervention for improving physical and psychosocial functioning after hip fracture in older people (Crotty et al. 2010). While most trauma providers are highly supportive of the benefit to trauma patients in receiving post-acute care in a specialized trauma rehabilitation center, evidence is limited. Post-acute rehabilitation is not an entitlement and patients may receive post-acute care in rehabilitation centers, skilled nursing facilities, at home, or not at all. Additional research is warranted in this area as scant evidence has been published comparing the outcomes of trauma patients by where they receive their post-acute care.

Systematic reviews or meta-analyses have also been performed to assess the body of evidence on trauma systems effectiveness. Champion et al. (1990) analyzed the MTOS, a retrospective study of injury severity and outcome and found an overall mortality rate of 9.0%. To assess the effectiveness of trauma systems based on registry comparisons, Jurkovich and Mock (1999) reviewed eight studies assessing outcomes compared with the MTOS norms and observed a 15–20% reduction in the risk of death. In a separate study, Mullins and Mann (1999) reviewed published evidence of trauma system effectiveness based on population-based studies and observed a 15–20% improved survival rate among seriously injured trauma patients following the implementation of a trauma system. MacKenzie (1999) reviewed ten panel studies of trauma systems effectiveness and reported weak evidence of trauma system effectiveness. Brown et al. (2009) conducted a systematic review of paramedic determinations of medical necessity and found little support for allowing paramedics to independently determine transport needs, thus reinforcing medical control and protocol needs for EMS systems. These systematic reviews offer insight into the validity of retrospective studies and offer some additional evidence in the absence of more rigorous prospective studies.

Although resource intensive, prospective studies, including randomized control trials, are increasingly contributing to the literature in the area of trauma system effectiveness. A prime example is the National Study of the Costs and Outcomes of Trauma (NSCOT) from which MacKenzie et al. (2006) found a significantly lower in-hospital mortality rate in trauma centers compared with non-trauma centers. In this analysis, propensity score weighting was used to control for observable differences in patients treated at trauma centers versus non-trauma centers. After controlling for severity and other potential confounders, in-hospital mortality at trauma centers was 7.6%, compared with 9.5% in non-trauma centers.

The NSCOT has also been used to study trauma outcomes and differences by hospital characteristics (trauma centers vs. non-trauma center) in areas such as:

- Value of trauma center care – MacKenzie et al. (2010) determined the value of trauma center care. The investigators found the added cost of treatment at a trauma center, compared with a non-trauma center, to be \$36,319 per life-year gained. The authors conclude that the regionalization of trauma care is both effective and cost-effective.
- Complications – Ang et al. (2009) compared complication rates at trauma centers to non-trauma centers and found a slightly greater rate of complications at trauma centers after adjusting for patient case mix. A possible explanation offered by the authors is that more aggressive treatment at trauma centers may be responsible for the difference, although the authors suggest that additional research is needed to understand the causes of complications and the observed differences between trauma centers and non-trauma centers.

- Intensity of care – Thompson et al. (2008) examined the relationship between age and intensity of care and mortality after traumatic brain injury in trauma centers and non-trauma centers. The authors report an inverse relationship between treatment intensity and age, with older patients receiving lower intensity of care and subsequently higher mortality in-hospital.
- Mortality risk for trauma – Thompson et al. (2010) developed and validated a comorbidity index to predict mortality risk for trauma patients. The authors identified six comorbidity factors (myocardial infarction, cerebrovascular disease, cardiac arrhythmias, severe liver disease, dementia, and depression) that were independently associated with mortality, and these were used as the basis for the index. This index is suggested to offer a simpler approach to controlling for case mix than the Charlson Comorbidity Index and others currently in use.
- Withdrawal of life-sustaining therapy – Cooper et al. (2009) compared withdrawal of life-sustaining therapy for injured patients treated at trauma centers and non-trauma centers. More than 60% of in-hospital trauma deaths occurred after a withdrawal of care order, with this more likely to occur in trauma centers.

Currently, studies of trauma system effectiveness concentrate on comparison of Level I trauma centers (often urban) to community non-trauma centers. Few studies have compared lower level trauma services in community settings to similar hospitals without trauma designation. Similarly, few studies have evaluated the effects of trauma systems on the care of seriously injured children. Injury death rates in rural settings are often double the rate in urban settings, yet few studies have examined how trauma systems benefit the rural trauma patient through the participation of rural EMS and hospital providers. A potential barrier to studying these areas continues to be the lack of comprehensive, linked data sets to follow patients from injury to definitive care.

Challenges and Future Direction

While significant advances have been made in the past four decades in the development and refinement of trauma systems, there is still much to learn. The majority of research occurs in an urban setting and there is a dearth of knowledge of trauma system effectiveness in rural areas. There is still a need to understand how the trauma system works to save the lives of special populations such as young children and older adults. There is no doubt that trauma systems are effective in improving survival of trauma patients, especially those that are severely injured, but what happens after a patient leaves an acute care facility? Now, in the era of accountability and performance measures, we need evidence to support trauma system effectiveness not only in reducing deaths but also in reducing disability and the severity of the injury outcome. While trauma systems were first developed with the idea of including rehabilitation, information on what happens to trauma patients who are discharged from an acute care facility is scant. Most trauma registries do not contain post-acute care information regarding rehabilitation. Further research is needed on how best to integrate post-acute outcomes data from rehabilitation hospitals into statewide trauma registries to allow more complete evaluations of trauma systems. This will continue to be a challenge in trauma systems research due to the different components and associated data systems that must be used collectively to show effectiveness. Linking data from the prehospital to rehabilitation is needed to understand how the different components of the trauma system work together.

As trauma systems continue to evolve, so will the research questions associated with evaluating outcomes in trauma care. Most of the research on trauma systems effectiveness is derived from trauma registries that are not population-based and that vary from region to region in their inclusion criteria. The development of the NTDB and the accompanying National Trauma Data Standard is promising and may improve our ability to use population-based trauma data. The role of prehospital

interventions in improving trauma outcomes is still unknown. EMS data collection is evolving although some agencies still use paper records. The development of standardized data elements through the NEMSIS program, as well as standardization of data submission for EMS agencies that bill for their services, are encouraging.

In addition, implementation research is needed to understand how effective each part of the trauma system is, how effective the parts are when combined into a complete system, where the opportunities for future improvement in survival are, when trauma systems have reached their point of maximum effectiveness, and how many additional lives would be saved by fine tuning the structure and function of trauma systems. As the evaluation and fine tuning of trauma systems continue, many more questions will arise. Few studies have examined the extent to which trauma system policy is enforced. How does that affect patient outcomes? Is there political will to de-designate a trauma center that underperforms or fails to comply with trauma system rules and regulations? Research is still needed to develop useful and valid outcome measures that permit accurate assessment of the effects of trauma systems that can be used to assist health care providers, policy makers, and community leaders in understanding the effect of trauma systems on the health of the community and to identify areas that require further work.

Lastly, even in a mature trauma system, preventable deaths will occur. As a result, trauma systems research will benefit the community by engaging in research focusing on the prevention of injuries as well as improved treatment of injured patients.

References

- American College of Surgeons (ACS). (2011). Retrieved from <http://www.facs.org/trauma/index.html>. Accessed 18 Jan 2011.
- American College of Surgeons Committee on Trauma (ACS-COT). (2006). *Resources for optimal care of injured patient*. Chicago, IL: American College of Surgeons.
- Ang, D. N., Rivara, P., Nathens, A., Jurkovich, G. J., Maier, R. V., Wang, J., et al. (2009). Complication rates among trauma centers. *Journal of the American College of Surgeons*, *295*, 595–602.
- Arthur, M., Newgard, C. D., Mullins, R. J., Diggs, B. S., Stone, J. V., Adams, A. L., et al. (2009). A population-based survival assessment of categorizing Level III and IV rural hospitals as trauma centers. *The Journal of Rural Health*, *25*, 182–188.
- Asaeda, G., Cherson, A., Giordano, L., & Kusick, M. (2001). Utilization of air medical transport in a large urban environment: a retrospective analysis. *Prehospital Emergency Care*, *5*, 36–39.
- Baker, S. P., Grabowski, J. G., Dodd, R. S., Shanahan, D. F., Lamb, M. W., & Li, G. H. (2006). EMS helicopter crashes: what influences fatal outcome? *Annals of Emergency Medicine*, *47*(4), 351–356.
- Bazzoli, G. J., Madura, K. J., Cooper, G., MacKenzie, E. J., & Maier, R. V. (1995). Progress in the development of trauma systems in the United States: results of a national survey. *JAMA*, *273*(5), 395–401.
- Bennett, K. M., Vaslef, S., Pappas, T. N., & Scarborough, J. E. (2010). The volume-outcomes relationship for United States Level I trauma centers. *The Journal of Surgical Research*, *167*(1), 19–23.
- Biewener, A., Aschenbrenner, U., Rammelt, S., Grass, R., & Zwipp, H. (2004). Impact of helicopter transport and hospital level on mortality of polytrauma patients. *The Journal of Trauma*, *56*(1), 94–98.
- Bowman, S. M., Zimmerman, F. J., Christakis, D. A., Sharar, S. R., & Martin, D. P. (2005). Hospital characteristics associated with the management of pediatric splenic injuries. *JAMA*, *294*, 2611–2617.
- Bowman, S. M., Zimmerman, F. J., Sharar, S. R., Baker, M. W., & Martin, D. P. (2008). Rural trauma: is trauma designation associated with better hospital outcomes? *The Journal of Rural Health*, *24*, 263–268.
- Branas, C. C., MacKenzie, E. J., Williams, J. C., Schwab, C. W., Teter, H. M., Flanigan, M. C., et al. (2005). Access to trauma centers in the United States. *JAMA*, *293*(21), 2626–2633.
- Brown, L. H., Hubble, M. W., Cone, D. C., Millin, M. G., Schwartz, B., Patterson, P. D. et al. (2009). Paramedic determinations of medical necessity: a meta-analysis. *Prehospital Emergency Care*, *13*, 516–527.
- Brown, J. B., Stassen, N. A., Bankey, P. E., Sangosanya, A. T., Cheng, J. D., & Gestring, M. L. (2010). Helicopters and the civilian trauma system: national utilization patterns demonstrate improved outcomes after traumatic injury. *The Journal of Trauma*, *69*(5), 1030–1034.

- Cales, R. H., & Trunkey, D. D. (1985). Preventable trauma deaths: a review of trauma care systems development. *JAMA*, 254(8), 1059–1063.
- Centers for Disease Control and Prevention (CDC). (2007). Web-based Injury Statistics Query and Reporting System (WISQARS). Retrieved from <http://www.cdc.gov/ncipc/wisquars>. Accessed 18 Jan 2011.
- Champion, H. R., Copes, W. S., Sacco, W. J., Lawnick, M. M., Keast, S. L., Bain, L. W Jr., et al. (1990). The Major Trauma Outcome Study: establishing national norms for trauma care. *The Journal of Trauma*, 30(11), 1356–1365.
- Chang, D. C., Bass, R. R., Cornwell, E. E., & MacKenzie, E. J. (2008). Undertriage of elderly trauma patients to state-designated trauma centers. *Archives of Surgery*, 143(8), 776–781.
- Chappell, V. L., Mileski, W. J., Wolf, S. E., & Gore, D. C. (2002). Impact of discontinuing a hospital-based air ambulance service on trauma patient outcomes. *The Journal of Trauma*, 52(3), 486–491.
- Claridge, J. A., Leukhardt, W. H., Golob, J. F., McCoy, A. M., & Malangoni, M. A. (2010). Moving beyond traditional measurement of mortality after injury: evaluation of risks for late death. *Journal of the American College of Surgeons*, 210(5), 788–796.
- Cooper, Z., Rivara, F. P., Wang, J., MacKenzie, E. J., & Jurkovich, G. J. (2009). Withdrawal of life-sustaining therapy in injured patients: variations between trauma centers and non-trauma centers. *The Journal of Trauma*, 66(5), 1327–1335.
- Crotty, M., Unroe, K., Cameron, I. D., Miller, M., Ramirez, G., & Couzner, L. (2010). Rehabilitation interventions for improving physical and psychosocial functioning after hip fracture in older people. *Cochrane Database of Systematic Reviews*, (1).
- Cudnik, M. T., Newgard, C. D., Sayre, M. R., & Steinberg, S. M. (2009). Level I versus level II trauma centers: an outcomes-based assessment. *The Journal of Trauma*, 66(5), 1321–1326.
- Davis, D. P., Peay, J., Serrano, J. A., Buono, C., Vilke, G. M., Sise, M. J., et al. (2005). The impact of aeromedical response to patients with moderate to severe traumatic brain injury. *Annals of Emergency Medicine*, 46(2), 115–122.
- Durham, R., Pracht, E., Orban, B., Lottenburg, L., Tepas, J., & Flint, L. (2006). Evaluation of a mature trauma system. *Annals of Surgery*, 243(6), 775–783.
- Edwards, R. R., Magyar-Russell, G., Thombs, B., Smith, M. T., Holavanahalli, R. K., Patterson, D. R., et al. (2007). Acute pain at discharge from hospitalization is a prospective predictor of long-term suicidal ideation after burn injury. *Archives of Physical Medicine and Rehabilitation*, 88(12 Suppl 2), S36–42.
- Garshick, E., Kelley, A., Cohen, S. A., Garrison, A., Tun, C. G., Gagnon, D. et al. (2005). A prospective assessment of mortality in chronic spinal cord injury. *Spinal Cord*, 43(7), 408–416.
- Garwe, T., Cowan, L. D., Neas, B. R., Sacra, J. C., & Albrecht, R. M. (2010). Directness of transport of major trauma patients to a level I trauma center: a propensity-adjusted survival analysis of the impact on short-term mortality. *The Journal of Trauma*, 70(5), 1118–1127.
- Gearhart, P. A., Wuerz, R., & Localio, A. R. (1997). Cost-effectiveness analysis of helicopter EMS for trauma patients. *Annals of Emergency Medicine*, 30(4), 500–506.
- Gonzalez, R. P., Cummings, G., Mulekar, M., & Rodning, C. B. (2006). Increased mortality in rural vehicular trauma: identifying contributing factors through data linkage. *The Journal of Trauma*, 61(2), 404–409.
- Gorra, A. S., Clark, D. E., Mullins, R. J., & Delorenzo, M. A. (2008). Regional variation in hospital mortality and 30-day mortality for injured Medicare patients. *World Journal of Surgery*, 32(6), 954–959.
- Graham, J. E., Chang, P. F., Berges, I. M., Granger, C. V., & Ottenbacher, K. J. (2008). Race/ethnicity and outcomes following inpatient rehabilitation for hip fracture. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 63, 860–866.
- Harti, R., Gerber, L. M., Iacono, L., Ni, Q., Lyons, K., & Ghajar, J. (2006). Direct transport within an organized state trauma system reduces mortality in patients with severe traumatic brain injury. *The Journal of Trauma*, 60(6), 1250–1256.
- Haut, E. R., Chang, D. C., Hayanga, A. J., Efron, D. T., Haider, A. H., & Cornwell, E. E., III. (2009). Surgeon- and system-based influences on trauma mortality. *Archives of Surgery*, 144(8), 759–764.
- Holbrook, T. L., Anderson, J. P., Sieber, W. J., Browner, D., & Hoyt, D. B. (1999). Outcome after major trauma: 12-month and 18-month follow-up results from the Trauma Recovery Project. *The Journal of Trauma*, 46, 765–771.
- Jurkovich, G. J., & Mock, C. (1999). Systematic review of trauma system effectiveness based on registry comparisons. *The Journal of Trauma*, 47, S46–S55.
- Kannikeswaran, N., Mahajan, P. V., Dunne, R. B., Compton, S., & Knazik, S. R. (2007). Epidemiology of pediatric transports and non-transports in an urban Emergency Medical Services system. *Prehospital Emergency Care*, 11(4), 403–407.
- Kortebein, P., Granger, C. V., & Sullivan, D. H. (2009). A comparative evaluation of inpatient rehabilitation for older adults with debility, hip fracture, and myopathy. *Archives of Physical Medicine and Rehabilitation*, 90(6), 934–938.

- Lansink, K. W. W., & Leenen, L. P. H. (2007). Do designated trauma systems improve outcome? *Current Opinion in Critical Care*, 13(6), 686–690.
- Macias, C. A., Rosengart, M. R., Puyana, J. C., Linde-Zwirble, W. T., Smith, W., Peitzman, A. B., et al. (2009). The effects of trauma center care, admission volume, and surgical volume on paralysis after traumatic spinal cord injury. *Annals of Surgery*, 249(1), 10–17.
- MacKenzie, E. J. (1999). Review of evidence regarding trauma system effectiveness resulting from panel studies. *The Journal of Trauma*, 47, S34–S41.
- MacKenzie, E. J., Hoyt, D. B., Sacra, J. C., Jurkovich, G. J., Carlini, A. R., Teitelbaum, S. D., et al. (2003). National inventory of hospital trauma centers. *JAMA*, 289(12), 1515–1522.
- MacKenzie, E. J., Bosse, M. J., Castillo, R. C., Smith, D. G., Webb, L. X., Kellam, J. F., et al. (2004). Functional outcomes following trauma-related lower-extremity amputation. *The Journal of Bone and Joint Surgery*, 86-A, 1636–1645.
- MacKenzie, E. J. Ph.D., Rivara, F. P. M.D., M.P.H., Jurkovich, G. J. M.D., Nathens, A. B. M.D., Ph.D., Frey, K. P. M.P.H., Egleston, B. L. M.P.P. et al. (2006). A national evaluation of the effect of trauma-center care on mortality. *The New England Journal of Medicine*, 354, 366–378.
- MacKenzie, E. J., Rivara, F. P., Jurkovich, G. J., Nathens, A. B., Egleston, B. L., Salkever, D. S., et al. (2008). The impact of trauma-center care on functional outcomes following major lower-limb trauma. *The Journal of Bone and Joint Surgery*, 90(1), 101–109.
- MacKenzie, E. J. Ph.D., Weir, S. Ph.D., Rivara, F. P. M.D., M.P.H., Jurkovich, G. J. M.D., Nathens, A. B. M.D., Ph.D., M.P.H., Wang, W. Ph.D. et al. (2010). The value of trauma center care. *The Journal of Trauma*, 69(1), 1–10.
- Mann, N. C., Mullins, R. J., MacKenzie, E. J., Jurkovich, G. J., & Mock, C. N. (1999). A systematic review of published evidence regarding trauma system effectiveness. *The Journal of Trauma*, 47, S25–S33.
- Mann, N. C., MacKenzie, E., Teitelbaum, S. D., Wright, D., & Anderson, C. (2005). Trauma system structure and viability in the current healthcare environment: a state by state-by-state assessment. *The Journal of Trauma*, 58(1), 136–147.
- Matshes, E. W., & Lew, E. O. (2010). Do resuscitation-related injuries kill infants and children? *The American Journal of Forensic Medicine and Pathology*, 31(2), 178–185.
- McCarthy, M. L., MacKenzie, E. J., Durbin, D. R., Aitken, M. E., Jaffe, K. M., Paidas, C. N., et al. (2006). Health-related quality of life during the first year after traumatic brain injury. *Archives of Pediatrics & Adolescent Medicine*, 160(3), 252–260.
- Moore, L., Hanley, J. A., Turgeon, A. F., & Lavoie, A. (2010). Evaluation of the long-term trend in mortality from injury in a mature inclusive trauma system. *World Journal of Surgery*. doi:10.1007/s00268-010-0588-z.
- Mullins, R. J. (1999). A historical perspective of trauma system development in the United States. *The Journal of Trauma*, 47, S8–S14.
- Mullins, R. J., & Mann, N. C. (1999). Population-based research assessing the effectiveness of trauma systems. *The Journal of Trauma*, 47, S59–S66.
- Mullins, R. J., Veum-Stone, J., Helfand, M., Zimmer-Gembeck, M., Hedges, J. R., Southard, P. A. et al. (1994). Outcome of hospitalized injured patients after institution of a trauma system in an urban area. *JAMA*, 271(24), 1919–1924.
- Mullins, R. J., Hedges, J. R., Rowland, O. J., Arthur, M., Mann, N. C., Price, D. O. et al. (2002). Survival of seriously injured patients first treated in rural hospitals. *The Journal of Trauma*, 52, 1019–1029.
- Nathens, A. B., Jurkovich, G. J., Rivara, F. P., & Maier, R. V. (2000a). Effectiveness of state trauma systems in reducing injury-related mortality: a national evaluation. *The Journal of Trauma*, 48(1), 25–31.
- Nathens, A. B., Jurkovich, G. J., Cummings, P., Rivara, F. P., & Maier, R. V. (2000b). The effect of organized systems of trauma care on motor vehicle crash mortality. *JAMA*, 283(15), 1990–1994.
- National Academy of Sciences. (1966). *Accidental death and disability: the neglected disease of modern society*. Washington, DC: National Research Council.
- National Academy of Sciences. (1999). *Reducing the burden of injury: advancing prevention and treatment*. Washington, DC: National Research Council.
- National Emergency Medical Services Information System (NEMSIS). (2011). Retrieved from <http://www.nemsis.org/>. Accessed 7 Mar 2011.
- Nirula, R., Maier, R., Moore, E., Sperry, J., & Gentilello, L. (2010). Scoop and run to the trauma center or stay and play at the local hospital: hospital transfer's effect on mortality. *The Journal of Trauma*, 69(3), 595–599.
- Owens, P. L., Zodet, M. W., Berdahl, T., Dougherty, D., McCormick, M. C., & Simpson, L. A. (2008). Annual report on health care for children and youth in the United States: focus on injury-related emergency department utilization and expenditures. *Ambulatory Pediatrics*, 8(4), 219–240.
- Papa, L., Langland-Orban, B., Kallenborn, C., Tepas, J. J. 3rd, Lottenberg, L., Celso, B. et al. (2006). Assessing effectiveness of a mature trauma system: association of trauma center presence with lower injury mortality rate. *The Journal of Trauma*, 61(2), 261–266.

- Pracht, E. E., Tepas, J. J., III, Langland-Orban, B., Simpson, L., Pieper, P., & Flint, L. M. (2008). Do pediatric patients with trauma in Florida have reduced mortality rates when treated in designated trauma centers? *Journal of Pediatric Surgery*, *43*(1), 212–221.
- Riggs, R. V., Roberts, P. S., Aronow, H., & Younan, T. (2010). Joint replacement and hip fracture readmission rates: impact of discharge destination. *PMR*, *2*(9), 806–810.
- Rivara, F. P., Mackenzie, E. J., Jurkovich, G. J., Nathens, A. B., Wang, J., & Scharfstein, D. O. (2008). Prevalence of pain in patients 1 year after major trauma. *Archives of Surgery*, *143*(3), 282–287.
- Shackford, S. R., Hollingsworth-Fridlund, P., Cooper, G. F., & Eastman, A. B. (1986). The effect of regionalization upon the quality of trauma as assessed by concurrent audit before and after institution of a trauma system: a preliminary report. *The Journal of Trauma*, *26*, 812–820.
- Shah, M. N., Cushman, J. T., Davis, C. O., Bazarian, J. J., Auinger, P., & Friedman, B. (2008). The epidemiology of emergency medical services use by children: an analysis of the National Hospital Ambulatory Medical Care Survey. *Prehospital Emergency Care*, *12*(3), 269–276.
- Thompson, H. J., Rivara, F. P., Jurkovich, G. J., Wang, J., Nathens, A. B., & MacKenzie, E. J. (2008). Evaluation of the effect of intensity of care on mortality after traumatic brain injury. *Critical Care Medicine*, *36*(1), 282–290.
- Thompson, H. J., Rivara, F. P., Nathens, A., Wang, J., Jurkovich, G. J., & MacKenzie, E. J. (2010). Development and validation of the mortality risk for trauma co-morbidity index. *Annals of Surgery*, *252*(2), 370–375.
- Trowbridge, M. J., Gurka, M. J., & O'Connor, R. E. (2009). Urban sprawl and delayed ambulance arrival in the U.S. *American Journal of Preventive Medicine*, *37*(5), 428–432.
- Utter, G. H., Maier, R. V., Rivara, F. P., Mock, C. N., Jurkovich, G. J., & Nathens, A. B. (2006). Inclusive trauma systems: do they improve triage or outcomes of the severely injured? *The Journal of Trauma*, *60*(3), 529–535.
- Ventura, T., Harrison-Felix, C., Carlson, N., Diguiseppi, C., Gabella, B., Brown, A. et al. (2010). Mortality after discharge from acute care hospitalization with traumatic brain injury: a population-based study. *Archives of Physical Medicine and Rehabilitation*, *91*(1), 20–29.
- West, J. G., Williams, M. J., Trunkey, D. D., & Wolferth, C. C. (1988). Trauma systems: current status-future challenges. *JAMA*, *259*(24), 3597–3600.

Chapter 33

Systems Approach to Patient Safety

Sneha Shah, Michelle Patch, and Julius Cuong Pham

Introduction

Patient safety research has become an international priority, fueled in large part by the 1999 release of the landmark report, “To Err is Human,” by the Institute of Medicine. This report galvanized support for patient safety improvements with the estimate of 44,000–98,000 deaths per year from medical errors in the US health-care system (Kohn et al. 1999). International medical error rates are equally high. The UK found that 11.7% of patients had an adverse event during their admission, with about half of those errors judged as preventable (Vincent et al. 2001). The Danish National Patient Register found that 9% of admissions were associated with an adverse event, 40.4% of which were potentially preventable (Schioler et al. 2001). In Australia, 16.6% of hospitalizations were associated with an adverse event (Wilson et al. 1996).

Medical errors occur in the outpatient setting as well. Approximately 5% of outpatients have an adverse drug reaction (ADR) (Hutchinson et al. 1986). In elderly patients, 95% of adverse drug events are potentially avoidable (Hanlon et al. 1997). In the Boston area, approximately 18% of chart reviews and patient surveys revealed an ADR. Interestingly, chart review found only 3% of these errors (Gandhi et al. 2000a). Based on electronic patient records, as many as 5.5% of patients who present for care experience an adverse drug event (Honigman et al. 2001).

Numerous high-profile cases have also alerted the media and public to medical mistakes such as wrong-site surgeries (Nundy et al. 2008), inadvertent administration of vincristine, a chemotherapeutic medication, into the spinal column instead of by intravenous infusion (Alcaraz et al. 2002; Schochet et al. 1968; Gilbar and Carrington 2004), and incorrect dosing with potassium chloride and heparin (Institute for Safe Medication Practices (ISMP)). These cases have galvanized many organizations to champion patient safety: Joint Commission International Center for Patient Safety

S. Shah, MD • M. Patch, MSN, RN
Department of Emergency Medicine, The Johns Hopkins Hospital,
600 N. Wolfe St./Marburg B-186, Baltimore, MD 21287, USA
e-mail: sneha.harshad.shah@gmail.com; mpatch1@jhmi.edu

J.C. Pham, MD, PhD (✉)
Department of Emergency Medicine, Johns Hopkins University School of Medicine,
5801 Smith Ave Suite 220, Baltimore, MD 21209, USA

Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine,
5801 Smith Ave Suite 220, Baltimore, MD 21209, USA
e-mail: jpham3@jhmi.edu

(USA, global), the Agency for Healthcare Research and Quality (USA), the National Patient Safety Agency (UK), the Canadian Patient Safety Institute (Canada), and the Australian Commission on Safety and Quality in Health Care (Australia) (Hofoss and Deilkas 2008). The US federal government stepped up with a pledge of \$50 million/year for patient safety research in 2001, and multiple organizations, including thousands of health-care professionals, have become stakeholders in patient safety (Leape and Berwick 2005).

The unique challenges of researching patient safety require unique methods. The purpose of this chapter is to provide a broad overview of methods for evaluating adverse events in the health-care field. First, we address some of the challenges to conducting research in the field of patient safety. Next, we provide a conceptual model for organizing patient safety research methods. This model involves methods to (1) identify hazards in health care, both prospectively (before the error has occurred) and retrospectively (after the error has occurred); (2) analyze and prioritize these identified hazards; and (3) reduce the likelihood of, mitigate the effects of, and prevent medical adverse events.

Challenges to Patient Safety Research

Although the importance of patient safety is evident, its research offers challenges not encountered in other fields. First, variability exists among clinicians on what constitutes a “medical error.” For example, it may not be entirely clear whether a known procedural complication (e.g., pneumothorax from a central venous catheter placement) represents an adverse event or a medical error. Many other such decisions and consequences within health care are filled with ambiguity (Pronovost et al. 2006a). This ambiguity often makes defining and measuring the problem difficult.

Second, it is very difficult to measure medical errors validly as rates (Pronovost et al. 2006a, 2009). Medical errors (the numerators) are relatively uncommon, denominators (i.e., population at risk, time periods of exposure such as patient day or device day) are generally unknown, and an active surveillance system is largely nonexistent. Health care generally relies on self-reporting of adverse events, a mechanism that is fraught with inherent biases.

Third, the most rigorous method of research, the randomized controlled trial, is difficult to conduct in patient safety research because the intervention is often assumed to help rather than harm, and it may not be ethical to withhold such treatment from patients (Pronovost et al. 2009; Brown et al. 2008). We are then left with quality improvement study designs, which have significant potential biases (Pronovost et al. 2009). Currently, very few researchers have the necessary educational background to perform this type of work (Pronovost et al. 2009).

Fourth, there has been and continues to be a certain amount of skepticism regarding the magnitude of medical errors. This skepticism has been driven by a combination of self-denial within the community, a fear of undermining patient confidence and public trust, and human resistance to change (Leape and Berwick 2005).

Finally, institutions generally lack the resources for patient safety research. Many of these interventions are complex. To be effectively implemented, interventions require funding, supplies/tools/training, and dedicated, expert staff. Because of tight budgets, much more funding is directed at discovering and developing new interventions, rather than ensuring that interventions are implemented safely and effectively.

Conceptual Framework

Some general definitions are warranted to assist in this discussion. The Institute of Medicine defines an adverse event as “injury caused by medical management rather than the underlying condition of the patient” and a medical error as “failure of a planned action to be completed as intended or the use

of a wrong plan to achieve an aim” (Kohn et al. 1999; Murff et al. 2003). A near-miss is defined as “an event or situation that did not produce patient injury, but only because of chance” (<http://www.psnet.ahrq.gov/glossary.aspx#N>).

Several frameworks exist for approaching patient safety and guiding related research projects. One framework published by Reason (Reason 2000) describes two categories of human error: the person (or active) perspective and the systems (or latent) perspective. The person perspective is centered on cognitive limitations behind error: forgetfulness, unethical behavior, inattention, and lack of knowledge or skill base. It places the individual as the focus for the mistake. Methods of reducing person-errors are aimed at reducing variability in behavior through measures such as protocols and disciplinary actions. The systems perspective focuses on the environment in which the person operates, such as clinical operations and workflow. Although the systems perspective has more promise for permanent change, it is much harder to identify and modify failures of an organization than of an individual.

The structure–process–outcome model by Donabedian (1980) is a frequently used and universally accepted tripartite concept for measuring quality (Battles and Lilford 2003). Structure refers to the health-care environment, including the physical facilities and equipment, available resources, characteristics of health-care staff, and attributes of patients. Processes are the stepwise actions that occur to achieve a result. They involve interpersonal relations (communication, therapeutic bond, rapport, and teamwork) among health-care providers and between providers and patients and encompass procedural and technical skills of the providers. Outcome is the final product or the effect of structure and process on the health of the patient. It includes patient outcomes, behavioral change, patient satisfaction, and health-related knowledge. The aim of safety research is to improve outcomes by decreasing/eliminating harm through failures of structure or process.

For the purpose of organizing this chapter, we use a practical model of evaluating patient safety (Fig. 33.1). Our model centers on a four-step process of patient safety improvement: (1) identify hazards, (2) analyze the identified hazards and prioritize the different risks, (3) implement interventions to mitigate these risks, and (4) evaluate the effectiveness of the risk reduction. These steps are taken at various levels of health care: the unit level, the hospital or trust level, and the industry/national level (Pham et al. 2010a). All three communities go through the same process, though each has a unique set of priorities and stakeholders that affect the process, and each level builds upon the other to create a multidisciplinary safety culture.

Identification of Hazards

Retrospective Methods

Retrospective review can be used to help identify error-prone practices and illuminate vulnerabilities in the process that led to an adverse event. The findings can then be used to develop recommendations for patient safety improvements. In this section, we will review five commonly used methods: medical/administrative record review, direct observation, patient or provider surveys/focus groups/patient interviews, malpractice claims review, and adverse event reporting systems (AERSs).

Medical Record and Administrative Record Review

The most common method used to identify the prevalence of adverse events is review of medical and/or administrative records. Medical records are readily available and may be accessed at any time. A trained clinician reviews the records to determine the nature and prevalence of medical

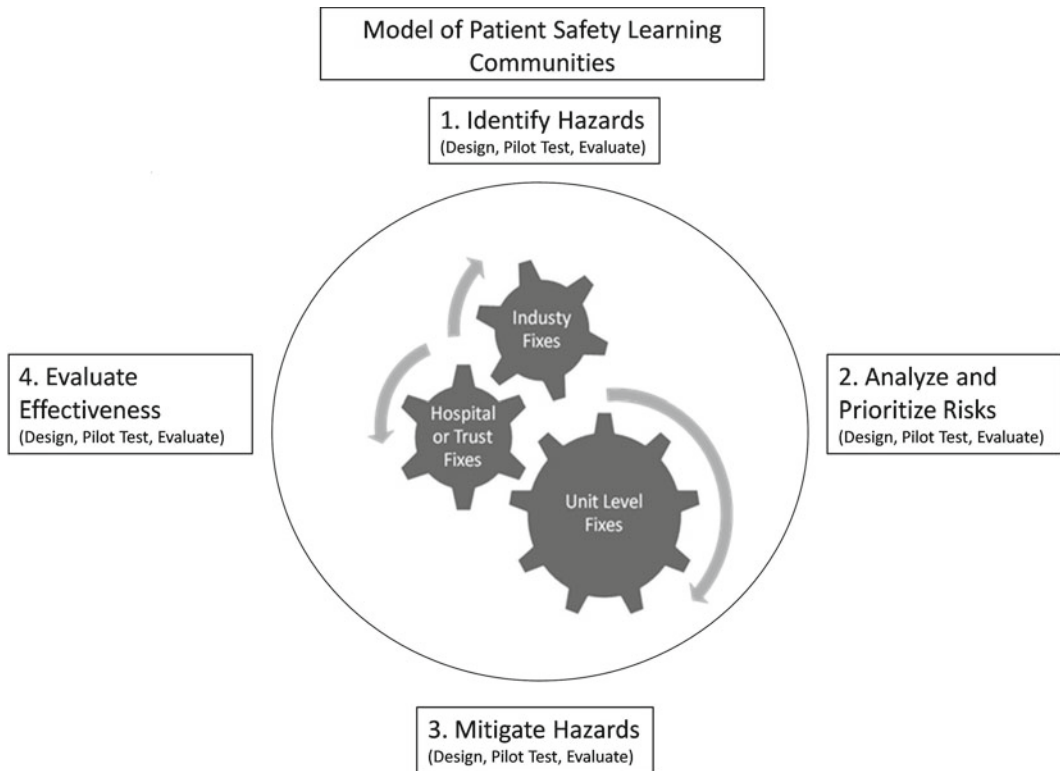


Fig. 33.1 Ideally, patient safety learning communities relate to each other in a gear-like fashion: as the identified hazards require stronger levels of intervention to achieve mitigation, the next learning community is engaged in action, eventually feeding back to the group that provided the initial thrust. Each group (unit, hospital, industry) follows the same four-step process but engages unique matrices of stakeholders to mitigate hazards that are within its locus of control (Pham et al. 2010a)

errors. Using this method, the Medical Insurance Feasibility Study (Mills 1978) identified that 4.65% of hospitalizations in California were associated with an adverse event. Similarly, the landmark Harvard Medical Practice Study (Brennan et al. 1991) identified a 3.7% adverse event rate among hospitalizations in New York; 27.6% of those cases were determined to be caused by negligence (Brennan et al. 1991). Of the 2.9% of hospitalizations associated with adverse events in Utah and Colorado, negligence was deemed a factor in 27.4% (Colorado) to 32.6% (Utah) of cases (Thomas and Petersen 2003).

One drawback of this methodology is that it is labor intensive. Because medical errors are not common, a tremendous number of records must be reviewed in the initial screening to identify an error (Murff et al. 2003). The time requirement to review a large number of records and the need for trained clinicians adds up to high resource utilization and cost (Murff et al. 2003; Shojania et al. 2002). Several strategies have been identified to decrease this burden. The first is to have a trained, nonphysician reviewer (e.g., a research nurse, pharmacist, or other health-care discipline) screen the records based on predetermined criteria. A physician or clinician conducts a final review to determine if a medical error indeed exists (Murff et al. 2003; Shojania et al. 2002). The Institute for Healthcare Improvement's Global Trigger Tool is an example of this, and has also shown promise in improving detection of adverse events (Classen et al. 2011). Adoption of electronic health records and advances in computer algorithms now allow automatic screening of charts for errors. Combining modalities that utilize various electronic filters and triggers to screen and identify charts that require

more specific manual review shows some promise in helping to reduce reviewer time and, therefore, expense (Murff et al. 2003).

Another drawback of record review is the potential for variability in inter-rater reliability (Murff et al. 2003; Shojania et al. 2002). What one reviewer considers an error may be considered a complication of the procedure/condition. Often, both views are right.

Finally, documentation limitations may exist. Clinicians may not document the medical error, either because they did not know it occurred or out of fear of medical liability. Documentation may not be integrated across individual health systems and generally is not integrated across nonaffiliated health-care entities. This lack of integration makes it difficult to connect errors with adverse outcomes which may occur weeks, months, or years later (Muething et al. 2010).

Direct Observation

Trained observers can be placed in clinical areas to detect adverse events and errors. The main advantage of direct observation over chart review is that errors are witnessed in real time, allowing for immediate feedback and correction. Furthermore, a second “set of eyes” allows objective evaluation of what is actually occurring, as opposed to what is being documented. For example, administration of the wrong medication may go undetected by the patient, nurse, physician, and medical record but be caught by an independent observer. Therefore, direct observation is better for capturing certain types of errors (such as medication administration errors) rather than other types (such as prescribing medication prescribing errors) (Gandhi et al. 2000b). What’s more, direct observation is better at detecting errors that do not result in patient injury (near misses) than other methods.

Direct observation has been used mostly in the study of medication errors. One study compared three methods of detecting medication error: incident report review, chart review, and direct observation. The investigators found that direct observation was more accurate and efficient (300 of 457 research pharmacist-confirmed errors detected) than chart review (17 of 457 errors detected) and incident report review (1 of 457 errors detected) (Flynn et al. 2002). Others have had similar findings and consider it the “best error detection method” (Allan and Barker 1990). Direct observation by human factors engineers and ethnographers in the intensive care unit (ICU) has identified similar patterns (Donchin et al. 1995; Andrews et al. 1997).

One barrier to direct observation is that the observers require training in qualitative observational methods, a specialized clinical background, and/or mentored hands-on experience within the clinical environment. Such training may be costly. Additionally, direct observation is not very efficient; many hours or days of observation may be required before an error or adverse event is detected.

An alternative to using dedicated on-site observers is videography. Videos may be reviewed at a convenient time/place. Video recording potentially costs less than observers, but hospitals would still incur the cost of purchasing and maintaining video equipment. Trained individuals would still be required to review and interpret the video in a meaningful way. Patient privacy laws may hinder this method from being widely utilized directly in the patient care environment.

Random safety auditing is a more focused means of direct observation. It involves frontline clinical staff monitoring designated high-risk, error-prone processes and procedures on a random basis rather than all the time. Advantages include lower cost (minimal additional training for participants since frontline clinical staff are observers), minimization of Hawthorne effect (staff change practice when they know they are being observed), engagement of clinical staff in safety outcomes, higher yield in error detection (especially those that are difficult to detect by other means), and adaptability to those high-risk processes and procedures (Ursprung and Gray 2010). Random safety audits have been extensively and effectively used by the National Center for Patient Safety in the neonatal ICU (Ursprung and Gray 2010; Ursprung et al. 2005).

Patient or Provider Surveys/Focus Groups/Patient Interviews

This technique refers to surveying or direct interviewing of health-care providers or patients by researchers to identify and/or gather information regarding adverse events. Physicians and patients are aware of errors that occur in their medical care, and these errors have resulted in substantial morbidity and mortality (Blendon et al. 2002). Structured patient telephone interviews often identify adverse events prevalent in the discharge period that are not identified by medical record review (Forster et al. 2003). Moreover, patient complaints can reveal safety defects in a way that is complementary to that of other reporting systems (Levtzion-Korach et al. 2010). Patient surveys and interviews can aid in detection of adverse events and can provide insight into existing barriers which may prevent patients from engaging in an open discussion of safety concerns with their health-care team (Schwappach 2008). Having this unique patient perspective on medical errors is beneficial, especially where medical records are less robust, such as in the outpatient setting. However, similar to direct observation, focus groups and patient interviews can be both time and labor intensive.

Malpractice Claims

The review of legal claims data alleging malpractice and/or negligence can be used to identify and/or analyze potential medical errors and preventable adverse events. Claims data may be useful in revealing medical errors and/or adverse events not previously identified through other mechanisms (Levtzion-Korach et al. 2010). Some examples include the medical liability component (Study III) of the Harvard Medical Practice Study (Localio et al. 1991) and a recent study involving claim files from the Netherland's largest medical liability insurer (van Noord et al. 2010). Claims data shared across institutions may reveal common areas of patient safety risk. CRICO-RMF and MCIC Vermont, Inc. are examples of medical malpractice insurer groups who have partnered with their insured institutions to focus patient safety efforts based, in part, on claims data. Yet the use of claims data has limitations: (1) access to claims data for research purposes may be limited until appropriate adjudication; (2) claims may be delayed by years, as adverse outcomes may not manifest immediately (Levtzion-Korach et al. 2010); (3) malpractice claims may not represent an actual error, it may just be a tragic event; and (4) patients who sustain injury related to medical error may never file a claim for compensation (Localio et al. 1991).

Adverse Event and “Near Miss” Reporting Systems

Health-care reporting systems may involve voluntary or mandatory reporting of adverse events and/or “near misses” by clinicians, depending on state or regulatory requirements. Specific examples include: MEDMARX system for capturing medication-related errors in the USA, Intensive Care Unit Safety Reporting System, National Reporting and Learning System used in the UK, and the University Health System Consortium Patient Safety Net. Under the US Patient Safety and Quality Improvement Act of 2005, patient safety organizations (PSOs) are being developed to support clinicians in voluntary, confidential reporting of information. This plan is intended to increase reporting rates, allow for improved identification and analysis of errors, and make health care safer (<http://www.pso.ahrq.gov/regulations/regulations.htm>).

Some challenges encountered with these reporting systems include potential bias associated with the types of events reported, such as hindsight bias; increased reporting in voluntary systems, which does not necessarily equate with increased frequency of errors/events – rather, it may reflect a more robust culture of safety (Thomas and Petersen 2003); difficult-to-determine denominators (Shojania 2008); and a focus on encouraging health-care providers to submit events rather than a focus on

reducing risk (Pronovost et al. 2009). Many events may be underreported because of workflow interruption, clinicians' concern about litigation, belief that reporting will not result in effective change, and lack of information that an adverse event has occurred (Murff et al. 2003). In addition, methods to analyze large groups of events are lacking (Pronovost et al. 2009). Categorization of events within various reporting systems has historically lacked uniformity. However, event reporting via PSOs and the adoption of evidence-based common definitions and reporting formats, known as "Common Formats" is expected to improve analysis of aggregate data (Clancy 2010).

Prospective Methods

Health care tends to be more reactive than proactive. Therefore, prospective approaches to patient safety research are newer than the retrospective methods discussed above. Prospective methods of medical injury research focus on decreasing the risk for potential errors and adverse events through prospective analysis and modification of high-risk practices. Two frequently utilized methods are simulation and failure mode and effects analysis (FMEA).

Simulation

Simulation builds on models used in various other fields, such as aviation (crew resource management), the military (Link Trainer, Combat Trauma Patient Simulation Program), and nuclear power. In medical simulation, patient encounter experiences are realistically recreated for use as a learning tool. These experiences can be simulated via standardized patients (actors portraying patients), specialized curriculum, and through the use of mannequins and enhanced simulation environments to reduce errors on actual patients. The recent advent of high-fidelity computerized mannequins has had a significant impact on medical simulation. These mannequins are so lifelike that they can be programmed to talk, breathe, deliver babies, and undergo surgery. They can respond to interventions with an extreme likeness to actual patients, enabling instructors to evaluate both cognitive and procedural skills. Simulation has been used in the fields of surgery, obstetrics/gynecology, and anesthesia. It can improve procedural skills such as adult and pediatric resuscitation (Schwid et al. 1999; Perkins 2007; Nadel et al. 2000), laparoscopic cholecystectomy (Seymour et al. 2002), and lumbar puncture (Lammers et al. 2005). Advantages to simulation are that it provides a safe practice environment for real-time feedback; can help improve competency for complicated, rare, life-threatening situations; and poses no harm to patients. However, simulation is resource intensive (it is a relatively expensive technology, time-consuming, and needs institutional and leadership support) and may not capture all of the elements of a real-life encounter.

Failure Mode and Effects Analysis

FMEA is a nonstatistical risk analysis technique that evaluates each potential failure in a system to determine both causes and effects of the failure and what actions need to be taken to repair the failure. It has been employed in manufacturing, computer software design, and aviation for decades and has recently spread to the health-care industry. There are two types of FMEA – design FMEA (which evaluates all of the individual components of a product to identify failures) and process FMEA (which evaluates the workflow to create a product); the latter is more applicable to health care. The general FMEA process involves five major steps: choosing a clinical process to study, creating a team to conduct the evaluation, organizing information about the process, conducting a

hazard analysis, and finally, implementing actions and outcome measures (Spath 2003). A health-care-specific FMEA has been developed by the Veterans Health Administration (VHA) (DeRosier et al. 2002). Advantages to FMEA are that (1) it is a prospective process (proactive) that prevents errors from happening rather than responding to an event after it has occurred (reactive), (2) it obviates hindsight bias, and (3) there is no blame or fear on the part of participants for an error that has already occurred. It is limited by participants needing basic training and skills in this mode of analysis and variability in the participant rating of risk. It is important to remember that the goal of an FMEA is “cost avoidance, not cost reduction” (Spath 2003). Although limited evidence is available regarding the validity, reliability, and effectiveness of the FMEA process, early results are promising. For example, FMEA has been used in a large pediatric hospital to decrease harm from tubing misconnections (Kimehi-Woods and Shultz 2006) and to improve the registration process for trauma patients (Day et al. 2007).

Adverse Event Analysis and Prioritization

Analysis Methods

After adverse events and hazards have been identified, they must be analyzed before methods of prevention can be explored. Through the analysis, causes and contributing factors are identified, and the risks are prioritized. This section explores research methods for adverse event analysis and risk prioritization.

Root Cause Analysis/Learning from Defects

Root cause analysis (RCA) is a retrospective method of evaluating adverse events based on finding the underlying “root cause” or contributing/causal factors responsible for an error. The basic principle is that correcting the “root cause” can prevent recurrence of the problem. The focus is on changes in procedures, systems, and processes rather than on blaming the individual. It has been widely used in the nuclear power and aviation fields to discover latent errors. RCA is becoming one of the main tools for error analysis in the health-care setting (Taitz et al. 2010; Pham et al. 2010b) and is required by various regulatory bodies, including the Joint Commission on Accreditation of Healthcare Organizations. Adverse events resulting in serious harm or death (sentinel events) are often the subject of RCAs. RCAs have been found to reduce rates of adverse drug events (Rex et al. 2000), reduce mortality from hip surgery (McGinn et al. 2005), and increase patient and graft survival after liver transplant (Perkins et al. 2005).

One challenge of RCA is obtaining sufficient evidence to make recommendations for improvement. The actual risk reduction of interventions is often unclear, and methods of effective evaluation are lacking (Pham et al. 2010b; Wreathall and Nemeth 2004). Second, when developing interventions, it is difficult to identify causal relationships between factors. Third, the determination of a root cause may be ambiguous because of the multitude of organizational levels that contribute to errors (Wreathall and Nemeth 2004). Fourth, the development and implementation of interventions often require an investment of money, time, and expertise that may not have institutional support (Pham et al. 2010b; Wreathall and Nemeth 2004). Fifth, many of the events investigated by RCA may be rare unusual occurrences. RCAs have even been described as scapegoat hunting (Hofoss and Deilkas 2008). However, RCAs remain one of the main methods of injury analysis in the health-care sector and recommendations for improvements are increasing the effectiveness of interventions (Pham et al. 2010b).

A variation on RCA is the Learning from Defects tool (Pronovost et al. 2006b). This is another structured retrospective approach to evaluating errors and can improve the efficiency of RCA by allowing an increased number of event reviews (Pronovost et al. 2009). It asks three basic questions: (1) What happened? This question is answered through a system/process-based perspective. (2) Why did it happen? This question serves to identify factors that increased and decreased risk to the patient. (3) What can we do to prevent it from happening again? This question serves to create a list of specific corrective actions that may help to decrease recurrence of the error and to develop a plan for implementation. The plan includes a project leader, follow-up dates, and a method of evaluating risk reduction. Evaluation can be either quantitative or qualitative based on the intervention design. Benefits include a simple methodology, real-time analysis, and unique evaluation of protective factors that reduce patient harm. The Learning from Defects tool is gaining popularity in many ICUs and is widely applied as part of the Comprehensive Unit-Based Safety Program (Pronovost et al. 2006c).

Case Series

A case series is an observational descriptive study that can be retrospective (most common) or prospective. A case report reflects a single individual, whereas a case series reflects a group of patients. Their clinical characteristics and outcomes are evaluated to develop an association between an effect and an environmental exposure. Case series are often employed when the disease process is uncommon, the disease process is thought to be linked to a specific exposure, and development of a randomized controlled trial may be difficult for ethical or resource-intensive reasons. For example, a case series of 227 RCAs from the VHA found that patient misidentification accounted for the majority of mislabeled specimens (Dunn and Moga 2010). Misidentification was often caused by selecting the wrong medical record when patients with similar names were on the same unit and “batching” of specimens and printed labels for multiple patients before submission.

Risk Prioritization Methods

At an institutional or national level, the challenge is not to identify hazards but rather to prioritize multiple competing risks. Unfortunately, in health care, resources are too limited to address every hazard.

Frequency of Occurrence and Proportions

Basic descriptive statistics can be used to help this risk prioritization. The frequency of event occurrence (total counts) refers to the summation of the number of errors or adverse events. It can be presented as a proportion, or ratio, of a subset [e.g., number of events (the numerator)] to a whole [e.g., total population, number of potential events, and the number of patients (the denominator)]. These values can be further analyzed by statistical methods such as means and percentages, depending on the nature of the data being collected. An example of this method comes from the UK’s National Patient Safety Agency, which analyzed more than 1,800 reports of serious incidents that potentially resulted in death. The agency determined that nearly 600 of these incidents were potentially preventable and used “counts” to determine the most frequently occurring errors as a means to direct future safety efforts (National Patient Safety Agency 2007; Pronovost et al. 2006c).

Table 33.1 Proportional reporting ratios

	Specific drug	All other drugs
Specific reaction	<i>a</i>	<i>b</i>
All other reactions	<i>c</i>	<i>d</i>

Proportional reporting ratios (PRR) = $[a/(a+c)]/[b/(b+d)]$. Denominators are unknown or uncertain – proportional mortality ratios. Reprinted with permission from Evans et al. (2001)

Proportional Reporting Ratios

Some of the largest spontaneous ADRs are maintained by the Food and Drug Administration (FDA), AERS, the Centers for Disease Control (CDC)/FDA Vaccine Adverse Events (VAERS), and the World Health Organization (WHO) international pharmacovigilance program. When suspected ADRs are reported, the event itself may or may not be related to the drug. False positives are common in these large databases (Edwards and Biriell 1994). Only when a true “signal” is detected can action be taken. Quantitative signal detection methods often use measures of disproportionality to quantify unexpectedness – which means that they compare the observed frequency of reports for a specific drug’s adverse event to the frequency of that adverse event for all drugs in the database to see if it is “higher than expected” (Bate and Evans 2009). Disproportionality analysis has four main methods: proportional reporting ratios (PRRs), Multi-Item Gamma Poisson Shrinker (MGPS), reporting odds ratio (ROR), and the Bayesian confidence propagation neural network (BCPNN). PRR and MGPS, the two best-described methods, are discussed below.

PRR uses a quantitative statistical method to evaluate the signal-to-noise ratio (Table 33.1) (Evans 2000). The null value for PRR is 1, and the higher the PRR the greater the signal strength (Evans et al. 2001). PRR has been successfully used to analyze the WHO Collaborating Program for International Drug Monitoring database, which contains two million reports (Bate et al. 1998). The PRR is simple to calculate and useful in situations where it may be difficult to obtain a denominator (Evans et al. 2001). Because PRRs are based on spontaneous reports, they also carry with them the limitations of spontaneous reporting: the data are observational; they include voluntary and passive surveillance, which leads to underreporting; there is no certainty that a reported reaction is causal; treatment is not randomized; data sets may be incomplete; changes in reporting patterns occur over time; classification methods impact analysis; and duplicate reports may exist (Bate and Evans 2009).

MGPS is another common quantitative method of analyzing data from spontaneous ADR databases. It uses a statistical model to compute signal scores (adjusted ratios of observed-to-expected drug event reports) while adjusting for confounders such as demographics. The MGPS algorithm calculates an empirical Bayes geometric mean (EBGM). It (MGPS) creates replicable, consistent, redundant signals and decreases random patterns (Szarfman et al. 2002). It can also create signal scores for higher-order (triplet or quadruplet) combinations of adverse events and drugs and identify possible synergistic interactions and syndromes (Szarfman et al. 2002). An EBGM of ten corresponds to a drug-event combination that is found ten times more frequently than if no relationship existed between the drug and the event (Almenoff et al. 2006). The MGPS tends to be more specific but less sensitive than the PRR (Almenoff et al. 2006). Several studies have used this method to investigate potential drug interactions (Almenoff et al. 2003).

Harm Susceptibility Model

Developed by Pham et al. (Pham et al. 2010c) in 2010, the harm susceptibility model is a quantitative method of analyzing large patient safety reporting systems. It is a statistical model that provides information about potential variation in organizational resiliency and harm within a sector (Pham

et al. 2010c). A harm susceptibility ratio (HSR) is calculated for the organizational level of interest (unit, hospital, medical specialty) and compared with similar organizational levels either inside a single institution or among institutions. The HSR is determined by the odds of harm reported (proportion of harmful events vs. nonharmful events) compared with the mean odds of harm for similar organizational levels (Pham et al. 2010c). This HSR provides a quick summary of risk within different departments and can help prioritize limited resources to maximize patient safety. This model has been applied to 20 trusts within the National Reporting and Learning Systems (Pham et al. 2010c). Limitations of this model are primarily related to reporting biases (variations in types of events reported, safety culture, hospital characteristics, etc.); in addition, all errors are treated with equal weight.

Injury Prevention and Mitigation

In section “Identification of Hazards,” we reviewed methods to identify safety risks, how often they occur, and what specific risk factors for error may exist. In this section, we will discuss methods used to study interventions intended to mitigate or prevent injury (e.g., performing “time-outs,” employing a unit-based pharmacist, providing teamwork training). In traditional scientific research, a randomized controlled trial is considered the “gold standard.” However, in patient safety research, ethical considerations preclude the use of this approach. Two commonly used methods of patient injury prevention and mitigation are quality improvement (QI) studies and quasi-randomized studies.

The QI study is by far the most commonly used method to test patient safety interventions (Fan et al. 2010). The design is a single-arm crossover study, often referred to as a before–after or pre–post study. QI studies can be prospective or retrospective. Key elements include identifying quality indicators, measuring baseline performance, designing interventions to improve the quality measure, assessing the impact on quality measures, and finally, sharing the results with others (Fan et al. 2010).

Several major interventions have been developed and tested with QI. Efforts to reduce wrong-site surgery (wrong place, wrong patient, wrong procedure) have used this design. The intervention is the Joint Commission/Veteran Health Affairs’s three-step process: (1) perioperative verification (confirming the patient’s name, identifying information, consent, and relevant diagnostic studies in the operating theater), (2) site marking (done prior to surgical prep by one of the surgeons involved in the procedure so that the patient can participate in confirming the planned procedure), and (3) time-out (before initiating the procedure, the patient’s name, identifying information, and operative plans are communicated to anesthesia and nursing teams in the operating theater) (Veterans Affairs National Center for Patient Safety 2004). Furthermore, use of a preoperative and postoperative debriefing checklist is associated with increased adherence to preoperative antibiotic orders and deep venous thrombosis prophylaxis (Paull et al. 2010). Teamwork training (VHA medical team training) that focuses on improving communication and teamwork has been shown to be associated with a lower surgical mortality (Neily et al. 2010). Similarly, use of a surgical safety checklist is associated with reductions in the rate of death and inpatient complications (Haynes et al. 2009).

Interventions to reduce central line-associated bloodstream infections (BSIs) have also employed a quality improvement design. BSIs account for approximately 82,000 infections and 28,000 related deaths each year (Klevens et al. 2007). Implementation of a multifaceted intervention and local culture change can reduce BSIs significantly (11.1 vs. 0 infections per 1,000 central line days) (Berenholtz et al. 2004). The multifaceted intervention includes: staff education, catheter supply cart, daily rounding about catheters that can be removed, procedural checklist, and empowering nurses to stop a procedure if violations occur (Berenholtz et al. 2004). When the BSI prevention program was replicated across an entire state (108 ICUs in Michigan), the results were equally

effective (66% incidence rate ratio) and sustained (up to 36 months) (Pronovost et al. 2006c, 2010). This program has been expanded to all states in the USA through a grant by the Agency for Healthcare Research and Quality and is gaining international attention (Sawyer et al. 2010).

In patient safety research, a quasi-randomized study design often involves randomly assigning groups of patients, rather than individual patients, to different interventions or forms of care. These groupings can be hospital units or entire hospitals. Grouping is used to overcome the challenge of individual subject randomization when the intervention involves system-wide changes in care as opposed to a medication or procedure. For example, when the intervention is a system of care that identifies decompensating patients, activates an emergency response team (medical emergency team, MET), and intervenes on the patient's behalf, it is difficult to randomize individual patients. This safety intervention was implemented in 23 hospitals in Australia to evaluate its effect on cardiac arrest, unplanned ICU admissions, and unexpected deaths (Hillman et al. 2005). All hospitals ultimately received the intervention (to overcome ethical concerns of denying treatment to patients), but the timing of implementation was randomized. Although the study revealed no differences in outcomes, enthusiasm and research into the MET concept continues.

Summary/Conclusion

Research on human subjects entails unique challenges, particularly when risks to patient safety are under scrutiny. Therefore, many factors must be considered in study design. This chapter serves to provide a broad overview of commonly used frameworks and methods for injury research related to patient safety. Beyond these methods, it is imperative that safety interventions be evaluated with appropriate data and that those interventions are customized specifically to the problem (Pronovost et al. 2009). As best summarized by Pronovost et al. (2009), future directions for patient safety will require funding prioritization from sources such as federal government and insurers, and should include:

1. Developing valid measures to evaluate patient safety progress.
2. Developing methods to reliably translate evidence into practice.
3. Studying the link between culture, behaviors, and patient outcomes.
4. Evaluating teamwork and leadership behaviors.
5. Using simulation to evaluate teamwork and technical work, train staff to translate evidence into practice, and identify and mitigate hazards.
6. Coordinating national-level efforts to investigate and implement industry-wide changes (e.g., commercial aviation safety teams).
7. Exploring ways to efficiently and effectively use patient safety resources at the unit, department, hospital, and health system levels.
8. Advancing the science of how to measure and reduce diagnostic errors in health care.
9. Developing patient safety measures that provide a more comprehensive view of the safety and quality in a product line (e.g., cardiac surgery).

References

- Alcaraz, A., Rey, C., Concha, A., et al. (2002). Intrathecal vincristine: Fatal myeloencephalopathy despite cerebrospinal fluid perfusion. *Journal of Toxicology. Clinical Toxicology*, 40, 557–561.
- Allan, E. L., & Barker, K. N. (1990). Fundamentals of medication error research. *American Journal of Hospital Pharmacy*, 47, 555–571.

- Almenoff, J. S., DuMouchel, W., Kindman, L. A., et al. (2003). Disproportionality analysis using empirical Bayes data mining: A tool for the evaluation of drug interactions in the post-marketing setting. *Pharmacoepidemiology and Drug Safety*, 12, 517–521.
- Almenoff, J. S., LaCroix, K. K., Yuen, N. A., et al. (2006). Comparative performance of two quantitative safety signaling methods: Implications for use in a pharmacovigilance department. *Drug Safety*, 29, 875–887.
- Andrews, L. B., Stocking, C., Krizek, T., et al. (1997). An alternative strategy for studying adverse events in medical care. *The Lancet*, 349, 309–313.
- Bate, A., & Evans, S. J. (2009). Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, 18, 427–436.
- Bate, A., Lindquist, M., Edwards, I. R., et al. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54, 315–321.
- Battles, J. B., & Lilford, R. J. (2003). Organizing patient safety research to identify risks and hazards. *Quality & Safety in Health Care*, 12(Suppl 2), ii2–ii7.
- Berenholtz, S. M., Pronovost, P. J., Lipsett, P. A., et al. (2004). Eliminating catheter-related bloodstream infections in the intensive care unit. *Critical Care Medicine*, 32, 2014–2020.
- Blendon, R. J., DesRoches, C. M., Brodie, M., et al. (2002). Views of practicing physicians and the public on medical errors. *The New England Journal of Medicine*, 347, 1933–1940.
- Brennan, T. A., Leape, L. L., Laird, N. M., et al. (1991). Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *The New England Journal of Medicine*, 324, 370–376.
- Brown, C., Hofer, T., Johal, A., et al. (2008). An epistemology of patient safety research: A framework for study design and interpretation. Part 2. Study design. *Quality & Safety in Health Care*, 17, 163–169.
- Clancy, C. M. (2010). Common formats allow uniform collection and reporting of patient safety data by patient safety organizations. *American Journal of Medical Quality*, 25, 73–75.
- Classen, D. C., Resar, R., Griffin, F., et al. (2011). ‘Global trigger tool’ shows that adverse events in hospitals may be ten times greater than previously measured. *Health Affairs (Millwood)*, 30, 581–589.
- Day, S., Dalto, J., Fox, J., et al. (2007). Utilization of failure mode effects analysis in trauma patient registration. *Quality Management in Health Care*, 16, 342–348.
- DeRosier, J., Stalhandske, E., Bagian, J. P., et al. (2002). Using health care Failure Mode and Effect Analysis: The VA National Center for Patient Safety’s prospective risk analysis system. *The Joint Commission Journal on Quality Improvement*, 28(248–67), 209.
- Donabedian, A. (1980). *Explorations in quality assessment and monitoring: The definition of quality and approaches to its assessment*. Ann Arbor, MI: Health Administration Press.
- Donchin, Y., Gopher, D., Olin, M., et al. (1995). A look into the nature and causes of human errors in the intensive care unit. *Critical Care Medicine*, 23, 294–300.
- Dunn, E. J., & Moga, P. J. (2010). Patient misidentification in laboratory medicine: A qualitative analysis of 227 root cause analysis reports in the Veterans Health Administration. *Archives of Pathology & Laboratory Medicine*, 134, 244–255.
- Edwards, I. R., & Biriell, C. (1994). Harmonisation in pharmacovigilance. *Drug Safety*, 10, 93–102.
- Evans, S. J. (2000). Pharmacovigilance: A science or fielding emergencies? *Statistics in Medicine*, 19, 3199–3209.
- Evans SJ, Waller PC, Davis S (2001) Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Safety*, 10, 483–486.
- Fan, E., Laupacis, A., Pronovost, P. J., et al. (2010). How to use an article about quality improvement. *Journal of the American Medical Association*, 304(20), 2279–87.
- Flynn, E. A., Barker, K. N., Pepper, G. A., et al. (2002). Comparison of methods for detecting medication errors in 36 hospitals and skilled-nursing facilities. *American Journal of Health-System Pharmacy*, 59, 436–446.
- Forster, A. J., Murff, H. J., Peterson, J. F., et al. (2003). The incidence and severity of adverse events affecting patients after discharge from the hospital. *Annals of Internal Medicine*, 138, 161–167.
- Gandhi, T. K., Burstin, H. R., Cook, E. F., et al. (2000a). Drug complications in outpatients. *Journal of General Internal Medicine*, 15, 149–154.
- Gandhi, T. K., Seger, D. L., & Bates, D. W. (2000b). Identifying drug safety issues: From research to practice. *International Journal for Quality in Health Care*, 12, 69–76.
- Gilbar, P. J., & Carrington, C. V. (2004). Preventing intrathecal administration of vincristine. *The Medical Journal of Australia*, 181, 464.
- Hanlon, J. T., Schmader, K. E., Koronkowski, M. J., et al. (1997). Adverse drug events in high risk older outpatients. *Journal of the American Geriatrics Society*, 45, 945–948.
- Haynes, A. B., Weiser, T. G., Berry, W. R., et al. (2009). A surgical safety checklist to reduce morbidity and mortality in a global population. *The New England Journal of Medicine*, 360, 491–499.
- Hillman, K., Chen, J., Cretikos, M., et al. (2005). Introduction of the medical emergency team (MET) system: A cluster-randomised controlled trial. *The Lancet*, 365, 2091–2097.

- Hofoss, D., & Deilkas, E. (2008). Roadmap for patient safety research: Approaches and roadforks. *Scandinavian Journal of Public Health, 36*, 812–817.
- Honigman, B., Lee, J., Rothschild, J., et al. (2001). Using computerized data to identify adverse drug events in outpatients. *Journal of the American Medical Informatics Association, 8*, 254–266.
- Hutchinson, T. A., Flegel, K. M., Kramer, M. S., et al. (1986). Frequency, severity and risk factors for adverse drug reactions in adult out-patients: A prospective study. *Journal of Chronic Diseases, 39*, 533–542.
- Institute for Safe Medication Practices (ISMP). *High alert medications*. Accessed 4, 2011, from <http://www.ismp.org/Tools/highalertmedications.pdf>.
- Kimehi-Woods, J., & Shultz, J. P. (2006). Using HFMEA to assess potential for patient harm from tubing misconnections. *Joint Commission Journal on Quality and Patient Safety, 32*, 373–381.
- Klevens, R. M., Edwards, J. R., Richards, C. L., Jr., et al. (2007). Estimating health care-associated infections and deaths in U.S. hospitals, 2002. *Public Health Reports, 122*, 160–166.
- Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (1999). *To err is human: Building a safer health system*. Washington, DC: National Academies Press.
- Lammers, R. L., Temple, K. J., Wagner, M. J., et al. (2005). Competence of new emergency medicine residents in the performance of lumbar punctures. *Academic Emergency Medicine, 12*, 622–628.
- Leape, L. L., & Berwick, D. M. (2005). Five years after To Err Is Human: What have we learned? *JAMA, 293*, 2384–2390.
- Levzion-Korach, O., Frankel, A., Alcalai, H., et al. (2010). Integrating incident data from five reporting systems to assess patient safety: Making sense of the elephant. *Joint Commission Journal on Quality and Patient Safety, 36*, 402–410.
- Localio, A. R., Lawthers, A. G., Brennan, T. A., et al. (1991). Relation between malpractice claims and adverse events due to negligence. Results of the Harvard Medical Practice Study III. *The New England Journal of Medicine, 325*, 245–251.
- McGinn, T., Conte, J. G., Jarrett, M. P., et al. (2005). Decreasing mortality for patients undergoing hip fracture repair surgery. *Joint Commission Journal on Quality and Patient Safety, 31*, 304–307.
- Mills, D. H. (1978). Medical insurance feasibility study. A technical summary. *Western Journal of Medicine, 128*, 360–365.
- Muehling, S. E., Conway, P. H., Kloppenborg, E., et al. (2010). Identifying causes of adverse events detected by an automated trigger tool through in-depth analysis. *Quality & Safety in Health Care, 19*, 435–439.
- Murff, H. J., Patel, V. L., Hripcsak, G., et al. (2003). Detecting adverse events for patient safety research: A review of current methodologies. *Journal of Biomedical Informatics, 36*, 131–143.
- Nadel, F. M., Lavelle, J. M., Fein, J. A., et al. (2000). Assessing pediatric senior residents' training in resuscitation: Fund of knowledge, technical skills, and perception of confidence. *Pediatric Emergency Care, 16*, 73–76.
- National Patient Safety Agency. (2007). Safer care for the acutely ill patient: Learning from serious incident 559. <http://www.nrls.npsa.nhs.uk/resources/?EntryId45=59828>
- Neily, J., Mills, P. D., Young-Xu, Y., et al. (2010). Association between implementation of a medical team training program and surgical mortality. *JAMA, 304*, 1693–1700.
- Nundy, S., Mukherjee, A., Sexton, J. B., et al. (2008). Impact of preoperative briefings on operating room delays: A preliminary report. *Archives of Surgery, 143*, 1068–1072.
- Paull, D. E., Mazzia, L. M., Wood, S. D., et al. (2010). Briefing guide study: Preoperative briefing and postoperative debriefing checklists in the Veterans Health Administration medical team training program. *American Journal of Surgery, 200*, 620–623.
- Perkins, G. D. (2007). Simulation in resuscitation training. *Resuscitation, 73*, 202–211.
- Perkins, J. D., Levy, A. E., Duncan, J. B., et al. (2005). Using root cause analysis to improve survival in a liver transplant program. *Journal of Surgical Research, 129*, 6–16.
- Pham, J. C., Gianci, S., Battles, J., et al. (2010a). Establishing a global learning community for incident-reporting systems. *Quality & Safety in Health Care, 19*, 446–451.
- Pham, J. C., Kim, G. R., Natterman, J. P., et al. (2010b). ReCASTing the RCA: An improved model for performing root cause analyses. *American Journal of Medical Quality, 25*, 186–191.
- Pham, J. C., Colantuoni, E., Dominici, F., et al. (2010c). The harm susceptibility model: A method to prioritise risks identified in patient safety reporting systems. *Quality & Safety in Health Care, 19*, 440–445.
- Pronovost, P. J., Miller, M. R., & Wachter, R. M. (2006a). Tracking progress in patient safety: An elusive target. *JAMA, 296*, 696–699.
- Pronovost, P. J., Holzmüller, C. G., Martinez, E., et al. (2006b). A practical tool to learn from defects in patient care. *Joint Commission Journal on Quality and Patient Safety, 32*, 102–108.
- Pronovost, P., Needham, D., Berenholtz, S., et al. (2006c). An intervention to decrease catheter-related bloodstream infections in the ICU. *The New England Journal of Medicine, 355*, 2725–2732.
- Pronovost, P. J., Goeschel, C. A., Marsteller, J. A., et al. (2009). Framework for patient safety research and improvement. *Circulation, 119*, 330–337.

- Pronovost, P. J., Goeschel, C. A., Colantuoni, E., et al. (2010). Sustaining reductions in catheter related bloodstream infections in Michigan intensive care units: Observational study. *BMJ*, *340*, c309.
- Reason, J. (2000). Human error: Models and management. *BMJ*, *320*, 768–770.
- Rex, J. H., Turnbull, J. E., Allen, S. J., et al. (2000). Systematic root cause analysis of adverse drug events in a tertiary referral hospital. *The Joint Commission Journal on Quality Improvement*, *26*, 563–575.
- Sawyer, M., Weeks, K., Goeschel, C. A., et al. (2010). Using evidence, rigorous measurement, and collaboration to eliminate central catheter-associated bloodstream infections. *Critical Care Medicine*, *38*, S292–S298.
- Schioler, T., Lipczak, H., Pedersen, B. L., et al. (2001). Incidence of adverse events in hospitals. A retrospective study of medical records. *Ugeskrift for Laeger*, *163*, 5370–5378.
- Schochet, S. S., Jr., Lampert, P. W., & Earle, K. M. (1968). Neuronal changes induced by intrathecal vincristine sulfate. *Journal of Neuropathology and Experimental Neurology*, *27*, 645–658.
- Schwappach, D. L. (2008). “Against the silence”: Development and first results of a patient survey to assess experiences of safety-related events in hospital. *BMC Health Services Research*, *8*, 59.
- Schwid, H. A., Rooke, G. A., Ross, B. K., et al. (1999). Use of a computerized advanced cardiac life support simulator improves retention of advanced cardiac life support guidelines better than a textbook review. *Critical Care Medicine*, *27*, 821–824.
- Seymour, N. E., Gallagher, A. G., Roman, S. A., et al. (2002). Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Annals of Surgery*, *236*, 458–463. discussion 463–464.
- Shojania, K. G. (2008). The frustrating case of incident-reporting systems. *Quality & Safety in Health Care*, *17*, 400–402.
- Shojania, K. G., Wald, H., & Gross, R. (2002). Understanding medical error and improving patient safety in the inpatient setting. *The Medical Clinics of North America*, *86*, 847–867.
- Spath, P. L. (2003). Using failure mode and effects analysis to improve patient safety. *AORN Journal*, *78*, 16–37. quiz 41–44.
- Szarfman, A., Machado, S. G., & O’Neill, R. T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA’s spontaneous reports database. *Drug Safety*, *25*, 381–392.
- Taitz, J., Genn, K., Brooks, V., et al. (2010). System-wide learning from root cause analysis: A report from the New South Wales Root Cause Analysis Review Committee. *Quality & Safety in Health Care*, *19*(6), e63.
- Thomas, E. J., & Petersen, L. A. (2003). Measuring errors and adverse events in health care. *Journal of General Internal Medicine*, *18*, 61–67.
- Ursprung, R., & Gray, J. (2010). Random safety auditing, root cause analysis, failure mode and effects analysis. *Clinics in Perinatology*, *37*, 141–165.
- Ursprung, R., Gray, J. E., Edwards, W. H., et al. (2005). Real time patient safety audits: Improving safety every day. *Quality & Safety in Health Care*, *14*, 284–289.
- van Noord, I., Eikens, M. P., Hamersma, A. M., et al. (2010). Application of root cause analysis on malpractice claim files related to diagnostic failures. *Quality & Safety in Health Care*, *19*(6), e21.
- Veterans Affairs National Center for Patient Safety. (2004). *Ensuring correct surgery and invasive procedures* (Vol. 4, Issue 4). http://www.patientsafety.gov/TIPS/Docs/TIPS_SeptOct04.pdf.
- Vincent, C., Neale, G., & Woloshynowych, M. (2001). Adverse events in British hospitals: Preliminary retrospective record review. *BMJ*, *322*, 517–519.
- Wilson, R. M., Runciman, W. B., Gibberd, R. W., et al. (1996). Quality in Australian Health Care Study. *The Medical Journal of Australia*, *164*, 754.
- Wreathall, J., & Nemeth, C. (2004). Assessing risk: The role of probabilistic risk assessment (PRA) in patient safety improvement. *Quality & Safety in Health Care*, *13*, 206–212.

Chapter 34

Intervention in Low-Income Countries

Samuel N. Forjuoh

Introduction

In this resource-constrained world, every effort must be made to avoid reinventing the wheel in injury prevention. This idea led some researchers to suggest that injury intervention efforts in developing or low- and middle-income countries (LMICs) should borrow from the successes of those implemented in industrialized or high-income countries, while paying attention to contextual issues (Forjuoh 1996; Zwi 1996). The World Health Organization recently echoed this suggestion, endorsing the tailoring of interventions found to be effective in high-income countries to LMICs, followed by rigorous evaluation (Peden et al. 2004). And just as LMICs could learn a lot from the injury interventions developed largely in high-income countries, learning from the failures is equally important.

Much of what we currently know about the science of injury prevention has been made possible by the excellent pioneering works of several individuals. These individuals include the great Americans Herbert Heinrich, Hugh de Haven, Edward Press, John Gordon, James Gibson, William Haddon Jr, and Susan Baker (Heinrich 1931; De Haven 1942; Press 1948; Gordon 1949; Gibson 1961; Haddon 1970, 1980; Baker and Haddon 1974; Baker et al. 1974). Together, these researchers paved the way for our current understanding of the causation of injury and how to prevent it. Possibly, other researchers from some other high-income countries also may have made substantial contributions, which are unfortunately not well publicized. The dedicated work of these researchers, along with the existence of adequate resources and high literacy rates, has resulted in the great success many high-income countries have had in injury prevention.

The same cannot be said of LMICs. The combination of poor resources, low literacy rates, and lack of personnel trained in injury prevention in LMICs dictates that they learn injury prevention from the experiences of high-income countries. Many LMICs are still hanging between the stages of epidemiologic polarization and protracted epidemiologic transition (Omran 1971; Frenk et al. 1989; Agyei-Mensah and de Graft Aikins 2010). This means that these relatively poor-resourced countries still have the onerous task of dealing with their long-standing problems of infections and malnutrition, alongside injury and other recently emerging health problems like obesity, diabetes, and HIV/AIDS.

This chapter discusses injury intervention in LMICs in the context of the theories, methods, and approaches to injury prevention presented in earlier chapters. The goal is to highlight effective and highly promising interventions that have already been transferred from high-income countries,

S.N. Forjuoh, MD, DrPH, MPH (✉)

Department of Family & Community Medicine, Scott & White Healthcare, Texas A&M Health Science Center College of Medicine, 1402 West Avenue H, Temple, TX 76504-5342, USA
e-mail: sforjuoh@swmail.sw.org

evaluated in the LMIC setting and found to be successful, and subsequently adopted. Another goal is to catalogue other promising interventions that require some consideration. First, injury interventions, most of which have been developed largely in high-income settings, are reviewed by their relative level of effectiveness. The challenges of transferring these interventions to LMICs are then discussed. Some key factors affecting the transfer of interventions and approaches to maximizing the effectiveness of transferred interventions are also discussed. Appropriate attention is given to the effect of ethics, politics, and other controversies on the transfer of interventions. Finally, selected interventions that have been transferred successfully to LMICs are described in greater detail, while some promising ones for serious consideration are catalogued.

Interventions by Level of Effectiveness

Although the basis of injury intervention is to focus on the exposures and specific risk factors causing injuries, oftentimes the most effective interventions may not necessarily be directed to mitigate the effects of the most obvious risk factors. For example, the most obvious factor causing a motorcycle crash injury in a rural setting might appear to be excessive speed. However, the most effective intervention, in terms of injury reduction, could be a motorcycle helmet law.

In addition, not all interventions developed to prevent injury are known to work effectively. One of the most important outcomes of a 1987 meeting of injury prevention experts from across the USA was the identification and documentation of four categories of injury interventions with respect to their effectiveness (The National Committee for Injury Prevention and Control 1989). These four categories of interventions, which have served as a useful guide in the science of injury prevention, are effective interventions, promising interventions, interventions with unknown effectiveness, and counterproductive interventions.

Effective Interventions

An effective intervention, also known as an intervention with proven effectiveness, is one whose implementation has been found to convincingly result in obvious injury reduction or some other major tangible and discernible positive effect. Examples of effective interventions are seat belts, air bags, safety helmets, sidewalks, roadway barriers, speed limits, speed bumps, smoke alarms, child-proof containers, laws regulating hot-water temperatures, residential sprinkler systems, and establishment of poison control centers. Seat belts, for instance, have been found to reduce fatalities in motor vehicle crashes by 50% and serious injury by 55% (Latimer and Lave 1987; Rivara et al. 2000). Smoke alarms have also been associated with over 70% reduction in deaths from fires and burns, particularly in conflagrations (Runyan et al. 1992). These are the interventions that should be transferred to LMICs and their effects monitored routinely. Already many of these effective interventions have been transferred and have been successfully implemented in some LMICs following their evaluation to measure their effectiveness in the local settings (Krug et al. 1994; Conrad and Bradshaw 1996; Petridou et al. 1999; Afukaar 2003).

Promising Interventions

A promising intervention is one that has been found to be associated with some injury reduction or some other mild discernible positive effect. Thus, the effectiveness of injury reduction of promising

interventions is not very convincing. Examples of promising interventions are licensure suspension laws, running light conspicuity measures, bicycle paths and lanes, one-way streets, edge lines, wrong-way signs, separation of cooking areas from living areas, promoting the use of safer lamps and stoves, developing standards and codes for fire-retardant garments, banning the manufacture and sale of fireworks, and locking away medicines and other toxic substances. While further research is needed to evaluate the effectiveness of these promising interventions, there is anecdotal evidence of the relative effectiveness of some of them. For example, a review of the effectiveness of the deterrence effect of licensure suspension laws in reducing the recidivism of drunk driving was not very convincing (McArthur and Kraus 1999). These interventions may also be candidates for transfer to LMICs. However, their adoption and implementation must be monitored more closely, and their outcomes evaluated more rigorously to assess their effectiveness in these new settings.

Interventions with Unknown Effectiveness

An intervention with unknown effectiveness is one that appears intuitively to have some discernible positive effect but has not been studied sufficiently. Examples of such interventions are bicycle safety programs, the designated driver concept, rumble strips, providing education about the risks of swimming, and reducing the storage of flammable substances in the home. It is doubtful whether these interventions are worth considering for transfer to LMICs at this time. However, as these interventions later transition into promising and, eventually, effective ones, they could then be ready for transfer and testing in LMICs. Therefore, researchers in LMICs need to test the effectiveness of these interventions in their local settings but should give priority to interventions known to be effective.

Counterproductive Interventions

A counterproductive intervention is one whose adoption and implementation has had little impact on injury reduction or that does not seem to work. Such interventions are simply ineffective! Examples of counterproductive interventions are painted crosswalks alone, driver education for young drivers, and applying butter to a burn injury. Painted crosswalks alone, particularly on two-lane roads, have not been found to be effective in reducing pedestrian crash rates (Zegeer et al. 2005). The butter-for-a-burn folk remedy presumably owes its origin and persistence, at least in part, to the soothing nature of a cool, greasy substance like butter and its immediate availability where minor burns often occur – in the kitchen. However, there are no studies exploring the healing or pain-reducing properties of butter. Placing butter or similar greasy ointments directly on a burn has rather been found to be counterproductive since it can seal in the heat. These interventions should not be considered at all for transfer to LMICs.

Actual Interventions vs. Strategies to Enhance Interventions

There is a need to differentiate between actual *interventions* that have been invented or developed to prevent an injury or ameliorate the effect of an injury and the *strategies* designed or developed to catalyze the actions of the actual interventions or promote the utility of the interventions. For example, in addressing the problem of motor vehicle injuries, the *seat belt* represents an actual intervention, while *mandatory seat belt use legislation and enforcement* represents one strategy to promote

the utility of the seat belt to protect an occupant from injury in case of a crash. Similarly, while the *smoke alarm* represents an actual intervention for fires and burns, *educating individuals to change the batteries of smoke alarms* represents a strategy to augment the utility of the smoke alarm. This distinction is important and relevant to the transfer of interventions across settings since local strategies may sometimes find a better fit with some effective transferred intervention vs. an accompanying transferred strategy. However, in this chapter, actual interventions and strategies will all be referred to as interventions.

Synergism of Interventions

Finally, it is also important to note that in order to effect successful and effective injury prevention in a particular situation, two or more interventions would often have to be implemented simultaneously or in succession. This is particularly so in LMICs where researchers are still testing for the best interventions or combinations of interventions to combat injuries. For example, visibility aids such as use of a flashing amber light, coupled with painted crosswalks, have been found to be a better intervention for prevention of pedestrian injuries than just painted crosswalks alone (Van Houten and Malenfant 1992; Kwan and Mapstone 2006; Karkee et al. 2010).

Challenges of Transferring Interventions

Transferring potentially effective or even highly promising interventions from high-income countries to LMICs faces several challenges. The major challenges include lack of adequate resources to import an intervention or conduct evaluation of the transferred intervention, inimical cultural beliefs regarding the causation of injury, competing health problems, existence of distinctive causes of injury in LMICs, low literacy rates, and peculiar political situations, which are often dominated by dictatorships and nondemocratic governments. These factors not only affect the adoption and implementation of transferred interventions, but they also can sometimes impede the diffusion of the transferred interventions in the local settings.

Limitation of Resources

In many LMICs, the resources to import and implement interventions or to assess the effectiveness of transferred interventions are often limited or even nonexistent. For example, there may be only a handful of researchers trained in the science of injury prevention in many of these LMICs. And oftentimes funds may be so limited that the budgetary allocation for safety promotion and injury prevention is rather deplorable. This situation does not augur well for sound injury intervention in LMICs.

It has been estimated that Pakistan spent a meager \$0.07 per capita or 0.015% of the country's gross domestic product (GDP) per capita on road safety in 1998. The corresponding figures for Uganda were \$0.09 per capita or 0.02% of GDP per capita (Bishai et al. 2003). The low percent-of-GDP per capita allocated to injury interventions in these LMICs simply does not make most interventions possible, even in situations where anecdotes or findings from empirical research reveal the need to intervene to prevent a particular type of injury. In certain circumstances, the issue becomes that of simply a misplaced priority.

Cultural Beliefs Regarding Fatalism and Injury

The problem of injury prevention is further compounded in many LMICs by people's fatalistic attitude toward injury causation – injuries are generally viewed as random or haphazard events, or even as acts of God. This fatalistic view of injury hampers efforts at prevention. For example, it may hinder the cooperation of communities that may have a deep belief in fatalism. Instead of spending resources to further the research and evaluation of an injury intervention, for instance, the scarce resources may rather be used to pacify gods or goddesses who are believed to be responsible for road traffic crashes on specific road segments. This practice is rampant throughout many LMICs.

Low Literacy Rates

The literacy rates in many LMICs are still very low. Such low literacy rates, which translate to people's apparent lack of knowledge about the causation of injuries, coupled with their cultural beliefs, have contributed to people's adherence to the fatalistic theory of injury as acts of God. It is therefore not surprising that there is a Ghanaian saying that "The dead (with reference to a pedestrian victim of a motor vehicle crash) is always guilty" – clearly victims of injury are blamed for their injuries! Another consequence of the low literacy rates in LMICs is that the vast majority of the population may not be able to read and understand simple road signs. Inability to comprehend and navigate road signs is indeed a recipe for disaster on our roadways.

Competing Health Problems

Despite the recent recognition of the importance of injury prevention by some governments and policymakers in many LMICs, there is still competition for the scarce resources with other pressing health problems such as recurrent infections, malnutrition, and HIV/AIDS. Currently, many low-income tropical countries spend a substantial portion of their health budget on malaria treatment. Therefore, prevention of injury, which is responsible for a substantial portion of the health burden, is relegated to the back burner.

Peculiar and Distinctive Causes of Injury

There are many other distinctive causes of injury or distinctive situations that may cause injury in LMICs. For example, there is a vastly different and distinct traffic mix in many LMICs that is not at all conducive to effective injury prevention. The roadways are often shared by petty traders, pedestrians, animals, cyclists, and motorized and nonmotorized vehicles. It is almost impossible to tell who has the right of way. Additionally, people are often transported in extremely unsafe manners such as in overloaded trucks, minibuses, and passenger-ferrying buses (Afukaar et al. 2003). The transfer of interventions must therefore take the local setting into consideration. For example, in high-income countries, drivers represent the group mostly involved in motor vehicle crashes, making air bags and seat belts excellent interventions. However, in most LMICs, victims of motor vehicle crashes are mostly pedestrians and occupants other than the driver, making air bags and seat belts less relevant in these settings. Other effective interventions to protect pedestrians and check overloading of vehicles become more relevant.



Fig. 34.1 Open hearth at floor level used for cooking in rural Ghana. Photo: Courtesy of Samuel N. Forjuoh

In many LMICs, open flames are a common feature of many households, particularly in the rural areas with no electrification or even among the slums in inner cities. Cooking devices may often include unsafe kerosene stoves and lanterns as well as open hearths at the floor level that are used for both cooking and warming. Among the important risks for fire and burns in such situations are the obvious lack of protective enclosures for the open fires and flames; availability of numerous flammable housing materials and highly flammable substances; instability of lanterns, stoves, and candles used for lighting and heating; and lack of exits in case of fire.

Figure 34.1 shows a typical cooking source in rural Ghana consisting of a tripod firestone made of dried mud and fueled by burning firewood that stays at the ground level in the kitchen within easy reach of children. This type of cooking source is a cause of many childhood burns in Ghana and other similar settings (Forjuoh et al. 1995).

Peculiar Political Situations

The peculiar political situations in many LMICs, often dominated by dictatorships or nondemocratic governments, do not augur well for injury prevention. First, adequate budgets may not be allocated for injury prevention. Second, leaders of dictatorships may flout international efforts at injury prevention. Finally, the right personnel may not always be allocated to the right office. In many of these countries, whose leaders came into power through the gun, soldiers who may not have the expertise are appointed as ministers without the requisite knowledge and expertise in the subject areas. Such situations impede injury prevention efforts. All these factors may act in unison to lessen the effectiveness of transferring potentially effective and promising interventions. For example, the continual importation of used vehicles without functional seat belts or with deployed air bags from high-income countries to LMICs for political and economic reasons does not promote seat belt use – an intervention with proven effectiveness.

Factors Affecting the Transfer of Interventions

Even if the challenges of transferring potentially effective and promising interventions can be overcome, there still remain other factors to seriously consider. First and foremost of these factors is whether the evidence for effectiveness generated in high-income countries must be imported directly into a low- or middle-income setting or whether the evidence should be produced locally. Next, is the cost involved in the transfer of interventions across different socioeconomic settings. Then there are feasibility issues as well as sustainability factors and barriers to implementation in the context of the local settings. These latter factors were elaborated in an extensive review in 1996. The authors used the four criteria – efficacy, affordability, feasibility, and sustainability of an intervention – to evaluate several transportation and home-related injuries with the intent of making recommendations for which interventions could be transferred to LMICs. They assessed several successful interventions using the four criteria and a three-scale rating (Forjuoh and Li 1996). Perhaps an overarching factor is whether an intervention is passive or active. The impact of development on health and the ethical and political issues of transferring interventions are also important factors. Finally, there are a few controversies regarding socioeconomic factors. Some of these factors are discussed below.

Local Evidence of Effectiveness of Transferred Interventions

With respect to the transfer of technology across settings, what is good for the goose may not necessarily be good for the gander. Therefore, transferred interventions from high-income countries should ideally be rigorously evaluated in the local setting before being extensively adopted and implemented. However, systematic evaluations of interventions in LMICs are limited. To date, only a handful of effective and promising interventions transferred from high-income countries have been evaluated sufficiently to assess their effectiveness in LMICs. Nonetheless, there are several innovative interventions that have been designed and developed in LMICs that appear intuitively effective and seem to work successfully. Therefore, the big question becomes whether we have to wait for evaluation of the effectiveness of available interventions in LMICs before adopting them.

Affordability of Interventions

In the transfer of an effective or a highly promising intervention from a high-income setting to a low- or middle-income setting, an important consideration is the cost involved to transfer and implement the intervention with reference to the health budget of the local setting (Forjuoh and Li 1996). A few questions then come up. For example, is it going to be expensive, somewhat affordable, or cheap to implement a successful transferred intervention? As an illustration, retrofitting all existing hotels in LMICs with fire sprinklers, an effective intervention for fires and burns, may be too expensive to even consider. On the other hand, constructing one of the effective speed-calming measures such as building speed strips or humps at built-up residential areas may be within the budget of many LMICs.

Feasibility of Interventions

The feasibility of transferring an effective or promising intervention, particularly in relation to the socioeconomic context of the local setting, is equally important (Forjuoh and Li 1996). Given that

the cost of a successful intervention is within the budget of LMICs, the next question is whether its transfer, adoption, implementation, and diffusion are possible and can be accomplished within the local setting. For example, air bags are an effective intervention and may not cost anything extra with the purchase of a new car. However, it may not be feasible to mandate all cars operating in LMICs to have these supplemental restraint systems at this time since most existing automobiles in many of these LMICs are not equipped with them and retrofitting them may not be realistic. Nonetheless, one has to start somewhere since that was the case in the USA when air bags were first introduced.

Sustainability of Interventions

Besides the cost of an intervention and whether or not it is feasible in a local setting, whether it can be sustained over a long period of time is another important consideration (Forjuoh and Li 1996). There are many situations in safety promotion and injury prevention where sustainability of an intervention is difficult to achieve. A typical example is an educational campaign to promote helmet use, especially without any accompanying legislation. All too often, initial gains in helmet use rates following an intensive campaign erode over time. In fact, sustainability is a big issue with active interventions that almost invariably involve behavioral change.

Barriers to Implementation of Transferred Interventions in Local Settings

Barriers to implementation of transferred interventions in local settings must never be discounted. It is, therefore, important to consider and even anticipate any potential barriers to the implementation of any transferred intervention with respect to the sociocultural milieu of LMICs even after overcoming cost, feasibility, and sustainability issues. Whether there may be any taboos with implementation of a transferred intervention should be brainstormed with community leaders and activists. The presence of any other potential socioeconomic controversies with the transfer of a particular intervention must be dealt with.

Passive vs. Active Interventions

Whether an intervention is a passive or an active one, in relation to whether an individual action is required, is also a factor to consider in the transfer of interventions. The air bag is an example of a passive intervention because it deploys automatically during a vehicular frontal crash with no action required of the occupant. An approved safety helmet, on the other hand, is an example of an active intervention because the individual rider must actually pick it up to wear, and must wear it correctly in order to reduce the risk of a head injury in case of a crash.

Due to their mechanism of action, passive interventions may not need to be really tested in local settings before adoption in LMICs. Of course, evaluation of their effectiveness in local settings should be done and is in fact desirable. However, it appears that compared to active interventions, passive interventions are more likely to be affordable, feasible, and sustainable in LMICs than active interventions.

Impact of Development on Health

Indisputably, development or advancement in economic productivity, technological sophistication, and industrial capability has a profound effect on the health of populations. However, health status changes with development have not corresponded to a change for the better in all sectors of the population and in all settings. In addition, many development policies designed to improve the living conditions and standards of communities can have unintended or unexpected consequences on health. Therefore, transferring even the most effective interventions without due consideration of the culture of the local setting may sometimes result in unintended consequences. For instance, road barriers may be less effective in preventing pedestrian injuries if the intended “safe” routes for pedestrians are not designed to be culturally acceptable, most convenient, and the easiest routes.

A classic example was provided by Baker in a pedestrian safety study in Rio de Janeiro, where instead of using a newly constructed pedestrian bridge, many people chose to climb over the concrete divider topped with a wire fence that was constructed to separate pedestrians and bicyclists from the motorized traffic in a superhighway (Baker 1975). The people, especially those with a child or a bicycle, presumably chose to run across the traffic lanes instead of using the pedestrian bridge because of its long flight of stairs.

Ethical and Political Issues of Transferring Interventions

Since product safety standards for consumer protection are vital (Van Weperen 1993), there is a need to develop mechanisms to regularly inform policymakers in LMICs about unsafe products that are marketed internationally. The political situation of LMICs may also play an important role in the transfer of interventions since politicians have the final say in effecting laws and many effective interventions stem from legislation and strict enforcement.

Maximizing the Effectiveness of Transferred Interventions

To maximize the effectiveness of a transferred intervention, a careful evaluation of what might work in the local setting is important. This is because what has been found to be effective in a high-income setting may not necessarily be effective in a low-income setting. A modification or adaptation of the intervention may be needed in order to achieve maximum effectiveness vis-à-vis the sociocultural milieu through improvisation or innovation.

Improvisation of Transferred Interventions

Improvisation of a transferred intervention is its modification or adaptation using available local materials. For example, in some LMICs, bamboo beams have been used in place of iron bars to separate pedestrians from the motorized traffic. This improvisation has been used successfully throughout Ghana to enhance traffic safety.

Innovation of Transferred Interventions

Innovation of a transferred intervention is its modification or adaptation using an entirely novel approach. An example is the safe bottle lamp invented by Dr. Wijaya Godakumbura, a Consultant General Surgeon in Sri Lanka who won the Rolex Award in 1998. Having witnessed so many cases of burns from overturned unsafe home-made lamps in Sri Lanka, Dr. Godakumbura initiated an action that resulted in the invention of the safe bottle lamp, which is described in detail in section “Selected Transferred Interventions.”

Selected Transferred Interventions

Selected effective and highly promising interventions that have already been transferred and implemented in some LMICs are presented below. They are limited to traffic and home-related injuries since these account for the greatest injury burden in terms of mortality and disability-adjusted life years in LMICs. In addition, most interventions that have been developed have focused on these injuries.

Traffic-Related Injury Interventions

Although there have been no studies using sound randomized controlled trials to show the effectiveness of transferred traffic-related interventions in LMICs, studies using other epidemiologic designs have shown the effectiveness of some of these transferred traffic-related interventions in some LMICs including Brazil, China, Ghana, Greece, Hungary, Indonesia, Malaysia, Singapore, South Africa, Taiwan, and Thailand (Table 34.1). Examples are seat belts, speed bumps, daytime running

Table 34.1 Selected traffic-related interventions that have already been transferred and tested in LMICs

Intervention	Level of effectiveness	Type of intervention	Country tested
Air bags	Effective	Passive	–
Seat belts	Effective	Active	Greece China
Speed limits	Effective	Passive	Brazil South Africa
Speed bumps	Effective	Passive	Ghana
Legislation and enforcement of motorcycle helmets	Effective	Active/passive	Indonesia Taiwan Thailand
Road safety education	Promising	Active	Singapore
Daytime running lights on vehicles	Promising	Passive	Hungary
Daytime running lights on motorcycles	Promising	Passive	Malaysia Singapore
Increases in fines and suspension of driver's license	Promising	Active	Brazil
Increased legal age of motorcyclists from 16 to 18 years	Promising	Active	Malaysia

lights on motorcycles, increases in fines and suspension of driver licenses, legislation and enforcement of motorcycle helmets, and increasing the legal age of motorcyclists from 16 to 18 years. In this section, traffic-related interventions that have been transferred and implemented in some LMICs are discussed, along with selected effective interventions for which no evaluations have been done (e.g., air bags).

Air Bags

To date, there have been no studies evaluating the effectiveness of air bags in any LMICs. Yet, as a passive intervention that requires no action on the part of the individual to be protected, the air bag is an effective intervention that is highly recommended for transfer to LMICs. Air bags have proven to be very effective in increasing vehicular occupant safety (O'Neill 1984, 1992, 2009), although there have been problems of early models of air bags hurting or even killing infants, children, and small adults, even in some low-speed collisions. Due to their high level of effectiveness, air bags have been mandated in all vehicles manufactured or imported into the USA since 1994. The same situation applies to many other high-income countries. The fact that there is really no barrier to using it if a vehicle is already equipped with an air bag makes it affordable and feasible in LMICs. Therefore, government authorities of LMICs must also begin to mandate only the importation of cars with air bags, which is very feasible.

Seat Belts

The effectiveness of seat belts in reducing injury and death in motor vehicle crashes has also been well established (Latimer and Lave 1987; Rivara et al. 2000). Seat belts are indeed estimated to help reduce motor vehicle fatalities by 50% and serious injury by 55%. Due to their proven effectiveness in preventing injuries, affordability, and feasibility, they are highly recommended as a transferred intervention to LMICs. As an active intervention for vehicular occupants where some action is required on the part of the user vs. a passive intervention like the air bag that requires no action on the part of the user, accompanying behavioral and legislation interventional strategies are also required to maximize effectiveness. Such strategies include mandatory seat belt laws, public education about the benefits of seat belt use, and legislation on the availability of functional seat belts in vehicles.

To date, there have been no studies to assess the effectiveness of seat belts in reducing injuries in LMICs. However, there is some empirical evidence on using educational interventions to increase seat belt use. In Greece, moderate increases in seat belt use were observed following a comprehensive intervention campaign to increase seat belt use (Petridou et al. 1999). Another study in China also demonstrated a 20% increase in seat belt use following enhanced training and enforcement of seat belt use (Stevenson et al. 2008). These data are encouraging given the low usage of seat belts, even in some high-income countries. Seat belt usage is abysmal in many LMICs. A Malaysian study reported that 60% of apparently restrained taxicab drivers observed at the curbside did not fasten the latch of their seat belts (Hauswald 1997).

While it may not appear appropriate to mandate the use of seat belts because most cars may not be equipped with functional seat belts, governments of LMICs should seriously consider issuing policies to ban the importation of automobiles without functional seat belts into their countries. However, this is a very delicate political issue. Seat belt use is clearly effective, affordable, and feasible and can be sustained in LMICs. Increasing seat belt use in LMICs could help to reduce motor traffic injuries and deaths among vehicular occupants.

Speed Limits

Speeding on highways is a major cause of motor vehicle crashes. Limiting the traveling speeds of motorists using speed limits and other speed-calming measures has proven effective in reducing motor traffic crashes (Moore et al. 1995). As an intervention, the enactment of speed limits has been associated with reduced pedestrian and vehicular occupant injuries. In Johannesburg, South Africa, a significant decrease in the number of motor vehicle crash patients admitted to a city hospital was observed following enforcement of a speed limit law (Wilkson 1974). The reduction in traffic crashes and deaths in Brazil has also partly been attributed to posted speed limits beginning in 1998 (Poli de Figueiredo et al. 2001). Speed limits are an effective intervention that is definitely useable, affordable, and feasible in LMICs.

However, like seat belts, the effectiveness of speed limits as an intervention can be enhanced by accompanying intervention strategies. Such strategies may include the use of traffic-calming measures as described below and enforcement of speed limits, which may not be feasible in many LMICs due to myriad reasons, including limited resources available to the police.

Speed Bumps in Reducing Pedestrian Injuries

Traffic-calming or physical speed-reducing measures such as use of speed bumps on the roadway have been shown to allow pedestrians to coexist with motor vehicles in relative safety in many high-income settings (Kjemtrup and Herrstedt 1992). Traffic-calming or physical speed reduction assists in slowing motor vehicular speeds, particularly at roundabouts, narrow sections of roadways, and congested segments of roadways. Luckily, many of the principles that have been used to design guidelines for traffic-calming in high-income countries may also be applicable to LMICs. This has indeed been shown in Ghana.

Using a before-and-after study design in Ghana, it has been shown that speed bumps are effective in reducing traffic-related injuries, especially pedestrian injuries. The use of rumble strips and speed bumps on a crash hot spot segment of a major highway reduced the number of traffic-related crashes by 35%, fatalities by 55%, and serious injury by 76% (Afukaar 2003). The relative success of this intervention has led to its widespread diffusion with improvisation and innovation. For example, a wide range of materials is now being used in the whole country to construct speed bumps on the roadways including vulcanized rubber, hot thermoplastic materials, bituminous mixes, mud mixed with stones, concrete, and bricks. In addition, speed bumps are being combined with rumble strips and speed humps to drastically slow down vehicles and improve pedestrian safety at potentially dangerous segments on the roadways in built-up areas. Speed-calming measures are indeed one inexpensive but effective intervention for LMICs, and their use must be propagated widely.

Legislation and Enforcement of Motorcycle Helmet Wearing

There is considerable evidence that mandatory helmet laws with enforcement not only lead to increased helmet use but also greatly alleviate the burden of traffic-related injuries. Such evidence exists both in high-income countries and some LMICs. An evaluation of helmet use and traumatic brain injury before and after the introduction of legislation in Italy revealed a huge pay-off. Helmet use increased from 20% to over 96%, while traumatic brain injury admissions for motorcycle/moped crashes decreased by 66% (Servadei et al. 2003). An Indonesian study also reported a 1.7 relative risk of injury among nonhelmeted motorcyclists – 32% of head injuries among injured motorcyclists wearing helmets vs. 52% among those not wearing helmets (Conrad and Bradshaw 1996).

A comprehensive literature review focusing on the effectiveness of motorcycle helmet use, and on mandatory helmet laws and their enforcement reported successes in Taiwan and Thailand. In

Taiwan, there was a reported 14% decline in motorcycle fatalities and a 22% reduction of head injury fatalities with the introduction of a helmet law. Additionally, there were 6,240 quality-adjusted life years gained due to reduction in head injuries from motorcycle crashes as a result of helmet law enforcement (Tsauo et al. 1999). In Thailand, where 70–90% of all crashes involve motorcycles, helmet use increased fivefold, the number of injured motorcyclists decreased by 34%, head injuries decreased by 41%, and deaths decreased by 21% following enforcement of a helmet law (Ichikawa et al. 2003). For LMICs with high rates of motorcycle injuries, enforced, mandatory motorcycle helmet laws are potentially one of the most cost-effective interventions available (Hyder et al. 2007). Therefore, governments of LMICs should seriously consider legislating and enforcing the use of motorcycle helmets in their countries.

Road Safety Education

There is ample evidence to show that providing information and education to road users about the hazards on the roadways improves their knowledge and leads to subsequent reduction of pedestrian injuries. In fact, educating pedestrians on how to cope with the complex traffic environment has been touted as one of the most essential elements to improve pedestrian safety and reduce pedestrian injuries. However, road safety education translating to behavior change does not always result in reduction of road traffic crashes. Nonetheless, road safety education is a promising intervention that is affordable, feasible, and sustainable in LMICs. This intervention has already been transferred to some LMICs with measurable successes. For example, a study in Singapore reported a 52% reduction in serious injuries and 66% reduction in minor injuries over a 9-year period following an intensive road safety education programs for school children (Them and Lee 1993).

Daytime Running Lights on Vehicles

Daytime running lights are designed to increase visual contrast between vehicles and their backgrounds, thereby improving their probability of being easily noticed and detected. Several countries including Canada, Denmark, Finland, Hungary, Iceland, Norway, and Sweden require motor vehicles to have their lights on during the daytime. Studies from these and other countries have generally indicated that daytime running lights on vehicles are associated with modest reductions in multiple-vehicle daytime crashes, especially those involving vehicles approaching from the front or side. A Hungarian study showed a 13% reduction in frontal and side vehicle collisions during the daytime over a period when a partial obligation for using daytime running lights was in effect (Hollo 1998). As a passive intervention, the use of daytime running lights on vehicles is a promising intervention that appears feasible, and therefore, governments of LMICs must consider mandating all imported vehicles to be equipped with daytime running lights.

Daytime Running Lights on Motorcycles

Daytime running lights are similarly designed to increase visual contrast between motorcycles and their backgrounds to improve their conspicuity. Low conspicuity, or the inability of the motorcycle and rider to be seen by other road users, is thought to be associated with motorcycle crash-related injury and death. Despite the limited evidence base, several countries including Austria, Malaysia, and the USA have made daytime use of headlights mandatory, and riders in other countries have voluntarily adopted this and other strategies. A preliminary analysis of the short-term impact of a running headlights intervention in Malaysia revealed a significant drop in conspicuity-related motorcycle crashes by 29% (Radin et al. 1996). Another study in Singapore found a 15% reduction in fatal

motorcycle crashes 14 months following the passage of a legislation requiring all motorcyclists to switch on their headlamps (Yuan 2000). Therefore, to make this a passive intervention, the use of daytime running lights on motorcycles is also a promising intervention that appears feasible, and so governments of LMICs must consider mandating all imported motorcycles to be equipped with daytime running lights.

Increases in Fines and Suspension of Driver Licenses

Punitive measures have been utilized over the years in designing some interventions in high-income countries. For example, increases in fines and suspension of driver licenses for traffic offenses have been associated with some successes. Whether such interventions may be feasible and sustainable in LMICs at this time is unclear, albeit there have been some attempts in many LMICs. The reduction in traffic crashes and deaths in Brazil has partly been attributed to increases in fines and suspension of driver's licenses, along with posted speed limits beginning in 1998 (Poli de Figueiredo et al. 2001). As a promising intervention, increases in fines and suspension of drivers' licenses may be useable in some LMICs where strict enforcement can be guaranteed. However, in many LMICs, law-enforcement personnel are not adequately equipped with the resources to enforce even existing traffic laws. Therefore, it is obvious that enforcing new traffic laws may pose serious challenges.

Increasing the Legal Age of Motorcyclists from 16 to 18 Years

There is increasing use of many promising interventions that involve reduction of risk exposure. These interventions have been used to reduce many types of injuries. Among the measures used to reduce exposure to road injury risk is placing restrictions on motor vehicle users. This intervention has been used successfully to reduce motorcycle crashes as shown in Malaysia. Increasing the legal riding age from 16 to 18 years in Malaysia led to substantial reduction in motorcycle crashes. This intervention was also found to have the greatest benefit–cost ratio among several interventions proposed to reduce motorcycle crashes (Norghani et al. 1998). Other LMICs must therefore test this intervention as well as others that involve reduction of risk exposure.

Home-Related Injury Interventions

There is even less evidence regarding the effectiveness of interventions for home-related injuries than for traffic-related injuries. However, most interventions for home-related injuries are passive, requiring no action on the part of the individual to be protected. Therefore, a few home-related injury interventions have been transferred successfully to some LMICs without any evaluations. In this section, home-related interventions that have been transferred to LMICs are discussed, along with selected effective interventions for which no evaluations have been done (e.g., smoke alarms). The discussion is limited to interventions for burn-related, poisoning-related, drowning-related, and fall-related injuries.

Smoke Alarms

Smoke alarms have been shown to reduce deaths from fires and burns by about 70%, particularly in conflagrations (Runyan et al. 1992). In many high-income countries, because 75% of all deaths from

fires and burns are due to house fires or conflagrations, smoke alarms have been virtually mandated in all houses. Although conflagrations may not be much of a problem in many LMICs, smoke alarms as an efficacious intervention are still highly recommended. Many LMICs now have smoke alarms installed in residential dwellings. Thus far, there have been no studies to evaluate the effectiveness of smoke alarms in LMICs. Evidence from rigorous evaluation research based on local data may help facilitate increased use of smoke alarms in LMICs.

Residential Sprinkler Systems

Residential sprinkler systems automatically put out fires. As a passive intervention, their use is also highly recommended for LMICs. In fact, automatic sprinkler systems have been successfully used to protect industrial and commercial buildings and their occupants for more than 100 years in many high-income countries. Historically, the place that has offered the least amount of fire protection to occupants was, and still is, their own home. The purpose of a residential sprinkler system built to the required standard is to provide a sprinkler system that aids in the detection and control of residential fires, and thus provides improved protection against injury, life loss, and property damage. From a performance perspective, if the room of fire origin is sprinkled, a sprinkler system designed and installed in accordance with the residential sprinkler standards is expected to prevent flashover and improve the occupant's opportunity to escape or to be rescued. With the growth in the hotel industry in many LMICs, governments in these LMICs must ensure that all new buildings are appropriately with residential automatic sprinkler systems and smoke alarms. As fitted with smoke alarms, evidence of the effectiveness of this passive intervention in LMICs is lacking.

Hot-Water Temperature Regulation Laws

In many high-income countries, there are laws requiring that the thermostat of a new water heater offered for sale or lease for use in a residential unit be preset by the manufacturer to no higher than 120°F (or 49°C) or to the minimum setting on any water heater that cannot be set as low as that temperature. However, there is no law that prohibits an owner of an owner-occupied residential unit or resident of a leased or rented residential unit from readjusting the temperature setting after occupancy. Any readjustment of the temperature setting by the resident relieves the owner or agent of an individual residential unit and the manufacturer of water heaters from liability for damages attributed to the readjustment by the resident. Regulating hot-water temperatures has been found to avert many burn injuries in high-income countries.

This intervention, however, may make little sense in LMICs where many households have no electricity and even those with electricity may not use water heaters. However, with increasing urbanization and industrialization, many LMICs have begun or are beginning to rely on hot water from water heaters installed in residential homes and industrial settings. Therefore, to avert burn injuries, it behooves governments of LMICs to begin to pass similar laws as those passed in high-income countries on presetting temperatures on the thermostats.

Developing and Promoting Use of Safer Lamps and Stoves

In many LMICs, serious burns have been associated with use of unsafe lamps and stoves. This is because these unsafe lamps and stoves, which are the main sources of heating and cooking, use fossil fuel. Tipping of these lamps and stoves often results in thermal burns, particularly to infants and toddlers. Developing safer lamps and stoves has therefore been associated with tremendous reduction of burns. An example is the safe bottle lamp invented by Dr. Wijaya Godakumbura, a Consultant

General Surgeon in Sri Lanka that won a Rolex Award in 1998 (www.safebottlelamp.org). The safe bottle lamp was invented to replace the cheap, unsafe makeshift lamps that are used for lighting in rural Sri Lanka. Some of its design features include being short and heavy to prevent it from easily tipping over, having two flat sides so in case of tipping it does not roll over, being made of thick glass so it does not break if it tips over, and having a screw-on metal lid so the oil does not spill if it tips over. Obviously, developing safer consumer products such as safer lamps and stoves is a promising intervention that will go a long way to reduce home-related injuries in LMICs.

Banning the Manufacture and Sale of Fireworks

The use of fireworks is often associated with injuries. Fireworks cause about 10,000 injuries treated in US hospital emergency departments each year (American Academy of Pediatrics 2001). This observation increased the call to ban fireworks by many organizations including Prevent Blindness America, which supports the development and enforcement of bans on the importation, sale and use of all fireworks and sparklers, except those used in authorized public displays by competent licensed operators. However, while many high-income countries have banned the manufacture and sale of fireworks, burn injuries through fireworks are still common in many LMICs. In many of these LMICs, national holidays as well as religious and other festivities are still celebrated by setting off fireworks indiscriminately. Banning the manufacture and sale of fireworks – a promising intervention – in LMICs is an effective means of eliminating the social and economic impact of fireworks-related trauma and damage. In LMICs, where fireworks are not manufactured, governments should seriously consider banning their importation. This requires a lot of political will on the part of individual low- and middle-income country governments.

Child-Resistant Containers

Child-resistant packaging is one of the best documented successes in childhood unintentional injury prevention. Child-resistant packaging has been found to be an effective intervention for poisoning from medications, fuel, household chemicals, and pesticides. In many LMICs, paraffin oil or kerosene used for cooking and heating is frequently stored in bottles and other containers previously used for storing beverages. This practice exposes children to poisoning. In South Africa, a successful program to intervene in childhood poisoning involved free distribution of child-resistant paraffin containers. This 14-month intervention reduced the annual incidence of poisoning from 104 per 100,000 to 54 per 100,000 (Krug et al. 1994). Clearly, child-resistant packaging is an intervention that is affordable and feasible in LMICs.

Other Potentially Transferable Interventions

Besides the effective handful of highly promising interventions that have undergone some evaluation in LMICs, there are several others that have apparently been tested in some LMICs with documented success. There are still a few others that have been suggested based on risk factor analysis using epidemiologic studies in some LMICs – burn-related and poisoning studies in Ghana, Brazil, Bangladesh, India, Peru, Malaysia, and Zimbabwe (Forjuoh et al. 1995; Werneck and Reichenheim 1997; Daisy et al. 2001; Bawa Bhalla et al. 2000; Delgado et al. 2002; Azizi et al. 1993; Nhachi and Kasino 1994). In addition, a book published by Berger and Mohan presented several sound ideas as potential strategies for injury intervention in LMICs (Berger and Mohan 1996). These are all potentially transferable interventions that should be tested and adopted with caution and evaluated whenever it is possible. Table 34.2 catalogues some of these suggested interventions.

Table 34.2 Selected transferable home-related interventions available for testing and adoption in LMICs

Type of injury	Proposed intervention	Country of original study suggesting intervention	
Burn	Separating cooking areas from living areas	Ghana	
	Reducing use of indoor fires for cooking		
	Reducing storage of flammable substances in the home		
	Ensuring cooking surfaces are at appropriate heights		
	Closely supervising younger children especially during cooking		
	Reducing overcrowding in the kitchen		
	Poisoning	Storing cooking utensils in the kitchen out of reach of children	Brazil
		Developing standards and codes for fire-retardant garments	Peru
		Improving housing quality/safety for low-income people	Bangladesh
		Replacing high-pressure home cooking stoves with low-pressure wick stoves	India
Introducing inexpensive stands to stabilize bottle lamps		a	
Locking away medicines and other toxic substances		Malaysia	
Prohibiting use of secondhand household containers for storage		Malaysia	
Drowning		Designing inexpensive, childproof containers for kerosene, pesticides, etc.	Zimbabwe
		Establishing national and multinational regulations for toxic disposal	a
		Replace lead paint and glazes with unleaded substitutes	
	Covering residential wells with grills	a	
Fall	Increasing inspection of ferries for safety		
	Improving local and regional flood control measures		
	Fencing close-by lakes and riverbanks	–	
	Building flood control embankments	–	
	Fencing of domestic swimming pools	–	
	Providing more stable climbing devices at construction sites (e.g., welded ladders)	a	
	Modifying routines to reduce climbing of tall trees		
	Playground standards legislation	–	
	Window guards legislation	–	
	Using stair gates and guard rails	–	

*From Berger and Mohan (1996)

Conclusion and Discussion

Injury intervention in LMICs is still in its infancy, and much can be learned from the successes in high-income countries. One obvious advantage of transferring interventions across settings is conservation of resources. In considering options for technology transfer to LMICs, however, a careful evaluation of what might work in these settings is very important. The extent to which an intervention that has been found to work effectively in one setting can be successfully transferred to another depends on several factors. Although high-income countries have had great success in identifying, inventing, and implementing effective and many promising interventions, only a handful of these interventions have been tested in the LMIC setting. One reason for this has been the lack of trained personnel in injury intervention, which seems to be improving. Another is limited funding – public efforts at safety promotion and injury prevention are generally poorly funded. However, this also seems to be improving.

It is evident that effective injury interventions require multidimensional strategies including education, legislation, and environmental modification due to the multifactorial nature of injury causation. It is also clear that maximum effectiveness of injury intervention requires a combination of effective interventions and strategies as well as cooperation among several individuals working at different levels across several sectors of the community. Such multi-interventional and intersectoral approach has already been used to successfully intervene in many other health-related problems in LMICs. International assistance could assist with program development and the training needs of LMICs.

Earlier calls made for policy examination, changes, and response to the neglected problem of injury in LMICs as well as the need to conduct more research (Forjuoh and Gyebi-Ofosu 1993; Zwi et al. 1996) are still very relevant today as are more recent ones (Forjuoh 2003, 2006; Peden et al. 2004, 2008; Perel et al. 2007; Borse and Hyder 2009). The toll of human suffering from injury that is predictable and preventable should be minimized in LMICs with all available useable interventions.

References

- Afukaar, F. K. (2003). Speed control in developing countries: Issues, challenges and opportunities in reducing road traffic injuries. *Injury Control and Safety Promotion*, 10, 77–81.
- Afukaar, F. K., Antwi, P., & Ofosu-Amaah, S. (2003). Pattern of road traffic injuries in Ghana: Implications for control. *Injury Control and Safety Promotion*, 10, 69–76.
- Agyei-Mensah, S., & de Graft, A. A. (2010). Epidemiological transition and the double burden of disease in Accra, Ghana. *Journal of Urban Health*, 87, 879–897.
- American Academy of Pediatrics. (2001). Committee on Injury and Poison Prevention. Fireworks-related injuries to children. *Pediatrics*, 108, 190–191.
- Azizi, B. H., Zulkifli, H. I., & Kasim, M. S. (1993). Circumstances surrounding accidental poisoning in children. *The Medical Journal of Malaysia*, 49, 132–137.
- Baker, S. P. (1975). The man in the street: A tale of two cities. *American Journal of Public Health*, 65, 524–525.
- Baker, S. P., & Haddon, W., Jr. (1974). Reducing injuries and their results: The scientific approach. *The Milbank Memorial Fund Quarterly. Health and Society*, 52, 377–389.
- Baker, S. P., O'Neill, B., Haddon, W., Jr., & Long, W. B. (1974). The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma*, 14, 187–196.
- Bawa Bhalla, S., Kale, S. R., & Mohan, D. (2000). Burn properties of fabrics and garments worn in India. *Accident Analysis and Prevention*, 32, 407–420.
- Berger, L. R., & Mohan, D. (1996). *Injury control: A global view*. Delhi: Oxford University Press.
- Bishai, D., Hyder, A. A., Ghaffar, A., Morrow, R. H., & Kobusingye, O. (2003). Rates of public investment for road safety in developing countries: Case studies of Uganda and Pakistan. *Health Policy and Planning*, 18, 232–235.
- Borse, N. N., & Hyder, A. A. (2009). Call for more research on injury from the developing world: Results of a bibliographic analysis. *The Indian Journal of Medical Research*, 129, 321–326.
- Conrad, P., & Bradshaw, Y. S. (1996). Head injury comparisons between helmet users and non-helmet users. *Accident Analysis and Prevention*, 28, 193–200.
- Daisy, S., Mostaque, A. K., Bari, T. S., et al. (2001). Socioeconomic and cultural influence in the causation of burns in the urban children of Bangladesh. *Journal of Burn Care & Rehabilitation*, 22, 269–273.
- De Haven, H. (1942). Mechanical analysis of survival in falls from heights of fifty to one hundred and fifty feet. *War Medicine*, 2, 586–596.
- Delgado, J., Ramirez-Cardish, M. F., Gilman, R. H., et al. (2002). Risk factors for burns in children: Crowding, poverty and poor maternal education. *Injury Prevention*, 8, 38–41.
- Forjuoh, S. N. (1996). Injury control in developing nations: What can we learn from industrialized countries? *Injury Prevention*, 2, 90–91.
- Forjuoh, S. N. (2003). Traffic-related injury prevention interventions for low-income countries. *Injury Control and Safety Promotion*, 10, 109–118.
- Forjuoh, S. N. (2006). Burns in low- and middle-income countries: A review of available literature on descriptive epidemiology, risk factors, treatment, and prevention. *Burns*, 32, 529–537.
- Forjuoh, S. N., & Gyebi-Ofosu, E. (1993). Injury surveillance: Should it be a concern to developing countries? *Journal of Public Health Policy*, 14, 355–359.

- Forjuoh, S. N., & Li, G. (1996). A review of successful transport and home injury interventions to guide developing countries. *Social Science & Medicine*, *43*, 1551–1560.
- Forjuoh, S. N., Guyer, B., Strobino, D. M., et al. (1995). Risk factors for childhood burns: A case-control study of Ghanaian children. *Journal of Epidemiology and Community Health*, *49*, 189–193.
- Frenk, J., Bobadilla, J. L., Sepuulveda, J., & Cervantes, M. L. (1989). Health transition in middle-income countries: New challenges for health care. *Health Policy and Planning*, *4*, 29–39.
- Gibson, J. J. (1961). *The contribution of experimental psychology to the formulation of the problem of safety: A brief for basic research* (reprinted from Behavioral approaches to accident research) (pp. 77–89). New York, NY: Association for the Aid of Crippled Children.
- Gordon, J. E. (1949). The epidemiology of accidents. *American Journal of Public Health*, *39*, 504–515.
- Haddon, W., Jr. (1970). On the escape of tigers: An ecologic note. *American Journal of Public Health*, *60*, 2229–2234.
- Haddon, W., Jr. (1980). Advances in the epidemiology of injuries as a basis of public policy. *Public Health Reports*, *95*, 411–421.
- Hauswald, M. (1997). Seat belt use in a developing country: Covert compliance with a primary enforcement law in Malaysia. *Accident Analysis and Prevention*, *29*, 695–697.
- Heinrich, H. W. (1931). *Industrial accident prevention: A scientific approach*. New York: McGraw-Hill (quoted in Hollnagel E (2009) Safer complex industrial environments: A human factors approach. CRC Press).
- Hollo, P. (1998). Changes in the legislation on the use of daytime running lights by motor vehicles and their effects on road safety in Hungary. *Accident Analysis and Prevention*, *30*, 183–199.
- Hyder, A. A., Waters, H., Phillips, T., & Rehwinkel, J. (2007). Exploring the economics of motorcycle helmet laws – implications for low and middle-income countries. *Asia-Pacific Journal of Public Health*, *19*, 16–22.
- Ichikawa, M., Chadbunchachai, W., & Marui, E. (2003). Effect of the helmet act for motorcyclists in Thailand. *Accident Analysis and Prevention*, *35*, 183–189.
- Karkee, G. J., Namibisan, S. S., & Pulgurtha, S. S. (2010). Motorist actions at a crosswalk with an in-pavement flashing light system. *Traffic Injury Prevention*, *11*, 642–646.
- Kjemtrup, K., & Herrstedt, L. (1992). Speed management and traffic calming in urban areas in Europe: A historical view. *Accident Analysis and Prevention*, *24*, 57–65.
- Krug, A., Ellis, J. B., Hay, I. T., Mokgabudi, N. F., & Robertson, J. (1994). The impact of child-resistant containers on the incidence of paraffin (kerosene) ingestion in children. *South African Medical Journal*, *84*, 730–734.
- Kwan, I., & Mapstone, J. (2006). Interventions for increasing pedestrian and cyclist visibility for the prevention of death and injuries. *Cochrane Database of Systematic Reviews*, *18*(4), CD003438.
- Latimer, E. A., & Lave, L. B. (1987). Initial effects of the New York State auto safety seat belt law. *American Journal of Public Health*, *77*, 183–186.
- McArthur, D. L., & Kraus, J. F. (1999). The specific deterrence of administration per se laws in reducing drunk driving recidivism. *American Journal of Preventive Medicine*, *16*(1 Suppl), 68–75.
- Moore, V. M., Dolinis, J., & Woodward, A. J. (1995). Vehicle speed and risk of a severe crash. *Epidemiology*, *6*, 258–262.
- Nhachi, C. F., & Kasino, O. M. (1994). Household chemicals poisoning admissions in Zimbabwe's main urban centres. *Human and Experimental Toxicology*, *13*, 69–72.
- Norghani, M., Zainuddin, A., Radin Umar, R. S., & Hussain, H. (1998). *Use of exposure control methods to tackle motorcycle accidents in Malaysia* (Research Report 3/98). Serdang, Malaysia: Road Safety Research Center, University Putra Malaysia.
- O'Neill, B. (1984). *A note on air bag effectiveness*. Washington, DC: Insurance Institute for Highway Safety.
- O'Neill, B. (1992). Effectiveness of airbags. *The New England Journal of Medicine*, *326*, 1091.
- O'Neill, B. (2009). Preventing passenger vehicle occupant injuries by vehicle design – a historical perspective from IIHS. *Traffic Injury Prevention*, *10*, 113–126.
- Oman, A. R. (1971). The epidemiologic transition: A theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*, *49*, 509–538.
- Peden, M., Scurfield, R., Sleet, D., et al. (2004). *World report on road traffic injury prevention*. Geneva: World Health Organization.
- Peden, M., Oyegbite, K., Ozanne-Smith, J., et al. (2008). *World report on child injury prevention*. Geneva: World Health Organization and UNICEF.
- Perel, P., Ker, K., Ivers, R., & Blackhall, K. (2007). Road safety in low- and middle-income countries: A neglected research area. *Injury Prevention*, *13*, 227.
- Petridou, E., Trichopoulos, D., Stappa, M., Tsoufis, Y., & Shalkidou, A. (1999). Effectiveness of a comprehensive multisector campaign to increase seat belt use in the greater Athens area, Greece. *American Journal of Public Health*, *89*, 1861–1863.
- Poli de Figueiredo, L. F., Rasslan, S., Bruscahin, V., Cruz, R., & Rochae Silva, M. (2001). Increases in fines and driver license withdrawal have effectively reduced immediate deaths from trauma on Brazilian roads: First-year report on new traffic code. *Injury*, *32*, 91–94.

- Press, E. (1948). Epidemiologic approach to accident prevention. *American Journal of Public Health*, 38, 1442–1445.
- Radin, U. R., Mackay, M. G., & Hills, B. L. (1996). Modeling of conspicuity-related motorcycle accidents in Seremban and Shah Alam, Malaysia. *Accident Analysis and Prevention*, 28, 325–332.
- Rivara, F. P., Koepsell, T. D., Grossman, D. C., & Mock, C. (2000). Effectiveness of automatic shoulder belt systems in motor vehicle crashes. *JAMA*, 283, 2826–2828.
- Runyan, C. W., Bangdiwala, S. I., Linzer, M. A., Sacks, J. J., & Butts, J. (1992). Risk factors for fatal residential fires. *The New England Journal of Medicine*, 327, 859–863.
- Servadei, F., Begliomini, C., Gardini, E., et al. (2003). Effects of Italy's motorcycle helmet law on traumatic brain injuries. *Injury Prevention*, 9, 257–260.
- Stevenson, M., Yu, J., Hendrie, D., et al. (2008). Reducing the burden of road traffic injury: Translating high-income country interventions to middle-income and low-income countries. *Injury Prevention*, 14, 284–289.
- The National Committee for Injury Prevention and Control. (1989). *Injury prevention: Meeting the challenge*. New York: Oxford University Press.
- Them, M. M., & Lee, J. (1993). Road safety education for school children. *World Health Forum*, 14, 407–409.
- Tsauo, J. Y., Hwang, J. S., Chin, W. T., et al. (1999). Estimation of expected utility gained from the helmet law in Taiwan by quality adjusted survival time. *Accident Analysis and Prevention*, 31, 253–263.
- Van Houten, R., & Malenfant, L. (1992). The influence of signs prompting motorists to yield before marked crosswalks on motor vehicle-pedestrian conflicts with flashing amber. *Accident Analysis and Prevention*, 24, 217–225.
- Van Weperen, W. (1993). Guidelines for the development of safety-related standards for consumer products. *Accident Analysis and Prevention*, 25, 11–18.
- Werneck, G. I., & Reichenheim, M. E. (1997). Paediatric burns and associated risk factors in Rio de Janeiro, Brazil. *Burns*, 23, 478–483.
- Wilkson, A. E. (1974). Speed limits and accidents. *South African Medical Journal*, 48, 1323.
- Yuan, W. (2000). The effectiveness of the “ride bright” legislation for motorcyclists in Singapore. *Accident Analysis and Prevention*, 32, 559–563.
- Zegeer, C. V., Stewart, J. R., Huang, H. H., et al. (2005). Safety effects of marked versus unmarked crosswalks at uncontrolled locations: Final report and recommended guidelines. FHWA Publication Number: HRT-04-100.
- Zwi, A. B. (1996). Injury control in developing countries: Context more than content is crucial. *Injury Prevention*, 2, 91–92.
- Zwi, A. B., Forjuoh, S., Murugusampillay, S., et al. (1996). Injuries in developing countries: Policy response needed now. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 90, 593–595.

Chapter 35

Implementing and Evaluating Interventions

Caroline F. Finch

Why Intervention Research?

Advances in real-world injury prevention will be achieved only if research efforts are directed toward fully understanding the implementation context, while continuing to build the evidence base for the efficacy and effectiveness of interventions (Finch 2006). Throughout earlier chapters of this book, guidance has been given on the design, conduct, and analysis of research studies leading to understanding both injury causation and reduction. This chapter is concerned with theoretical underpinnings of research needed to address intervention implementation and effectiveness research. This is particularly important for injury prevention because understanding the barriers and facilitators to the widespread adoption and sustainability of interventions is vital to ensuring effective and sustainable injury prevention.

Overall, while there is a relatively large literature relating to the rationale, design, and development of injury interventions and their evaluation in efficacy studies, there have been few published effectiveness studies describing aspects of injury prevention implementation. This is a major gap because the studies that do describe the presence (or absence) of injury prevention benefits associated with interventions are unable to explain the reason for the benefits or lack thereof. Too often, we are left with only knowledge that something did or did not work in one study, and there is no guidance on how to translate those findings to another setting or a similar injury problem. For example, recent studies describing the benefits of an exercise training program to prevent injuries in community soccer have shown only limited success, because few of the targeted participants adopted the program and there was a perception that it was not relevant to the real-world community sport setting in which it was implemented (Kilding et al. 2008; Soligard et al. 2008; Steffen et al. 2008). Similarly, there are challenges when translating “ideal” falls-prevention interventions involving risk factor assessment and home-based interventions, because implemented versions of the same program need to be modified to suit community settings and hence may no longer be effective (Hendriks et al. 2008). There can also be suboptimal uptake of Tai Chi falls prevention interventions and low levels of ongoing adoption that can compromise study effectiveness (Logghe et al. 2011). Without additional information about some of the important implementation factors (e.g., program uptake, pragmatic

C.F. Finch, PhD (✉)

Australian Center for Research into Injury in Sport and its Prevention, Monash Injury Research Institute,
Monash University, Building 70, Clayton, VIC Australia, 3800
e-mail: caroline.finch@monash.edu

changes to interventions for delivery purposes, etc), reasons for the lack of success could not have been identified in these studies.

There are many reasons why implemented programs and their evaluations can fail and these can be summarized as (Bierrman 2006; Stame 2010):

- Program theory failure in which the developed intervention either (a) is too complex for the setting in which it is implemented or (b) does not lead to the desired behavior change because of the way it was designed.
- Implementation failure in which the intervention does not adequately address (a) the implementers' own behaviors in relation to intervention delivery or (b) the context in which it is to be delivered.
- Methodology failure in which (a) internal and/or external validity are compromised; (b) the evaluation plan and tools are not up to the task required of them to demonstrate the outcomes of the intervention; or (c) no concurrent process evaluation has been undertaken to explain unexpected observations or to confirm expectations.

When injury studies have considered implementation issues, this has typically been as a minor component of an effectiveness study, with most studies evaluating only some aspects of intervention implementation. There is no doubt that there are many complexities involved in conducting implementation research in real world settings. Many studies only report injury outcomes without also examining the required intermediary behavior change, such as exercise adoption or protective equipment use, which is necessary to link those reductions firmly to the implemented preventive measures. In contrast, others have only reported these proxy or intermediary outcomes and assumed that they will lead to the desired injury outcome (Rivara 2008). The vast majority of studies do not even consider whether the intervention target groups actually adopted, or complied with, the intervention. Nor do they recognize that individual safety behavior change is also significantly influenced by other factors such as the form of the intervention delivery, the person delivering it, and the broader ecological system in which the intervention has taken place.

There is no doubt that there is a complex relationship between desired injury reduction benefits and how interventions are packaged, delivered, and promoted (Nilsen 2004; MacKay and Vincenten 2009). It has been argued that the conduct of well-designed large-scale intervention effectiveness trials has been hampered because of a lack of theoretical considerations in their design, implementation, and evaluation (Thompson and Sacks 2001; Glasgow et al. 2003; Glasgow et al. 2004; Timpka et al. 2006; Armstrong et al. 2008; Catford 2009; Ward et al. 2009). Prevention research efforts will only develop further if they begin to incorporate such considerations, as has also been demonstrated to be the case for injury prevention research (Trifiletti et al. 2005; McGlashan and Finch 2010). Moreover, as also discussed elsewhere in this book, many different implementation and intervention delivery approaches could be considered to support prevention efforts, either in isolation or jointly. These range from educational/behavior change strategies (Christoffel and Gallagher 2006; Robertson 2007; Provvidenza and Johnston 2009) to environmental modifications (Christoffel and Gallagher 2006; Robertson 2007), to making policy/law changes (Scott 2004; Christoffel and Gallagher 2006; Robertson 2007), to public awareness/advocacy (Henley 2004; Christoffel and Gallagher 2006), and stakeholder engagement (Brussoni et al. 2006; Christoffel and Gallagher 2006; MacKay and Vincenten 2009).

To further injury prevention, it will be necessary for implementation studies to have a firm theoretical basis. Because of the general lack of international implementation research in any aspect of injury prevention, there is very little direct information about how best to conduct intervention studies in relevant community settings. While some theoretical considerations have been developed specifically for some safety programs (e.g., safe communities (Nilsen 2006)), and specific settings (e.g., sports injury prevention delivery contexts (Finch and Donaldson 2010)), most of the available

Table 35.1 Reported use of explicit behavioral and social science theory applications in injury prevention research

How the behavioral or social science theory was used	Review of 37 papers describing theory use in unintentional injury studies published during 1998–2001 (Trifiletti et al. 2005)	Review of 11 papers describing theory use in sports injury prevention studies published prior to June 2009 (McGlashan and Finch 2010) ^a
To guide program design and/or implementation and/or evaluation measures ^c	43 ^b	8
To develop or evaluate a measured theory or model constructs	7	7
To test application of a theory	5	4
Other (including not stated)	6	3

^aOnly one of the sports injury studies applied two theories; all others only reported use of a single theory

^bThis number exceeds the total number of papers reviewed because several papers used more than one theory and so this refers to the number of theory applications

^cThis is a large category that combines several types of studies but was used in both review papers to categorize the studies. Most of the reviewed studies in those two papers did not evaluate the effectiveness of injury prevention interventions

examples come from broader health promotion or behavioral science applications. Table 35.1 summarizes how behavioral and social science theory has been used to date in the small number of injury prevention studies that report it, highlighting this as a major knowledge gap. Overall, very few studies have reported theory use and, when they have, this has been most commonly in terms of program/implementation/evaluation design (Trifiletti et al. 2005; McGlashan and Finch 2010).

Theoretical considerations have important implications for how intervention studies are conducted and reported. Improved reporting standards for implementation studies are needed to provide a more comprehensive analysis of the factors affecting intervention uptake and effectiveness (Finch 2006; Roen et al. 2006). Application of health promotion frameworks to evaluate the public health impact of interventions could also potentially help to better understand contextual and policy influences in this setting.

Despite the availability of injury prevention interventions with proven or likely efficacy, it is clear that limited research attention has focused on understanding the intervention implementation context and processes, including barriers and facilitators to sustainable programs. To address this challenge, injury prevention research aimed at demonstrating real world uptake of interventions needs to:

- Draw on available evidence for the efficacy of interventions in terms of reductions in both injury and injury risk, as well as intermediate behavioral measures (sometimes referred to as impact measures).
- Engage relevant stakeholders and end user groups in implementation and injury prevention research from the outset.
- Continue to partner with these stakeholder groups in further intervention and intervention delivery developments.
- Develop multifaceted and multi-action strategic approaches toward injury prevention in relevant real-world culturally relevant settings.
- Develop and evaluate strategic implementation plans designed to address key barriers and facilitators toward intervention uptake at all levels.
- Adopt a multidisciplinary approach that embraces both qualitative and quantitative research methodologies.
- Include measures of cost-effectiveness for sustained program implementation.

Effectiveness Versus Efficacy

Research studies for demonstrating the preventive potential of injury interventions can be broadly categorized into two types: efficacy and effectiveness (Table 35.2). The differences between the design and conduct of efficacy and effectiveness studies have been discussed by a number of authors (Glasgow et al. 2003; Finch 2006; Mallonee et al. 2006; Prochaska et al. 2007; Glasgow 2008; van

Table 35.2 A comparison of the key features in the design and conduct of efficacy and effectiveness studies

Component	Efficacy studies	Effectiveness studies	Considerations for the design and evaluation of interventions in implementation studies
Study design	<ul style="list-style-type: none"> Highly controlled Examples are RCTs and controlled laboratory studies 	<ul style="list-style-type: none"> Level of control is much less Allow assessment of relevant implementation factors Examples include quasi-experimental, pre-post, interrupted-time series 	<ul style="list-style-type: none"> Include randomization of units to intervention implementation groups Control groups add strength and reduce the chance of ecological fallacy
Intervention delivery	<ul style="list-style-type: none"> Under full research team control Well-defined protocols must be adhered to Deliverers employed by the researchers 	<ul style="list-style-type: none"> Interventions and/or accompanying resources are delivered or implemented by others not directly employed by the research team 	<ul style="list-style-type: none"> Motivation and commitment of deliverers, as well as their usual practices, are important Potential barriers/enablers of the intended delivery to be assessed before finalization of the intervention design and its full implementation
Study participants, intervention allocation, and targeting	<ul style="list-style-type: none"> Under the strict control of researchers Analysis according to intention-to-treat principles Participants are a relatively homogenous group that meet specific criteria 	<ul style="list-style-type: none"> Allocation plan is determined by the researchers but undertaken by others Intervention is delivered to a defined group or population (i.e., a heterogeneous group) 	<ul style="list-style-type: none"> Different levels of intervention uptake need to be monitored Reasons for why there is/is not uptake should be assessed
Sample size and length of study	<ul style="list-style-type: none"> Adequate numbers of study participants needed to ensure power Follow-up over large amounts of time 	<ul style="list-style-type: none"> Of shorter duration Involves many more study participants 	<ul style="list-style-type: none"> Shorter-duration studies can show immediate behavior/knowledge change effects Longer studies needed to show sustainability and maintenance of these changes
Intervention protocol and setting constraints	<ul style="list-style-type: none"> Rigidly structured Must be adhered to Interventions cannot be modified but are developed specifically with the specific target population in mind No assessment of generalizability across settings 	<ul style="list-style-type: none"> Protocol and interventions must be flexible enough to allow adaptations for the specific context and setting/s if necessary during implementation Can assess the extent to which the intervention can be successfully used in different settings 	<ul style="list-style-type: none"> Engaging stakeholders in the development of the delivery plan Pilot testing of the intervention and delivery plan are needed Community feedback should be sought

(continued)

Table 35.2 (continued)

Component	Efficacy studies	Effectiveness studies	Considerations for the design and evaluation of interventions in implementation studies
Staffing, local infrastructure, and funding issues	<ul style="list-style-type: none"> • Very labor intensive • Require full funding for both intervention delivery and evaluation data • Involve a limited number of staff with specific training in the study protocol 	<ul style="list-style-type: none"> • Intervention delivery is usually the responsibility of the real-world agencies/individuals • Only limited support for implementation from research funds • Involve people with different training experiences • Evaluation often conducted and funded by researchers 	<ul style="list-style-type: none"> • Stakeholder engagement and buy-in needed from the outset • Intervention programs more likely successful if these groups are also involved as equal partners during all stages of an implementation trial and evaluation

Adapted from (Finch 2009)

Tiggelen et al. 2008; Finch 2009). Table 35.2 summarizes the key features of these study types and highlights some of the particular challenges that arise in the conduct of implementation studies.

In efficacy studies, the preventive effect of interventions is assessed under ideal and tightly controlled conditions and individual injury reduction outcomes are desired to be demonstrated. The highest form of this research evidence is from randomized controlled trials (RCTs), though other experimental designs can also contribute knowledge. The high level of control is necessary to ensure large effect sizes, corresponding to the preventive capacity of the intervention under study. The vast majority of injury prevention intervention trials are efficacy studies.

Effectiveness research is undertaken when the preventive effect of the intervention is assessed under everyday circumstances. This implies little or no control over how the intervention is implemented, though in practice this may be hard to ensure. The goal of effectiveness studies is to determine the extent to which the intervention actually prevents injuries when delivered as it would be used in real world practice. Broader implementation research studies measure and report factors such as how the intervention was delivered as well as how it was complied with and used. This focus is necessary because if efficacious interventions are not widely adopted, complied with and sustained as ongoing practice, then it is very unlikely they will have any significant or long-lasting injury prevention impact (Finch 2006).

Intervention Research Requires Appreciation and Understanding of Ecological Systems

The above discussion has highlighted that for full impact, any intervention aimed at individual-focused injury reductions must consider the broader context in which implementation of the intervention needs to occur. Individuals, while the target of prevention programs, are heavily influenced by the groups they belong to and the broader social and cultural norms related to the injury risk behavior being targeted. Recognition of this is conceptualized in ecological models of injury prevention (Eime et al. 2004; Sleet and Gielen 2004; Allegrante et al. 2006; Allegrante et al. 2010). Importantly, the more individual-based approaches cannot alter environmental (physical, social, or cultural) factors that influence the initiation and maintenance of safety behavior. Ecological models, on the other hand, identify intrapersonal factors, sociocultural factors, policies, physical environments, etc., as levels of influence on injury prevention behaviors. As such, they recognize

that many factors combine to influence an individual's protective or risk-reduction behavior (and any decisions to not adopt the behavior).

The injury iceberg model proposed by Hanson et al. (2005) is a conceptual explanation of this ecological model for the application to community safety interventions. It emphasizes that latent failures can occur when implementing community safety programs if interpersonal, organizational, community, and societal levels of influence of community safety are not considered from the outset. Too often injury intervention studies ignore most, if not all, of these influences and only focus on intrapersonal factors (Allegrante et al. 2006; Allegrante et al. 2010).

The only sports injury prevention study to apply the ecological model to date (McGlashan and Finch 2010) developed and evaluated a comprehensive protective eyewear promotion program for squash players (Eime et al. 2004). Through surveys of squash players and venue managers, it was determined that protective eyewear was not readily available, and that players' behaviors, knowledge, and attitudes did not favor its use. A protective eyewear promotion program was developed with components to inform and educate players and squash venue operators of the risk of eye injury and of appropriate protective eyewear. Other components of the program addressed the availability of the eyewear and incentives for players to use it. A reported structural strength of the ecological intervention was the strong collaborative links across multidisciplinary researchers, the squash sport governing body, eyewear manufacturers, squash venue personnel and players, from the outset. This also allowed some attempts toward longer-term dissemination and sustainability of more widespread eye injury prevention measures in the sport. The evaluation outcomes of the program, published separately, demonstrated significant effects on knowledge about appropriate eyewear use (Eime et al. 2005).

There is an apparent disconnect in the literature between what is called an "ecological design" and studies informed by the ecological model, as described here. It is important to realize that they are not necessarily the same thing. In the former, standard epidemiological study designs (including RCTs) are used but the unit of analysis is a group, rather than the individual (Hingson et al. 2001; Connor 2004; Rivara 2008). However, this does not mean that studies adopting this design necessarily consider the full range of ecological determinants of the outcome of interest. By definition, however, many studies using the ecological model of behavior change do need to adopt some aspects of ecological study designs because they are necessarily concerned with group or population-level outcomes, not just individual behavior change.

Rivara (2008) discussed a range of outcomes that were appropriate to injury research ranging from serious injury (such as death and hospitalization) to moderate/mild injury to injury-free events to behaviors and knowledge/attitudes. While he discussed the use of ecological study designs, he only considered these outcomes at the individual level. A review of the effectiveness of community-based injury prevention programs (Nilsen 2005) also found that most studies only reported injury rate reductions and were not concerned at all with contextual factors that could explain the study findings or provide additional information about the interventions being tested.

Figure 35.1 provides an extension of Rivara's (2008) pyramid of outcomes, which stresses the need for outcomes across other levels of the ecological context for injury prevention and also recognizes the overlapping influence that different levels can operate on each other. Thus, ecologically driven intervention implementation studies need to specifically focus on understanding drivers of behavior and related behavior change across multiple levels. Many behavioral models (Ajzen 1985; Ajzen 1991; Eime et al. 2005; Gielen et al. 2006a) emphasize that intention to undertake a behavior is an important outcome stage in its own right and so this has also been added as an outcome level to Rivara's original list. Importantly, intervention implementation studies do not ignore the injury outcomes or recording of injury-free events because they are effectiveness studies, but they do give more weight to the behaviorally orientated factors.

The remainder of this chapter presents three specific theoretical frameworks and approaches that show good promise for injury prevention intervention research. These include the use of Intervention

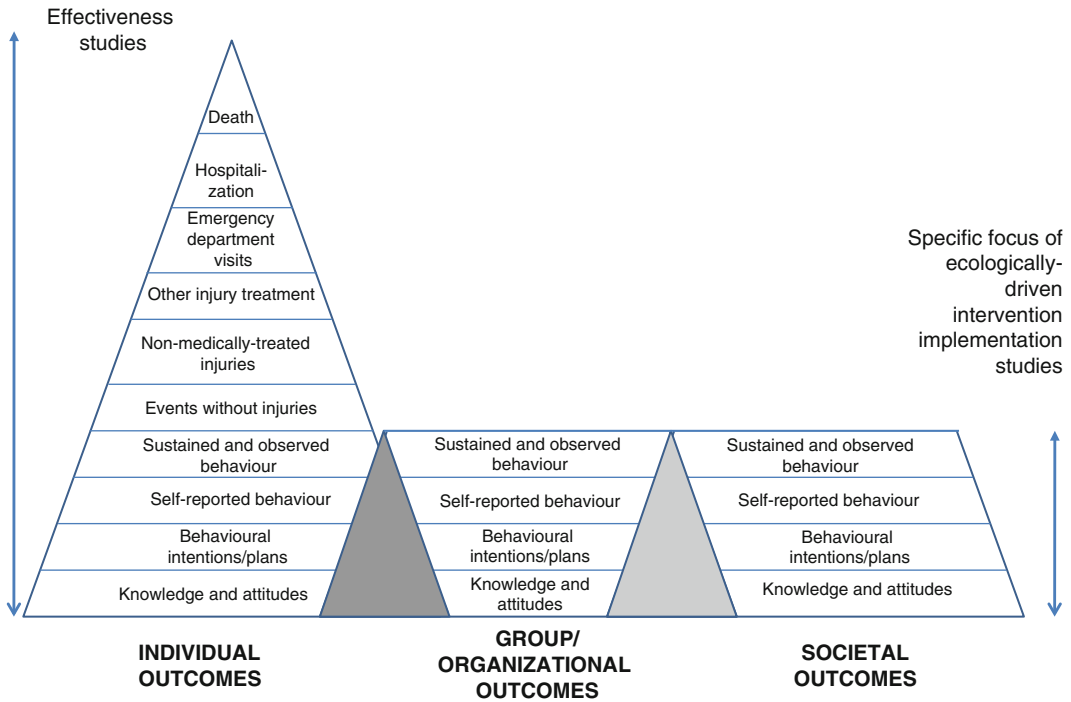


Fig. 35.1 The pyramid of outcomes used for ecologically driven intervention implementation studies. The shaded areas represent areas of mutually overlapping influences. Societal outcomes can also be directly related to individual outcomes. (Adapted from Rivara 2008)

Mapping is a tool to assist in the intervention development process itself, Diffusion of Innovations theory to guide efforts in the planning of intervention strategies, and the RE-AIM framework (reach, effectiveness, adoption, implementation, maintenance) most commonly (but not exclusively) used as an evaluation and evaluation planning tool. The injury literature that has applied these theoretical approaches to date is also summarized, particularly as it pertains to unintentional injury prevention.

It is acknowledged that these are not the only frameworks, models, or approaches that can be used for injury intervention research. While some studies may have a theoretical underpinning to their research, this is not always stated. Some authors describe other systematic approaches that could be used (e.g., Nold and Bochmann 2010; Winston and Jacobsen 2010), and these broadly mirror the systematic approaches advocated in the following sections.

Intervention Mapping

While there has been an increasing number of published papers describing so-called injury intervention implementation studies and evaluating the impact of interventions on both injury and process outcomes, there is surprisingly little information about how interventions were developed in the first place or how they were actually delivered. Most studies will mention that they have taken an evidence-based approach toward defining their intervention and that they have then implemented it according to scientific principles. However, specific information describing exactly how the intervention was packaged for its delivery, or how it was refined for the particular setting of its targeting, are details that are often not reported in the literature. Yet it is exactly this information that

has the greatest potential to further and improve implementation studies because it gives valuable additional cues as to why certain things do/do not work and what may need to be done to ensure program sustainability or translation to other groups or settings. Moreover, interventions that are effective in one setting may not necessarily be effective in others; some modification will be needed for each new contextual setting. Information about how the intervention was developed and delivered in the first place can inform this.

Developing interventions for implementation that will be fully effective is a complex process that involves many components, not just previous efficacy evidence that they should work. Interventions that are developed from a theoretical basis are likely to be more successful than others. However, it is also important that consideration be given to the practical strategies that will need to be adopted, or refined from the theoretical foundation, when considering any implementation study. This is the premise behind the Intervention Mapping approach toward the planning of interventions (Bartholomew et al. 2006). This approach considers intervention delivery to be necessary within an ecological framework in which behavioral and social science considerations are paramount. Intervention Mapping draws on previous behavioral change models as applied to complex societal systems such as the PRECEDE-PROCEED model (Green and Kreuter 1991) and Diffusion of Innovations theory (Rogers 2003). However, its authors also recognize that no single behavioral change theory is fully applicable to all contexts and interventions. Therefore, the approach allows intervention developers and implementers to draw on the best theoretical basis for their setting (Bartholomew et al. 2006).

The Intervention Mapping approach provides guidance for decision making across all stages of the intervention process from intervention planning and implementation processes to the final evaluation (Bartholomew et al. 2006). The Intervention Mapping protocol provides a systematic summary of the necessary steps and tasks that need to be undertaken to ensure the combining of empirical evidence, relevant theoretical constructs, contextual knowledge, and context-specific experience to inform the development, implementation and evaluation of health promotion/injury prevention interventions. Document matrices are advocated as a means of recording decisions about how to influence the desired behavior change within the specific social and physical environments embedded within ecological systems necessary to prevent injuries. While such a systematic approach has the potential to both help plan and implement effective interventions, it also has the added benefit of assisting with understanding why any intervention does or does not work.

Intervention Mapping achieves this through six steps:

1. Conducting a needs assessment or problem analysis.
2. Creating matrices of change objectives based on the determinants of behavior and environmental conditions.
3. Selecting relevant theory-based intervention methods and practical strategies.
4. Translating methods and strategies into an organized program.
5. Planning for adoption, implementation, and sustainability of the program.
6. Generating an evaluation plan.

Figure 35.2 summarizes the specific tasks that should be undertaken in each step. Importantly, while represented in a linear form, these steps and tasks should be undertaken in an iterative manner with new information being fed back to reinforce earlier steps. For detailed guidance on how to complete this Intervention Mapping process, with examples, the reader is referred to the book by Bartholomew and colleagues (Bartholomew et al. 2006).

Completion of each of the six Intervention Mapping steps requires working through the following core processes in an interactive way that incorporates feedback loops and revision of prior decisions, as appropriate (Bartholomew et al. 2006):

- Pose a relevant question.
- Brainstorm a provisional list of answers or range of possible solutions.

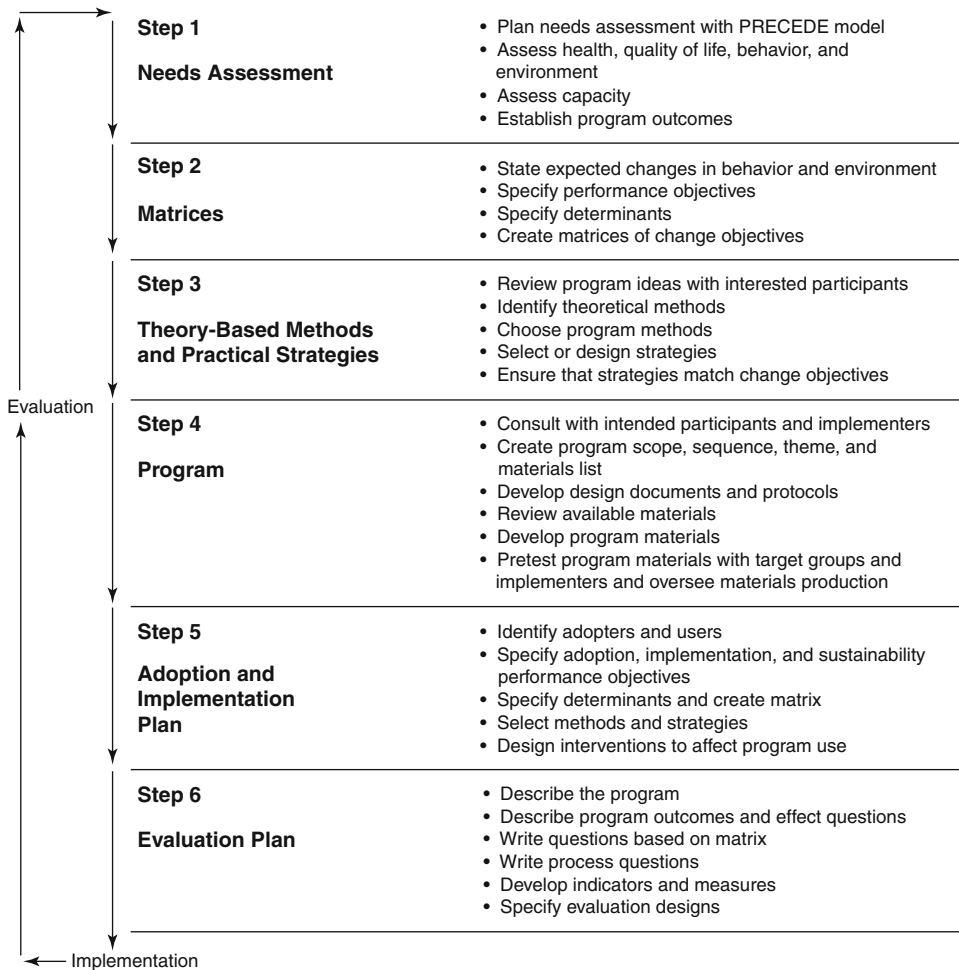


Fig. 35.2 The intervention mapping process reproduced from Bartholomew et al. (2006) *Planning health promotion programs*. An intervention mapping approach. Jossey-Bass, San Francisco. Reprinted with permission of John Wiley and Sons, Inc

- Review the available literature related to the topic – both peer and non-peer reviewed.
- Access and identify an appropriate theoretical approach and use this to further develop or justify the solutions.
- Conduct new research and acquire new data through the lens of the identified theoretical approach.
- Formulate the final answers/responses to the initial question posed, or its revised form.

In the injury context, Intervention Mapping has been reported in three contexts: a parent education intervention to prevent violence in school students in the USA (Murray et al. 1998); safety interventions in metalworking shops also in the USA (Brosseau et al. 2007); and a Dutch school-based physical activity injury prevention program (Collard et al. 2009a). Each of these studies used an iterative Intervention Mapping approach in which underpinning theoretical considerations appropriate to the specific context were integrated with existing evidence from published literature and new data collected from the target population during the intervention development phase. Two of the studies

reported use of a five-step process, as advocated in the first edition of *Intervention Mapping* (Bartholomew et al. 1998), which did not include the needs assessment phase as one of the major stages. In the most recent edition (Bartholomew et al. 2006), the needs assessment is added as the first of six crucial steps.

As part of an initial needs assessment, staff at two public middle schools in Texas identified parental monitoring of their children as a potential modifiable behavior that could influence the level of violence among adolescents and so this was recommended as the target for a brief school-based intervention (Murray et al. 1998). Data to inform the *Intervention Mapping* process were obtained through self-reported surveys completed by students and both qualitative and quantitative information collected through parental telephone interviews and focus group sessions; this was considered at all stages of the intervention development process. Staff from the two schools were involved in both the data collection phase and the development of the intervention and its delivery plan which had theoretical underpinnings from social cognitive theory (Bandura 1986), the theory of planned behavior (Ajzen 1985; Ajzen 1991), and the transtheoretical model (Prochaska et al. 1997). The identified intervention strategy was educational newsletters for parents, which was later shown to be popular and to be associated with higher levels of parental monitoring (Murray et al. 1999).

The *Intervention Mapping* approach was also found to provide valuable new insights into strategies for the development and delivery of interventions to reduce the risk of injuries in people engaged in metalworking businesses (Brosseau et al. 2007). Separate intervention variations were considered for owners of businesses and employees within them. Both considered personal, social, and environmental determinants of machine-related hazards and amputations and drew on the social cognitive theory. Information fed into the interactive information mapping process came from consultations with an advisory board with members across all relevant sectors; a review of machine safety documents, regulations, and standards; direct discussions with employees; safety audits of machines in businesses; presentation of draft materials to the expert groups and pilot businesses; and piloting of the intervention. The authors considered that the use of *Implementation Mapping* allowed them to develop a robust uniform intervention that could still be adapted to be applicable at multiple sites.

Focus group sessions, supplemented with a small number of interviews, were initially conducted with physical education teachers from 12 Dutch secondary schools to develop an intervention that could improve knowledge in teachers, students, and their parents about injury prevention in physical activity (Collard et al. 2009a, b). Knowledge needs and educational formats were determined through application of the attitude, social influence, and self-efficacy (ASE) model (De Vries et al. 1995), which combines the Theory of Planned Behavior (Ajzen 1985; Ajzen 1991) and Social Cognitive Theory (Bandura 1986). Although it is not clear from the paper how the target group members contributed to all stages of the mapping, the developed intervention materials were piloted with teachers and children from six schools, with the assumption that their acceptance by these two groups also indicated that they would be acceptable to the parents. The intervention was subsequently implemented in a cluster randomized trial in 40 Dutch schools (Collard et al. 2010b). While it had an overall nonsignificant effect on injury rates, there was a significantly reduced rate of injuries in children who were classified as being low-active. The design of this study has since been suggested as the basis of a more general approach toward intervention development for sports injury prevention (Collard et al. 2009b). Interestingly, the finding that the intervention was effective only in part of the target population mirrors the findings from the controlled evaluation of “RiskWatch,” another teacher-led UK school-based intervention covering different safety behaviors (Kendrick et al. 2007) and further justifies the need to conduct detailed process evaluations alongside each intervention.

Designing appropriate interventions and accompanying intervention strategies and evaluation plans is a complex and time-consuming process. While the *Intervention Mapping* approach does not

remove this, it does provide a systematic approach toward undertaking this important activity. When used fully and interactively, it ensures that the views, needs, and desired behavioral actions of each ecological level target group are considered at all stages of the planning and evaluating process.

Diffusion of Innovations

As mentioned in the last section, successful implementation and implementation studies require both a well-defined and targeted intervention and detailed information about the context in which it is to be implemented and how this will affect adoption of the intervention. Intervention Mapping provides a systematic approach for achieving both of these goals. Because full understanding of the implementation context is critical to successful interventions and their diffusion through the target groups, it is worth discussing this aspect further.

One of the most successful approaches toward understanding the uptake of interventions is the Diffusion of Innovations theory, first proposed by Rogers as early as 1962 (Rogers 2004). The importance of this theory is demonstrated by its subsequent underpinning of aspects of Intervention Mapping (Bartholomew et al. 2006) and the RE-AIM framework (Glasgow et al. 1999) to be discussed in the next section. Indeed, the theory is one of the most-cited social theories in public health application. However, despite its wide application in other areas of health promotion and public health and its clear relevance to injury prevention initiatives (Nelson and Moffitt 1988; Aldoory and Bonzo 2005; Gielen et al. 2006a, b; Collard et al. 2010a, b), it appears to have had only limited application to the injury field to date (Trifiletti et al. 2005; McGlashan and Finch 2010).

The strength of the theory lies with its focus on communication of new ideas (or innovations) within multi-level ecological structures that require some form of behavioral, social, or other change across one or more levels for the innovation to be considered effective. Rogers (2003) provides the following definitions for the main components in the diffusion process:

- Innovation – an idea, practice, or object perceived as new by an individual or other unit of adoption; attributes of the innovation are paramount to its subsequent adoption.
- Communication channel – the means by which messages get from one individual to another. There are various ways this can be achieved, depending upon the specific purpose such as social marketing/mass media (Henley 2004; Christoffel and Gallagher 2006), advocacy (Pitt and Spinks 2004; Christoffel and Gallagher 2006), or through public/policy agencies (Foster et al. 2004; Christoffel and Gallagher 2006).
- Innovation-diffusion process – whereby an individual passes from first knowledge of an innovation, to forming an attitude toward the innovation, to a decision to adopt or reject, to implementation of the new idea, and to confirmation of this decision. This relies on key societal members who are opinion leaders or change agents.
- Innovativeness – the degree to which an individual (or other adopter unit) is relatively earlier in adopting new ideas than other members of their social system. The concept of adopter category is relevant here as different individuals will respond to the intervention in different ways.
- Rate of adoption – relative speed with which an innovation is adopted by members of a social system. Not all interventions will be adopted at the same rate of uptake.
- Social system – a structured set of interrelated units (e.g., people) that are engaged in joint problem solving to accomplish a common goal. This includes defining opinion leaders, change agents, and other influencers of opinion or adoption.
- Consequences – the changes that occur to an individual or social system because of the adoption or rejection of an innovation.

According to the theory, the attributes of any new interventions that would need to be considered are (Rogers 2003):

- Relative advantage – the degree to which the new intervention is conceived to be better than existing programs or practices.
- Compatibility – the degree to which the new intervention is consistent with the existing values, past experiences, and needs of people targeted by it (i.e., the potential adopters).
- Complexity – the extent to which a new intervention is perceived to be easy (or difficult) to understand and use.
- Trialability – the extent to which a new intervention may be tested by potential adopters.
- Observability – the extent to which the new intervention and its benefits are visible to others.

Interventions which are ranked more positively with regard to advantage, compatibility, trialability, and observability and which are also perceived to be easier to use and understand will be taken up more readily and more rapidly than other interventions.

As Rogers (2003) himself defines it, diffusion is “the process through which an innovation, defined as an idea perceived as new, spreads via certain communication channels over time among members of a social system.” The Diffusion of Innovations model can be used to determine both the level and rate of intervention uptake, so that different interventions can be compared both within target groups and across them. Accordingly, members of a social system can be characterized as belonging to one of five ideal categories (Rogers 2003):

- Innovators – people who are very ready to adopt new innovations, even before the full value to society has been shown; they are very much ahead of most other people in terms of their willingness to try new ideas. Their behavior involves a certain amount of risk and they need to accept the consequences of adopting an innovation that may not be successful. Innovators have a very important role in terms of introducing new ideas to community groups and play “a gate-keeping role in the flow of new ideas into a system.”
- Early adopters – highly influential opinion leaders in any system who are seen as the people to give general advice about the suitability and usefulness of new innovations. For this reason, they are often seen as the change agents for ensuring rapid diffusion of new ideas. Once enough early adopters take on the innovation or intervention, they can then trigger rapid diffusion – i.e., they form a critical mass.
- Early majority – while they do not adopt interventions as rapidly as the two previously mentioned groups, they do so more rapidly than the average person within a societal system does. They are, therefore, a very important group in ensuring high uptake rates and comprise about one-third of any societal group. While it takes them longer to decide to take up an intervention than those earlier groups, once they do so they become very strong supporters and hence help convince other members also to take on the behavior.
- Late majority – like the early majority, this group also comprises about one-third of the population. They tend to be more skeptical about the innovation than earlier groups but generally will later adopt it if their concerns about the new idea are removed or if there are significant peer-influences or economic reasons for doing so.
- Laggards – these people tend to be suspicious of new ideas and interventions and of change agents operating to introduce them. It takes considerable time and persuasion, most commonly from their own peers, before they will adopt new innovations.

The practical implication of this adopter categorization is that different strategies will be needed to target different members of the same community groups depending upon their readiness-to-adopt category, somewhat akin to the implications of the transtheoretical model (Prochaska et al. 1997) for individual behavior change interventions. Importantly, Diffusion of Innovations theory considers the communication process to be one whereby ideas within a societal system converge to a common

understanding (or misunderstanding), as a result of individual members creating new knowledge and experiences and sharing this with other members of that system.

Of course, to have long lasting public health effects, any intervention that is adopted needs to be sustained and the desired behavior change and structural systems to support this maintained. With regard to sustained adoption of any prevention program with ongoing desired injury prevention benefits, intervention studies should monitor the level to which the innovation is taken up by members of the target group, including their knowledge about it and how they use it; how the intervention is used in practice and ongoing implementation and continued use of the innovation (Gielen et al. 2006a, b).

Trifiletti et al. (2005) undertook a review of the extent to which the Diffusion of Innovations (and other behavioral and social science theories) had been used in research on unintentional injury prevention. For the period 1988–2001, they were able to identify 12 studies that had applied the Diffusion of Innovations to injury problems, but only two of these papers had applied it to unintentional injury, in this case both were bicycle helmet use studies (Farley et al. 1996; Farley et al. 1997). More recently, a similar review of theory use in sports injury research published prior to mid-2009 (McGlashan and Finch 2010) found only two studies to have since applied the theory to sports injury prevention: one study related to helmet use in recreational activities undertaken in ski areas (Andersen et al. 2004) the other to coach education in relation to sports concussion (Sawyer et al. 2010).

In the helmet study, skiers and snowboarders were both observed and interviewed in ski fields in northwestern USA and Canada (Andersen et al. 2004). Collected data were used to test three specific hypotheses arising from application of the Diffusion of Innovations theory that (1) prevalence of helmet use by skiers and snowboarders would have increased over time; (2) helmet use would be greater among certain groups (i.e., in the more educated guests, frequent skiers/snowboarders, experts and intermediates, and snowboarders); and (3) the rate of increase in helmet use would be higher in some groups (i.e., guests residing in the Rocky Mountain region and Canada, who were experts, skied or snowboarded the largest proportion of days, and snowboarders). The results confirmed the first two hypotheses but there was no statistical support for the third. The authors interpreted this result as providing no support for the critical mass concept within Diffusion of Innovations which essentially states that there is a specific point at which enough people in a population undertake the desired behavior to make further diffusion of the innovation self-maintaining (Rogers 2003). Two possible explanations for this were provided: either that 1-year follow-up is not long enough to test for “critical mass” effects or that the marketing of helmets in the preceding 2 years had reached all adopter groups equally, so there was no differential uptake across them.

In the concussion education study, the Centers for Disease Control developed a toolkit entitled “Heads Up: Concussion in high school sport” to be used by coaches to prevent and manage concussion in school athletes in the USA (Sawyer et al. 2010). To inform the development, dissemination plan and evaluation, 497 high school athletic coaches were surveyed about their demographics; receipt of the toolkit; actual or intended use to the toolkit and reasons for this; their views on the overall appeal, ease of use and usefulness of the content; expected benefits of the toolkit, especially in relation to other prevention methods and resources; and whether they would recommend it to others. The responses were found to support the premises of the Diffusion of Innovations theory and provided clear guidance for the ongoing targeting of the toolkit to coaches.

The most recent injury intervention study to apply the Diffusion of Innovations involved assessment of the adoption and implementation of an educational program for the prevention of intentional incidents (with high potential for injury) (Henderson et al. 2006). The intervention was aimed at professionals who provided mental health programs to children who had been identified as firesetters, and hence were at risk of lighting future fires. The paper also described the dissemination characteristics of the program as a guide to wider diffusion in the future. The study concluded that a better understanding of the Diffusion of Innovations theory components was necessary to close the

research-practice gap, particularly with regard to educating and engaging health professionals and service providers in community programs for injury prevention.

As noted by Meyer (2004), much of the initial work using Diffusion of Innovations theory was based on quantitative methodologies with consequent limitations. With the increasing recognition that both qualitative and mixed-methods approaches are needed to fully understanding injury prevention interventions and their implementation settings, application of the Diffusion of Innovations theory has much potential to contribute to injury prevention intervention studies in the future.

The RE-AIM Framework

The RE-AIM framework is a health promotion model with high applicability to injury prevention because it could underpin much implementation research (Finch 2009). The RE-AIM Framework was first proposed by Glasgow and colleagues (Glasgow et al. 1999) as a tool for evaluating the effectiveness of implemented programs with a large behavior change focus (Glasgow et al. 1999; Glasgow et al. 2003). It has since been used in a variety of program implementation contexts, most commonly focusing on individually targeted behavior change through exercise programs for people with arthritis (Gyurcsik and Brittain 2006), lifestyle interventions targeting cardiovascular disease risk factors (Besculides et al. 2008), other community-based behavioral interventions (Dzewaltowski et al. 2004), and knowledge translation systems in emergency departments (Bernstein et al. 2009). It has recently been advocated as a suitable model for the delivery and evaluation of sports injury prevention interventions (Finch 2009).

The RE-AIM Framework has a strong underpinning of health promotion theory and approaches (such as Diffusion of Innovations theory) and so is very relevant to the evaluation of injury prevention interventions, though the extent of its use is still in its infancy in this context. It draws from health promotion concepts, such as Diffusion of Innovations theory, that stress that desired health behaviors will only be achieved if the delivered interventions are available to the target group, adopted by them, and used as they were intended and that this use is sustained over time for ongoing prevention benefits. It, therefore, incorporates important aspects relating to individuals' responses and readiness in relation to targeted interventions, as well as the more public health-oriented benefits. In both its development and application, it has been shown to be highly robust and translatable across implementation settings (Glasgow et al. 1999; Glasgow et al. 2006; Jilcott et al. 2007). For any implementation study, understanding and representing the context in which the intervention is to be implemented and evaluated is a key component in its success. As will be shown in the injury examples below, the actual measures chosen within each of the framework dimensions can be set according to the specific contextual implementation feature of interest.

The RE-AIM framework has five key dimensions for assessing interventions that are useful for guiding thinking about the full complexities of the implementation context (Glasgow et al. 1999; Glasgow et al. 2003; Glasgow et al. 2006):

- Reach – the proportion (number) of the target population who are approached to take up the intervention and the representativeness of that group; this domain is relevant at the level of individuals.
- Effectiveness – the success rate if implemented as intended, as well as documentation of both positive and negative outcomes of the intervention. In some studies, this component has collected intervention efficacy, which may be more appropriate if this aspect of an intervention has not yet been developed. Outcomes here have most commonly been focused on individuals.
- Adoption – the proportion or number and representativeness of people, settings, practices, and plans that adopt the intervention. This dimension includes setting level factors.

- Implementation – the extent to which the intervention is implemented as intended in the real world. This dimension considered important factors associated with the delivery of the intervention within the setting of its application. The dimension is a setting-level assessment.
- Maintenance – the extent to which the intervention is sustained over time. This aspect is often categorized according to both individual-level and setting-level maintenance.

An underutilized strength of the RE-AIM framework is the capacity for all dimensions to be applied across all levels of the ecological framework for injury prevention and not just the levels initially proposed by the authors (Glasgow et al. 1999; Glasgow et al. 2003; Glasgow et al. 2006). This has recently been expanded upon in detail specifically for the sports injury prevention context (Finch and Donaldson 2010).

To date, seven published injury prevention studies have reported use of the RE-AIM Framework, all within the past 2 years. Two studies were within the context of falls prevention in older people (Li et al. 2008; Day et al. 2010), and five within sports injury prevention applications (Collard et al. 2010a; Finch and Donaldson 2010; Saunders et al. 2010; Finch et al. 2011a, b). In these injury studies, the RE-AIM framework has been used in several ways and these provide models for its application to other injury problems:

- As a model for undertaking and evaluating contextual influences on injury prevention in ecological systems (Finch and Donaldson 2010). This paper explains how RE-AIM could be used to understand the implementation impact of sports safety interventions that need to be implemented across several settings to be fully effective. In particular, it stresses that care needs to be taken when directly applying the RE-AIM framework to safety interventions implemented in the community sport setting because the definition for each dimension will depend on the specific level targeted. While many interventions will be targeted at only the individual sports participant, implementation of most sports injury interventions is multifaceted and complex and often needs to be targeted at multiple levels, as it will involve actions on the part of others such as coaches, sports administrators, peak sports bodies, etc. For this reason, a Sports Setting Matrix adaptation of RE-AIM was developed that outlined evaluation dimensions against each level of the sports safety delivery system.
- As study protocols (Day et al. 2010; Finch et al. 2011a), these two papers explain how the RE-AIM Framework has been used to design program delivery and evaluation plans from the outset. The context for the Day et al. (2010) protocol is the design of an evaluation plan to assess a large-scale system-wide prevention program for falls in older people. The second protocol is for the design and evaluation of a national program (including both intervention and delivery plan development and testing) to prevent football-related lower limb injuries (Finch et al. 2011). This study adopts the sports setting matrix adaptation of RE-AIM (Finch and Donaldson 2010).
- To inform the development of an intervention delivery plan for a larger-scale effectiveness RCT (Finch et al. 2011b). This study used both the RE-AIM framework and health belief model (Janz and Becker 1984) to identify the likely barriers and facilitators that would be experienced by football players if they were targeted by an exercise program to prevent lower limb injuries in their sport.
- As part of a process evaluation – the most common application. In an American county, a Tai Chi group exercise program to prevent falls in community dwelling older people was delivered through community health services (Li et al. 2008). The study had a major focus on the reach, adoption, and implementation RE-AIM dimensions which were all monitored at the end of the 12-week implementation period in the exercise participants. Effectiveness dimensions were analyzed in older people who participated in the Tai Chi program through identified changes determined through a pre-post test design. Adoption and maintenance dimensions were assessed at both the level of the exercise participant and the community health center. At 12 weeks, the study was too short to assess maintenance effects but some indicators of likely

drivers of that longer term uptake were also assessed. A study in Dutch school children reported the translatability and flexibility elements from a RE-AIM evaluation of a school-based program aimed at preventing physical activity-related injuries (Collard et al. 2010a). While there were some positive intervention effects, these were small and the RE-AIM evaluation was able to demonstrate that this most likely related to the intervention not being fully implemented as planned. The third study applied RE-AIM in interpreting coaches' feedback on the implementation of a safe landings program through targeted coach education sessions followed by coach delivery of the principles to their teams of junior netball players (Saunders et al. 2010). Evaluation against the RE-AIM dimensions enabled the authors to identify aspects of the intervention that could be improved to maximize future uptake and sustainability of the trialed intervention.

As shown from the above injury examples, the RE-AIM framework has been most commonly applied as an evaluation tool and that has been the case across other health issue applications. However, as other of the above injury application examples show, it has broader application as a planning tool and as a method to review intervention studies as is also promoted by its authors on the comprehensive RE-AIM website (see <http://www.re-aim.org/>).

There has been some criticism about the scientific application of RE-AIM in an analytical sense, and the rigor with which various dimensions have been measured and reported in published studies (Hoepsell et al. 2011). A recent extension to the CONSORT guidelines for randomized trial reporting has included some new aspects relating to the reporting of results from so-called pragmatic trials which are designed to inform decisions about practice changes as the result of interventions and these are relevant to RE-AIM type studies (Zwarenstein et al. 2008). Hoepsell et al. (2011) have recently outlined an epidemiological framework for reporting the public health impact from studies using RE-AIM that should also assist with the quality of studies in the future, particularly with regard to the reach component and external validity considerations.

Translation of the Findings from Intervention Research into Policy Initiative and Sustained Programs

A major goal of all injury research is to prevent injuries, so it is important that the research does not stop with producing effectiveness evidence. While this chapter has focused on only three theory-driven approaches, it is acknowledged that other approaches have been reported in the injury prevention literature. Often these have adopted similar components to those discussed above. For example, a systematic staged and evidence-informed approach toward identifying what might work to promote smoke alarm installation was conducted in the UK based on guidelines developed by the UK Health agency to translate research into policy (Brussoni et al. 2006). Injury prevention practitioner and policy-maker engagement was ensured through a participatory project that considered issues such as policy drivers and funding opportunities; multi-agency partnerships; program design considerations, targeting of interventions, and likely program implementation barriers and facilitators. Similarly, an evaluation of knowledge transfer of sports concussion education assessed this in terms of: the optimal target audience, what message should be delivered, who should deliver the message, how the educational message/s should be delivered and the impact of the knowledge transfer on professionals' knowledge, awareness, and attitudes (Provvidenza and Johnston 2009).

Translation research can be seen as an extension of intervention research in which investigations are undertaken into the processes for ensuring that the evidence is formally integrated into policy and practice is undertaken. There is an emerging body of literature about how such studies could be undertaken but it is beyond the scope of this chapter to discuss it in detail. However, the interested reader is referred to recent health promotion and health policy literature on this topic (Bowen and Zwi 2005; Buse et al. 2005; Choi et al. 2009; Morandi 2009). Some recent injury examples include studies that have explored engaging policy makers in road safety (Tran et al. 2008), falls prevention (Finch et al. 2009), and other injury prevention efforts (Mitton et al. 2008).

Importantly, multi-agency engagement of all major stakeholders from the outset would enhance the long-term success of intervention programs, particularly in terms of their sustainability through incorporation into formal policies and practices. Translation research would include the documenting and analysis of this process to develop an understanding of why, how, and when specific decisions were made. Specific questions that could be addressed in the translation research activities, drawing from the excellent discussion of these issues by Christoffel and Gallagher (2006), include:

- Which groups are most likely to benefit from (a) adoption of the specific injury intervention and/or (b) the evaluated intervention package, including delivery plan?
- What are the key components to delivering evidence-based injury prevention packages that could be used to inform state/national strategic approaches to implementing other safety or health promotion interventions in the community setting?
- What unique, but complementary, role could each stakeholder agency play in a future strategic approach to safety?

Researchers can participate in discussions with stakeholder agencies to identify potential roles in any future strategic approaches (which will be determined from results of previous phases). This process should be documented and analyzed to develop an understanding of why, how, and when decisions were made. Lessons learned from the intervention delivery in the implementation trial should be reviewed and the direct relevance to other sports identified through these researcher and stakeholder consultations. Policy makers, in particular, require good effectiveness evidence about interventions they are considering but this must include information about their likely translatability to other contexts, with varying characteristics (Finch et al. 2009).

Active engagement of the stakeholder groups through all aspects of the research will also increase the profile of, and acceptance of, injury prevention activities more generally (MacKay and Vincenten 2009). They will also generate background support for safety initiatives within their organizations, structures, and cultural groups that will translate to increased knowledge and awareness among a range of relevant consumers (Peterson et al. 2007). These activities will include fostering research into the translation of safety evidence and should include dissemination of information through specific scientific sessions at relevant research and practitioner conferences and industry forums convened by stakeholder groups; such forums plan and deliver sports safety and injury risk management advice for community delivery bodies and participants. Finally, researchers should work with stakeholder agencies to write and publish regular plain-language articles describing latest advances in safety targeted at their members as well as publishing their high-quality science in appropriate forums.

If the injury research community does not rise to this challenge our field will continue to suffer from the major information gap already identified by Nilson and Yorkston (2007) as a “critical need to understand the reasons why some community-based programs succeed and seemingly equivalent programs fail.” Moreover, increased and sustained efforts will be needed to make sure that the results of our intervention implementation are then successfully disseminated to those who will need to put them to use – both policy makers and injury practitioners.

References

- Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Beckmann & J. Kuhl (Eds.), *Action control: From cognition to behavior* (pp. 11–39). Berlin: Springer.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human*, 50, 179–211.
- Aldoory, L., & Bonzo, S. (2005). Using communication theory in injury prevention campaigns. *Injury Prevention*, 11, 260–263.
- Allegrante, J., Hanson, D., Sleet, D., et al. (2010). Ecological approaches to the prevention of unintentional injuries. *Italian Journal of Public Health*, 7, 24–31.
- Allegrante, J., Marks, R., & Hanson, D. (2006). Ecological models for the prevention and control of unintentional injury. In A. Gielen, D. Sleet, & R. DiClemente (Eds.), *Injury and violence prevention Behavioral science theories, methods, and applications* (pp. 105–126). San Francisco: Wiley.
- Andersen, P. A., Buller, D. B., Scott, M. D., et al. (2004). Prevalence and diffusion of helmet use at ski areas in Western North America in 2001–02. *Injury Prevention*, 10, 358–362.
- Armstrong, R., Clark, R., Murphy, S., et al. (2008). Strategies to support knowledge translation and exchange. *Australasian Epidemiologist*, 15, 24–27.
- Bandura, A. (1986). *Social foundations of thought and action: a social cognitive theory*. Reading, MA: Addison Wesley Publishing Company.
- Bartholomew, L., Parcel, G., & Kok, G. (1998). Intervention mapping: a process for developing theory- and evidence-based health education programs. *Health Education and Behavior*, 1998, 545–563.
- Bartholomew, L., Parcel, G., Kok, G., et al. (2006). *Planning health promotion programs. An intervention mapping approach*. San Francisco: Jossey-Bass.
- Bernstein, E., Topp, D., Shaw, E., et al. (2009). A preliminary report of knowledge translation: lessons learnt from taking screening and brief intervention techniques from the research setting into regional systems of care. *Academic Emergency Medicine*, 16, 1225–1233.
- Besculides, M., Zaveri, H., Hanson, C., et al. (2008). Best practices in implementing lifestyle interventions in the WISEWOMAN program: adaptable strategies for public health programs. *American Journal of Health Promotion*, 22, 322–328.
- Bierman, G. (2006). Commentary on the pitfalls and pratfalls of evaluation research with intervention and prevention programs. In R. Parker & C. Hudley (Eds.), *New directions for evaluation No 110* (pp. 87–96). San Francisco: Jossey-Bass.
- Bowen, S., & Zwi, A. B. (2005). Pathways to “evidence-informed” policy and practice: a framework for action. *PLoS Medicine*, 2, 200–205.
- Brosseau, L., Parker, D., Samant, Y., et al. (2007). Mapping safety interventions in metalworking shops. *Journal of Occupational and Environmental*, 49, 338–345.
- Brussoni, M., Towner, E., & Hayes, M. (2006). Evidence into practice: combining the art and science of injury prevention. *Injury Prevention*, 12, 373–377.
- Buse, K., Mays, N., & Walt, G. (2005). *Making health policy*. Berkshire, UK: Open University Press.
- Catford, J. (2009). Advancing the “science of delivery” of health promotion: not just the “science of discovery”. *Health Promotion International*, 24, 1–5.
- Choi, B., Gupta, A., & Ward, B. (2009). Good thinking: Six ways to bridge the gap between scientists and policy makers. *Journal of Epidemiology and Community Health*, 63, 179–180.
- Christoffel, T., Gallagher, S., (2006). Injury prevention and public health. Practical knowledge, skills, and strategies. Jones and Bartlett Publishers, Inc., Sudbury.
- Collard, D., Chinapaw, M., van Mechelen, W., et al. (2009a). Design of the iPlay study. Systematic development of a physical activity injury prevention program for primary school children. *Sports Medicine*, 29, 889–901.
- Collard, D., Chinapaw, M., Verhagen, E., et al. (2010a). Process evaluation of a school based physical activity related injury prevention program using the RE-AIM framework. *BMC Pediatrics*, 10, 86.
- Collard, D., Singh, A., & Verhagen, E. (2009b). The behavioral approach. In E. Verhagen & W. van Mechelen (Eds.), *Sports injury research* (pp. 157–166). Oxford: Oxford University Press.
- Collard, D., Verhagen, E., Chinapaw, M., et al. (2010b). Effectiveness of a school-based physical activity injury prevention program. A cluster randomized controlled trial. *Archives of Pediatrics and Adolescent Medicine*, 164, 145–150.
- Connor, J. (2004). Risk factor identification: The role of epidemiology. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control* (pp. 125–143). Melbourne: IP Communications.
- Day, L., Finch, C., Hill, K., et al. (2010). A protocol for evidence-based targeting and evaluation of statewide strategies for preventing falls among community-dwelling older people in Victoria, Australia. *Injury Prevention*. doi:10.1126/ip.2010.03775.

- De Vries, H., Backbier, E., Kok, G., et al. (1995). The impact of social influence in the context of attitude, self-efficacy, intention and previous behavior as predictors of smoking onset. *Journal of Applied Social Psychology*, 25, 237–257.
- Dzewaltowski, D., Estabrooks, P., Klesges, L., et al. (2004). Behavior change intervention research in community settings: how generalizable are the results? *Health Promotion International*, 19, 235–245.
- Eime, R., Finch, C., Wolfe, R., et al. (2005). The effectiveness of a squash eyewear promotion strategy. *British Journal of Sports Medicine*, 39, 681–685.
- Eime, R., Owen, N., Finch, C., (2004). Protective eyewear promotion: Applying principles of behavior change in the design of a squash injury prevention program. *34*, 629–638
- Farley, C., Haddad, S., & Brown, B. (1996). The effects of a 4-year program promoting bicycle helmet use among children in Quebec. *American Journal of Public Health*, 86, 46–51.
- Farley, C., Otis, J., & Benoit, M. (1997). Evaluation of a four year bicycle helmet promotion campaign in Quebec aimed at children aged 8 to 12 years: impact on attitudes, norms, and behaviours. *Canadian Journal of Public Health*, 88, 62–66.
- Finch, C. (2006). A new framework for research leading to sports injury prevention. *Journal of Science and Medicine in Sport*, 9, 3–9.
- Finch, C., (2009). Chapter 16: Implementing studies into real life. In: E. Verhagen, W. van Mechelen (Eds.), *Sports injury research* (pp. 213–235). Oxford University Press.
- Finch, C., Day, L., Donaldson, A., et al. (2009). Determining policy-relevant formats for the presentation of falls research evidence. *Health Policy*, 93, 207–213.
- Finch, C., & Donaldson, A. (2010). A sports setting matrix for understanding the implementation context for community sport. *British Journal of Sports Medicine*, 44, 973–978.
- Finch, C., Gabbe, B., Lloyd, D., et al. (2011a). Towards a national sports safety strategy – addressing facilitators and barriers towards safety guideline uptake (the NoGAPS project). *Injury Prevention*. doi:10.1136/ip.2010.031385.
- Finch, C., White, P., Twomey, D., et al. (2011b). Implementing an exercise training program to prevent lower limb injuries – considerations for the development of a randomized controlled trial intervention delivery plan. *British Journal of Sports Medicine*, 45, 791–795.
- Foster, M., Mitchell, R., & McClure, R. (2004). Making policy. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control* (pp. 267–282). Melbourne: IP Communications.
- Gielen, A., Sleet, D., DiClemente, C. C., et al. (2006a). *Injury prevention and violence prevention Behavioral science theories, methods, and applications*. San Francisco: Wiley.
- Gielen, A., Sleet, D., & Green, L. (2006b). Community models and approaches for interventions. In A. Gielen, D. Sleet, & R. DiClemente (Eds.), *Injury and violence prevention: Behavior change theories, methods and applications* (pp. 65–82). San Francisco, CA: Jossey-Bass.
- Glasgow, R. (2008). What types of evidence are most needed to advance behavioral medicine? *Annals of Behavioral Medicine*, 35, 19–25.
- Glasgow, R., Klesges, L., Dzewaltowski, D., et al. (2004). The future of health behavior change research: What is needed to improve translation of research into health promotion practice? *Annals of Behavioral Medicine*, 27, 3–12.
- Glasgow, R., Klesges, L., Dzewaltowski, D., et al. (2006). Evaluating the impact of health promotion programs: using the RE-AIM framework to form summary measures for decision making involving complex issues. *Health Education Research*, 21, 688–694.
- Glasgow RE, Lichtenstein E, Marcus AC (2003) Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *93*, 1261–1267
- Glasgow, R., Vogt, T., & Boles, S. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health*, 89, 1322–1327.
- Green, L., & Kreuter, M. (1991). *Health promotion planning: An educational and environmental approach* (2nd ed.). Mountain View: Mayfield Publishing Company.
- Gyuresik, N., Brittain, D., (2006). Partial examination of the public health impact of the People with Arthritis Can Exercise (PACE) Program: Reach, adoption, and maintenance. *Public Health Nursing*, 23, 516–522.
- Hanson, D., Hanson, J., Vardon, P., et al. (2005). The injury iceberg: An ecological approach to planning sustainable community safety interventions. *16*, 5–10
- Henderson, J., MacKay, S., & Peterson-Badall, M. (2006). Closing the research-practice gap: Factors affecting adoption and implementation of children's mental health program. *Journal of Clinical Child and Adolescent Psychology*, 35, 2–12.
- Hendriks, M., Bleijlevens, M., van Haastregt, J., et al. (2008). Lack of effectiveness of a multidisciplinary fall-prevention program in elderly people at risk: A randomized, controlled trial. *Journal of the American Geriatrics Society*, 56, 390–1397.
- Henley, N. (2004). Social marketing. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control* (pp. 318–333). Melbourne: IP Communications.

- Hingson, R., Howland, J., Koepsell, T. D., et al. (2001). Ecologic studies. In (Ed.), *Injury control: A guide to research and program evaluation* (pp. 157–167) Cambridge University Press.
- Hoepsell, T., Zatzick, D., & Rivara, F. (2011). Estimating the population impact of preventive interventions from randomized trials. *American Journal of Preventive Medicine, 40*, 191–198.
- Janz, N., & Becker, M. (1984). The health belief model: A decade later. *Health Education Quarterly, 11*, 1–47.
- Jilcott, S., Ammerman, A., Sommers, J., et al. (2007). Applying the RE-AIM framework to assess the public health impact of policy change. *Annals of Behavioral Medicine, 34*, 105–114.
- Kendrick, D., Groom, L., Stewart, J., et al. (2007). “Risk watch”: cluster randomized controlled trial evaluating an injury prevention program. *Injury Prevention, 13*, 93–99.
- Kilding, A. E., Tunstall, H., Kuzmic, D. (2008). Suitability of FIFA’s “The 11” training program for young football players-impact on physical performance. *Journal of Sport Science and Medicine, 7*, 320–326.
- Li, F., Harmer, P., Glasgow, R., et al. (2008). Translation of an effective tai chi intervention into a community-based falls-prevention program. *American Journal of Public Health, 98*, 1195–1198.
- Logghe, I., Verhagen, A., Rademaker, A., et al. (2011). Explaining the ineffectiveness of a Tai Chi fall prevention training for community-living older people: a process evaluation alongside a randomised clinical trial (RCT). *Archives of Gerontology and Geriatrics, 52*, 357–362. doi:10.1016/j.archger.2010.1005.1013.
- MacKay, J. M., & Vincenten, J. (2009). Why isn’t more injury prevention evidence-based? *International Journal of Injury Control and Safety Promotion, 16*, 89–96.
- Mallonee, S., Fowler, C., & Istre, G. R. (2006). Bridging the gap between research and practice: A continuing challenge. *Injury Prevention, 12*, 357–359.
- McGlashan, A., & Finch, C. (2010). The extent to which behavioral and social sciences theories and models are used in sport injury prevention research. *Sports Medicine, 40*, 841–858.
- Meyer, G. (2004). Diffusion methodology: Time to innovate? *Journal of Health Communication, 9*, 59–69.
- Mitton, C., MacNab, Y., Smith, N., et al. (2008). Transferring injury data to decision makers in British Columbia. *International Journal of Injury Control and Safety Promotion, 15*, 41–43.
- Morandi, L. (2009). Essential nexus: How to use research to inform and evaluate public policy. *American Journal of Preventive Medicine, 36*, S53.
- Murray, N., Kelder, S., Parcel, G., et al. (1998). Development of an intervention map for a parent education intervention to prevent violence among Hispanic middle school students. *Journal of School Health, 68*, 46–52.
- Murray, N., Kelder, S., Parcel, G., et al. (1999). Padres Trabajando por la Paz: A randomized trial of a parent education intervention to prevent violence among middle school students. *Health Education Review, 14*, 421–426.
- Nelson, G., & Moffit, P. (1988). Safety belt promotion: Theory and practice. *Accident; Analysis and Prevention, 20*, 27–38.
- Nilsen, P. (2004). What makes community based injury prevention work? In search of evidence of effectiveness. *Injury Prevention, 10*, 268–274.
- Nilsen, P., (2005). Evaluation of community-based injury prevention programs: Methodological issues and challenges. *International Journal of Injury Control and Safety Promotion, 12*, 143–156
- Nilsen, P., (2006). The theory of community based health and safety programs: A critical examination. *Injury Prevention, 12*, 140–145
- Nilsen, P., & Yorkston, E. (2007). Uncovering evidence on community-based injury prevention: A review of program effectiveness and factors influencing effectiveness. *International Journal of Injury Control Safety Promotion, 14*, 241–250.
- Nold, A., & Bochmann, F. (2010). Examples of evidence-based approaches in accident prevention. *Safety Science, 48*, 1044–1049.
- Peterson, J., Rogers, E., Cunningham-Sabo, L., et al. (2007). A framework for research utilization applied to seven case studies. *American Journal of Preventive Medicine, 33*, S21–S34.
- Pitt, R., & Spinks, D. (2004). Advocacy. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control* (pp. 303–317). Melbourne: IP Communications.
- Prochaska, J., Evers, K., Prochaska, J., et al. (2007). Efficacy and effectiveness trials: examples from smoking cessation and bullying prevention. *Journal of Health Psychology, 12*, 170–177.
- Prochaska, J. O., Redding, C. A., & Evers, K. E. (1997). Chapter 4 the transtheoretical model and stages of change. In K. Glanz, F. M. Lewis, & B. K. Rimer (Eds.), *Health behavior and health education* (pp. 60–84). San Francisco: Jossey-Bass Inc.
- Providenza, C. F., & Johnston, K. M. (2009). Knowledge transfer principles as applied to sport concussion education. *British Journal of Sports Medicine, 43*(Suppl 1), i68–i75.
- Rivara, F. (2008). Evaluating the effect of an injury prevention intervention in a population. *American Journal of Preventive Medicine, 34*, S148–S152.
- Robertson, L. (2007). *Injury epidemiology. Research and control strategies*. New York: Oxford University Press.

- Roen, K., Arai, L., Roberts, H., et al. (2006). Extending systematic reviews to include evidence on implementation: Methodological work on a review of community-based initiatives to prevent injuries. *Social Science and Medicine*, *63*, 1060–1071.
- Rogers, E. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.
- Rogers, E. (2004). A prospective and retrospective look at the diffusion model. *Journal of Health Communication*, *9*, 3–19.
- Saunders, N., Otago, L., Romiti, M., et al. (2010). Coaches' perspectives on implementing an evidence-informed injury prevention program in junior community netball. *British Journal of Sports Medicine*, *44*, 1128–1132.
- Sawyer, R., Hamdallah, M., White, D., et al. (2010). High school coaches' assessments, intentions to use, and use of a concussion prevention toolkit: Centres for Disease Control and Prevention's heads up, concussion in high school sports. *Health Promotion Practice*, *11*, 34–43.
- Scott, I. (2004). Laws and rule-making. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control* (pp. 283–302). Melbourne: IP Communications Pty Ltd.
- Sleet, D., & Gielen, A. (2004). Developing injury interventions: the role of behavioral science. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control*. Melbourne: IP Communications.
- Soligard, T., Myklebust, G., Steffen, K., et al. (2008). Comprehensive warm-up programme to prevent injuries in young female footballers: Cluster randomized controlled trial. *British Medical Journal*, *337*, a2469.
- Stame, N. (2010). What doesn't work? Three failures, many answers. *Evaluation*, *16*, 371–387.
- Steffen, K., Myklebust, G., Olsen, O., et al. (2008). Preventing injuries in female youth football – a cluster-randomized controlled trial. *Scandinavian Journal of Medicine and Science in Sports*, *18*, 605–614.
- Thompson, R., & Sacks, J. (2001). Evaluating an injury intervention or program. In F. Rivara, P. Cummings, T. Koepsell, et al. (Eds.), *Injury control: A guide to research and program evaluation* (pp. 196–216). Cambridge, UK: Cambridge University Press.
- Timpka, T., Ekstrand, J., & Svanstrom, L. (2006). From sports injury prevention to safety promotion in sports. *Sports Medicine*, *36*, 733–745.
- Tran, N., Hyder, A., Kulanthayan, S., et al. (2008). Engaging policy makers in road safety research in Malaysia: A theoretical and contextual analysis. *Health Policy*, *90*, 58–65.
- Trifiletti, L., Gielen, A., Sleet, D., et al. (2005). Behavioral and social sciences theories and models: Are they used in unintentional injury prevention research? *Health Education Research*, *20*, 298–307.
- van Tiggelen, D., Wickes, S., Stevens, V., et al. (2008). Effective prevention of sports injuries: A model integrating efficacy, efficiency, compliance and risk taking behavior. *British Journal of Sports Medicine*, *42*, 648–652.
- Ward, V. L., House, A. O., & Hamer, S. (2009). Knowledge brokering: Exploring the process of transferring knowledge into action. *BMC Health Services Research*. doi:10.1186/1472-6963-9-12.
- Winston, F., & Jacobsen, L. (2010). A practical approach for applying best practices in behavioral interventions to injury prevention. *Injury Prevention*, *16*, 107–112.
- Zwarenstein, M., Treweek, S., Gagnier, J., et al. (2008). Improving the reporting of pragmatic trials: an extension of the CONSORT statement. *British Medical Journal*, *337*, a2390.

Chapter 36

Economic Evaluation of Interventions

Ted R. Miller and Delia Hendrie

Introduction

Setting priorities and allocating scarce resources among alternative uses always raise difficult choices. Injury prevention and control programs compete for funds with programs directed at such diverse topics as disease, road building, and property crime. Different kinds of injury and different prevention approaches also compete for funds from budgets earmarked for injury prevention.

In economic evaluation, the fundamental question is which interventions provide the best value for money compared with other interventions that could be provided with the same resources. Regardless of the sector to which it is applied, economic evaluation is characterized by two features (Drummond et al. 2005):

- It deals with both the costs and outcomes of interventions.
- It is about choices. Resource scarcity and the inability to fund all possible interventions mean that choices must be made. Economic evaluation is one tool that is available to assist decision makers to spend their money wisely when selecting from a range of alternative options.

These two characteristics lead to economic evaluation being defined as the comparative analysis of alternative courses of action in terms of both their costs and consequences. The tasks in economic evaluation are to identify, measure, value, and compare the costs and benefits of the options being considered. For example, a highway department annually upgrades the safety of numerous roads. It straightens curves, installs speed bumps and breakaway poles/signs, adds signals, replaces bridges, and clears the roadside of hazards. The list of desirable improvements always exceeds the budget available. To allocate its resources, the department needs to consider what it can afford and how to maximize the reductions in crashes, deaths, and injuries within that budget and politically imposed priorities. In contrast to economic evaluation, cost-of-injury studies identify and measure all costs of

T.R. Miller, PhD (✉)

Center for Public Health Improvement and Innovation, Pacific Institute for Research and Evaluation,
11720 Beltsville Drive, Suite 900, Calverton, MD 20705, USA
e-mail: miller@pire.org

D. Hendrie, MA

Population Health Research, Curtin Health Innovation Research Institute (CHIRI),
Curtin University, GPO Box U1987, Perth, WA 6845, Australia
e-mail: d.v.hendrie@curtin.edu.au

injury, including the direct, indirect, and intangible dimensions. The output, expressed in monetary terms, is an estimate of the economic burden of injury (see Chap. 19).

This chapter provides an introduction to the economic evaluation of injury prevention and control programs and its relevance to decision making. Section “What Represents Value for Money?” explains the purpose of economic evaluation. Section “Conducting an Economic Evaluation Study” describes the different types of economic evaluation and the measurement of costs and outcomes; section “Computing the Economic Outcome Estimates” adds details on benefits valuation and computation of the economic analysis measures. Guidelines for reporting an economic analysis appear in section “Reporting the Results of an Economic Evaluation.” Section “Incorporating Economic Evaluation in Decision Making” addresses the interpretation of economic analysis findings and the integration of economic analysis into the broader decision-making process. Section “Conclusion” concludes the chapter with some thoughts on the value of economic evaluations in the field of injury prevention and control.

All monetary values presented in the chapter are in 2009 US dollars unless otherwise indicated. Values for other years were converted to 2009 prices using country-specific price indexes. Local currency costs were converted to US dollars using purchasing power parity exchange rates.

What Represents Value for Money?

In economics, the concept of value for money is known as efficiency. Given budgetary challenges, attention must be paid to spending limited budgets on interventions that are comparatively good (i.e., efficient) at reducing deaths and injuries relative to their costs and away from those that require large expenditures to achieve comparatively low reductions. By comparing the resources used by an intervention and the benefits generated by that intervention, economic evaluation provides insight into how injury prevention resources can be allocated so as to maximize overall benefits to the community (Hendrie and Miller 2004).

The concept of value for money can be illustrated using a cost–effectiveness plane, which is a graphical method for comparing the cost–effectiveness of two or more interventions (Glick et al. 2007). The horizontal axis by convention measures differences in effectiveness, and the vertical axis measures differences in costs. In comparing two interventions, differences in costs and effect can fall into four quadrants (Fig. 36.1). Clearly, one would be allocating resources wisely by funding an injury prevention program that costs less than an alternative one and was more effective than it (lower right quadrant). In this case, the more cost–effective intervention is referred to as an economically “dominant” strategy. The opposite is a “dominated” strategy, and one generally would not allocate resources to an intervention that was less effective and cost more (upper left quadrant). The decision is more difficult when faced with the scenario of an intervention improving effectiveness at increased cost (upper right quadrant). This is a common situation. For example, consider a decision about whether to implement a falls prevention program for the low-risk elderly or maintain the status quo of no intervention. Implementing a new program will require resources such as staff, supplies and equipment, office accommodation, and the like. If effective, the program will reduce the number of fall injuries. In making a decision whether or not to fund the program, the question becomes whether the gains in effectiveness warrant the additional costs of funding the program. In the case of the lower left quadrant, an intervention is less effective than an alternative one but reduces costs. Here, the question is how the cost savings compare with the loss of effectiveness. In reality, however, people often are reluctant to consider a new intervention unless it has the prospect of being more effective than existing interventions (whatever it costs). They hesitate to make their lives less safe to save money.

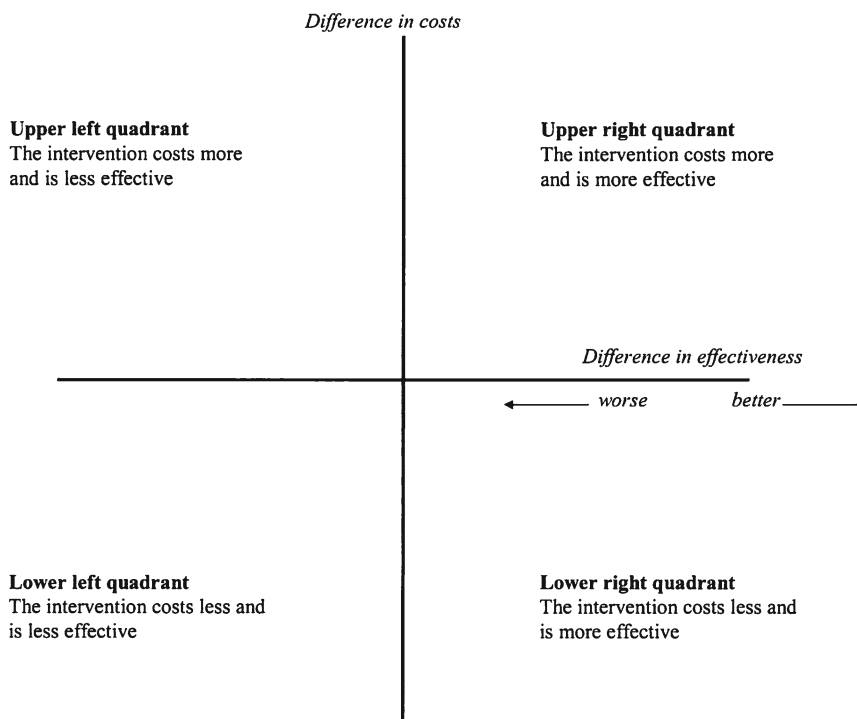


Fig. 36.1 The cost–effectiveness plane

Conducting an Economic Evaluation Study

An economic evaluation study comprises several interrelated elements. The study question must be specified, a research strategy must be developed, costs must be calculated, the effectiveness of the alternatives must be determined, and cost–outcome measures must be computed. Within each of these components, several steps can be identified.

Framing the Study

Objectives of the Study

The first step in conducting a study is to establish the objective. This involves defining the problem, the hypothesis to be tested, the interventions to be compared, and the viewpoint for the analysis.

Identifying the study problem is a crucial starting point, and several criteria must be considered. The problem must have economic importance in relation to resource utilization, and the study must have the potential to improve outcomes, reduce costs, or both. The options being compared must be relevant to the choices facing decision makers, and characteristics of these options must be well defined. The types of characteristics that need to be defined will depend on the analysis but include the nature of the interventions, the target population, the setting or delivery site, and the time period over which the costs and effectiveness will be evaluated.

Table 36.1 Types of economic evaluation studies

Type of study	Outcome measures	Evaluation question
Cost-minimization	Not measured as known to be the same	Least cost comparison of interventions with the same outcomes
Cost-effectiveness	Natural units (e.g., fatalities prevented, life-years saved, injuries prevented, burns prevented)	Comparison of interventions with same objective
Cost-utility	Utility (e.g., quality-adjusted life years (QALYs))	Comparison of interventions with same or different objectives
Cost-benefit	Money	Are the benefits worth the costs? How large are the benefits net of the costs?

The viewpoint for the analysis refers to whose perspective the study adopts. The viewpoint can be the employer (for an occupational safety intervention), the government, a specific level of government (national, State), the hospital, the health-care system, or society as a whole. The choice of viewpoint for an economic evaluation is determined by the question to be answered.

In setting the objectives of the study, an important factor is the decision-making context, in other words, who will use the information and how. These factors will define the information that needs to be collected and the perspective of the study, which in turn will affect many other decisions relating to the design of the study.

Types of Analysis

There are four main types of economic evaluation (Table 36.1). The identification of costs and their measurement in monetary values is similar across each type. The difference lies in how outcomes are measured and valued (see below). When to use each type of economic evaluation will depend on the nature of the question being addressed and the purpose for undertaking the analysis.

Cost-Minimization Analysis

Cost-minimization analysis takes into account only the costs of alternative options and is an appropriate type of analysis to use when the outcomes of the interventions being considered are known to be the same (e.g., domiciliary versus center-based community rehabilitation for falls patients). If evidence is available that each program produces comparable outcomes, then the decision regarding the most appropriate intervention can concentrate on finding the least costly option.

Cost-Effectiveness Analysis

While cost-minimization analysis is a useful technique for comparing programs with the same outcomes, few interventions are equally effective. Cost-effectiveness analysis is the most straightforward type of economic evaluation to take account of differences in outcomes. In cost-effectiveness analysis, outcomes are measured in naturally occurring units, which can be generic units (e.g., life years saved) that can be compared across all prevention programs that save lives or more specific units (e.g., reduction in the number of traumatic brain injuries or the number of assaults) that can be used only to compare interventions with the same objective.

Cost–Utility Analysis

Information from cost–effectiveness analyses is useful in clarifying choices between different interventions on the basis of an outcome measure that is unidimensional. However, a limitation of cost–effectiveness analysis is if comparisons need to be made between interventions to prevent different types of injuries where the outcome measure varies across the alternative options. For example, if the effectiveness of school playground resurfacing is measured in terms of the number of fractures avoided per year and the effectiveness of a school-based bicycle helmet distribution program is measured in terms of the number of traumatic brain injuries avoided, then the cost–effectiveness of the programs cannot be compared as the outcome units are different. On the other hand, if outcomes in cost–effectiveness analysis are being measured using a generic unit such as life years saved, then the outcome measure is only appropriate for treatments that are potentially life-saving, and even then the cost–effectiveness analysis will not take into account any differences in functional capacity that result from the interventions. This makes cost–effectiveness analysis inappropriate for comparing programs like helmet distribution that is primarily lifesaving with ones like playground resurfacing that primarily prevent nonfatal injuries and improve quality of life.

Cost–utility analysis is a form of cost–effectiveness analysis in which the outcome measure is utility, which is an economic term that relates to a person’s well-being. The most commonly used unit of utility is a quality-adjusted life year (QALY), which is a multidimensional concept measuring the physical, emotional, mental health, and social aspects that are relevant and important to a person’s well-being. Since outcomes are measured in commensurate units such as QALYs, comparisons in cost–utility analysis can span diverse interventions for different types of injuries. An alternative metric that is commonly used is the disability-adjusted life year (DALY), which can be thought of as one lost year of “healthy” life.

Cost–Benefit Analysis

In cost–benefit analysis, the costs and outcomes of interventions are both measured in monetary units. Cost–benefit analysis is similar to cost–utility analysis inasmuch as the effects of the alternatives are measured in commensurate units, except outcomes in cost–benefit analysis are measured in terms of money rather than utility. This allows for direct comparison across diverse interventions with different objectives. Also, as is the case with cost–utility analysis, multiple benefits can be captured in cost–benefit analysis if the interventions under consideration produce multidimensional outcomes. An additional advantage of cost–benefit analysis is that it provides an answer as to whether an intervention is worthwhile implementing. Any intervention where the benefits are greater than the costs is worthwhile. When comparing two alternatives, the intervention with the greatest net benefit or highest benefit–cost ratio (BCR) is the preferred option.

Table 36.2 illustrates the various elements of the research design for an economic evaluation of hip protectors for the elderly. Table 36.3 shows a study abstract that includes a variety of cost–outcome measures.

Estimating Costs in Economic Evaluation

Key Concepts

The Concept of “Cost” in Economic Evaluation

Safety programs often involve out-of-pocket expenses for staff, accommodation, equipment, and protective gear. When paid staff are assigned to work on a safety program, the program costs include

Table 36.2 Framing a study – an example of hip protectors for the elderly

Use of hip protectors in residential aged care facilities	
<i>Objective of the economic evaluation</i>	
Study problem	Falls in the elderly
Nature of intervention	Hip protectors
Comparator	Education for nurses who subsequently educated residents
Setting	Usual care
Target population	Residential aged care facilities
Study time frame	Residents with a high risk of falling
Perspective	18 months
Decision making context	Health sector
<i>Type of analysis</i>	Health policy makers and health insurers
Economic study type	Cost–effectiveness analysis
Outcome measure	Hip fractures prevented

Source: adapted from Meyer et al. (2005)

Table 36.3 Study abstract illustrating a range of cost–outcome measures and perspectives

The adoption of a Child Restraint System disbursement/education program could prevent up to 2 deaths, 12 serious injuries, and 51 minor injuries per 100,000 low-income children annually. When fully implemented, the program could save Medicaid over \$1 million per 100,000 children in direct medical costs while costing \$13 per child per year after all 8 years of benefit. From the perspective of Medicaid, the program would cost \$17,000 per life year saved, \$60,000 per serious injury prevented, and \$560,000 per death averted. The program would be cost saving from a societal perspective. These data are similar to published vaccination cost–effectiveness data

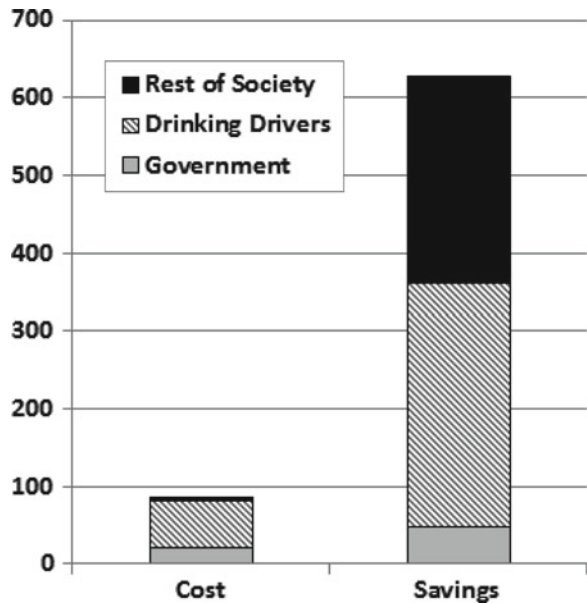
Source: Goldstein et al. (2008)

their wages, fringe benefits, and associated overhead expenses. That is true even if they are on salary and would have been paid anyway. Working on the safety program rather than a nutrition program or building a bicycle trail represents an opportunity cost because the community now lacks staff time to implement another worthwhile program.

When people volunteer their time, that time also represents a cost. Economists disagree about the best way to value volunteer's time. The preferred approach is to use the amount one would have to pay someone to do the work. Economists call this the *replacement cost*. The alternative is to use the amount the person could have earned by working instead of volunteering, which economists call the *opportunity cost*. A cost–effectiveness evaluation of bicycle helmet programs (Hatziaandreu et al. 1995) compared the cost per helmet worn of a community education campaign with price discounts, a motivational speaker at a school assembly, and passing and enforcing a rule requiring helmet use when bicycling. The evaluation priced the volunteer time using opportunity cost, and the motivational speaker was a professional athlete. The evaluation valued the athlete's time at half of his hourly salary when on the field. The time of volunteers in the community program was priced at a much lower value. Not surprisingly, the motivational approach was prohibitively more costly than the other approaches. A police officer, however, might have been an equally effective motivator at a fraction of the price. It makes no sense to price the time of a parent stuffing envelopes at her salary as a corporate executive. Fringe benefits generally are not added when valuing volunteer time.

A related issue arises when a program like conflict resolution or family therapy requires parents to transport their children to or attend sessions with their children. The travel costs are a program cost. Generally, the time of the children is not valued. Handling of the parental time is less clear. One credible approach is to state that it was not valued.

Fig. 36.2 Costs and cost savings from a sustained compulsory breath testing program in New Zealand by perspective (in millions of 2009 US dollars) (source: Miller and Blewden 2001)



The Impact of Alternative Perspectives

The choice of the perspective has an important effect on the cost components used in the evaluation. If the perspective is that of society, all costs of an intervention or injury count. If the perspective is narrower, say that of the government, then one would not value the loss of freedom, reduced mobility, and reduced enjoyment of intoxication that resulted from lowering the maximum allowable driver blood alcohol limit. Only the government's costs of passing, publicizing, implementing, and enforcing the law would be included. Similarly, from government's perspective, the cost savings of an unintentional injury prevention intervention include only emergency services, government paid health care, taxes lost if the injured person loses earnings, and safety net payments (welfare, housing assistance) if the injury results in a fall into financial poverty. Earnings loss borne by the injured and their employers do not count, nor do the out-of-pocket expenses, pain, and lost quality of life of the injured and their friends and families.

Another perspective is that of the health-care system. This perspective might help to sell the system on brief motivational interventions in the emergency department addressing suicidality or alcohol use, on regionalizing trauma care, or on a hospital-funded community safety initiative. The only relevant costs are the health-care system's costs to plan and operate the intervention. The only relevant cost savings are offsetting reductions in medical care costs. For example, in the USA, the average exposure call to a poison control center costs \$43 and prevents \$330 of unnecessary hospital services use (Zaloshnja et al. 2008).

When setting the stage for public policy intervention, an external cost perspective is of interest. This perspective evaluates the costs and cost saving benefits exclusive of costs and savings to the person whose behavior will be regulated. For example, a good reason to force motorcyclists to wear helmets is that other people pay \$340 extra per year in medical and wage replacement costs if a rider does not wear a helmet (Miller 1994).

Often one displays costs and savings from multiple perspectives as in Fig. 36.2. Showing savings from multiple perspectives helps to provide a more complete representation of the costs and benefits of a program.

Steps in Costing

Cost analysis involves identifying the range of resources used in implementing an injury prevention program and measuring and valuing these resources.

Identifying Resource Use

The first step in cost analysis is to identify the resource inputs used for the intervention. Intervention costs fall into five categories: (1) staff, (2) capital equipment, (3) disposables, (4) follow-up costs, and (5) impacts on the public. For example, a roadside sobriety check program uses police personnel to plan and staff the checkpoint, police vehicles, traffic cones, breathalyzers, small tokens of appreciation handed out to sober drivers, follow-up adjudication, and sanctioning for offenders. Beyond resource inputs, the checkpoints impose costs on the public by delaying their travel and intruding slightly on their privacy by forcing them to submit to testing.

In general, passage and enforcement of laws governing adult behavior impose intangible and hard-to-measure costs, but ones of great concern to some politicians. Laws interfere with personal freedom and privacy. They can cause discomfort (e.g., a helmet law), inconvenience, loss of the enjoyment of one more drink, loss of mobility, or loss of time. Road-building agencies generally prescribe how to value travel time, and mobility can be valued at the cost per kilometer of operating a motor vehicle. Many of the other costs are challenging to value in monetary units (e.g., dollars or francs).

When a new law restricts personal choices, the enjoyment foregone is a cost. Enforcing an existing law, however, generally is not considered to impose that cost because the offender lacks standing (Trumbull 1990). The legislature already has decided, for example, that the costs of preventing an assault do not include the brawler's loss of satisfaction from hitting someone.

The viewpoint of the economic evaluation will guide the range of costs to include in the study. If the results of the economic evaluation will be used only to compare the interventions under study, there is no need to estimate costs common to both because they will not affect the choice between the interventions.

If the relative order of magnitude of some cost elements is small, they are unlikely to make a difference in the results of the economic evaluation, so a substantial effort to measure them may be inadvisable. However, some justification should be given for not considering such costs or providing only a rough estimate.

Measuring and Valuing the Resources

Valuing the staff time, capital equipment, and disposables requires tracking how much of each input is used and attaching a unit cost to each input identified. Costs are typically valued in units of local currency, based on prevailing prices of, for example, staff time, bicycle helmets, smoke alarm batteries, educational materials, or media coverage. These unit prices can be obtained directly from budgets for the intervention or alternatively are often available from vendors or increasingly from web searches.

The objective in valuing costs is to obtain an estimate of the opportunity cost or worth of the resources used. Thus, it may be necessary to impute some unit costs (e.g., for volunteer time) or to make adjustments to the costs shown in budgets (e.g., for subsidized services). The costs usually should include costs to implement and monitor the effectiveness of the program in replication or scale-up, but not the costs to pioneer and evaluate the original program.

Program Costs Versus Cost Savings

The costs of a program include the value of all resources used to implement an intervention. If effective, an outcome of the intervention will be the reduction in one or several measures of injury such as burns prevented or life years saved in the case of a burns prevention program. Another benefit is the saving of resources such as medical care used in treating an injured person or rehabilitation aids and appliances required to assist in performing tasks following injury. These savings mirror the costs of implementing a program and are measured and valued in a similar way (Drummond et al. 2005).

Measuring Outcomes in Economic Evaluation Studies

Injury prevention and control reduces injury and associated death and disability. The economic analyst needs to list the likely intervention outcomes and decide which to include in the analysis. If the analysis is linked to an evaluation study, the outcomes measured in the evaluation will dictate the possible choices. For example, reducing the maximum blood alcohol levels for drivers may cause some people to drink less, resulting in less risky sex and fewer falls. It also may shift drinking from bars to homes, resulting in fewer barroom brawls but more domestic violence. If the evaluation only measures the effect on impaired driving, however, the economic analysis cannot value these other possible outcomes. The exclusion becomes a limitation of the analysis. Parts III and IV in this book describe how to identify intervention outcomes and evaluate their statistical significance.

A broader approach to outcome estimation is to perform a systematic review or meta-analysis of existing effectiveness estimates. Generally, such an evaluation should consider the quality of the studies as well as the mean estimates and their variance. The Guide to Community Preventive Services, *Cochrane Reviews*, Elvik et al. (2009), and, to a lesser extent, the Child Injury Prevention Tool (<http://childinjuryprevention.org/>) all provide meta-analytic or systematic review data that may be helpful.

Sometimes the economic analysis will be prospective. For example, Lestina et al. (2002) analyzed the economic feasibility of ultraviolet (UV) headlights and associated UV striping of roads and use of UV paint on bicycles. They estimated the costs of the different kinds of crashes that UV would reduce. Based on experimentation, they arrived at a possible range of effectiveness. They estimated the return on investment across that range and also computed break-even effectiveness – the minimum effectiveness needed for UV’s benefits to exceed its costs.

Computing the Economic Outcome Estimates

Cost–Effectiveness Analysis

Cost–effectiveness analysis (CEA) is the most straightforward type of economic evaluation to take account of differences in outcomes. It compares programs or strategies in terms of their cost per unit of outcome and measures outcomes in naturally occurring units such as life years saved, the number of injuries prevented, the number of fractures prevented, and so on.

Outcomes in cost–effectiveness analysis can be expressed in measures that represent (1) process outcomes that are known or believed to have a direct bearing on health gains, (2) intermediate outcomes that lie along the pathway to health gains, or (3) health gains. The simplest formula for a cost–effectiveness ratio (CER) is simply

$$\text{CER} = \frac{\text{intervention cost}}{\text{intervention outcome}}. \quad (36.1)$$

In injury prevention, CEAs work best for comparing alternative ways to achieve an intermediate outcome like cost per new safety belt user, cost per smoke alarm with working batteries, or cost per youth trained in conflict resolution. Process outcomes like cost per suicidal person referred to treatment or cost per person advised to put a child in a child seat are less certain to produce outcomes. Some people will not comply with the advice. Worse, the level of noncompliance may differ among the alternatives.

CEAs focused on a final outcome do not work very well in injury prevention for three reasons. First, interventions tend to prevent both fatal and nonfatal events. So, cost per life saved or life year saved misses many of prevention's benefits. Second, protective devices tend to convert deaths to nonfatal injuries and reduce the severity of nonfatal injuries that would have occurred without the device. Cost per injury prevented ignores those severity reductions. Third, depending on where force impacts the body during an injury event, a wide range of injuries can result. The CER in (36.1) gives the same value to a shattered skull and a fractured toe.

A more sophisticated CER has evolved that partially addresses these problems by valuing as many benefits as possible in monetary units and leaving one unmeasured aspect of the benefits – typically a final outcome – unquantified. The quantified benefits are subtracted from the costs when computing the CER. So, the improved CER formula is

$$\text{CER} = \frac{\text{intervention cost} - \text{savings in monetary units}}{\text{intervention outcome}}. \quad (36.2)$$

This version of CER requires knowing the effects of the intervention on a final outcome or on an intermediate outcome with known effects on a final outcome. For example, the medical cost savings associated with wearing a helmet when riding a motorcycle have been well documented (Lawrence et al. 2003).

Estimating Injury Cost Savings

Evaluating the benefits that result from the prevention of injury requires determining the cost of the injuries that will be prevented. The benefits are the cost savings.

The costs of an injury have many dimensions. They include a range of short-term and long-term resource costs – for police, fire, emergency medical, coroner, acute medical care, follow-up medical care, mental health care, public assistance, and funeral services, as well as processing of insurance claims and public assistance payments. Employers incur costs because employees are distracted by talk about the injury of a coworker or family member and because the supervisor of a worker who is killed or disabled must shuffle schedules, possibly hire a replacement, and compensate for the loss of specialized skills and knowledge. Friends and relatives of those injured and killed lose wages and incur travel costs. Injury incidents often also involve property damage, scene cleanup, incident investigation, and liability litigation. Motor vehicle crashes delay traffic, and crimes and crashes often result in adjudication and sanctioning costs.

In addition to the resource costs, injuries have opportunity costs. They prevent people from earning wages and doing housework. They also have intangible effects including pain, suffering, scarring, and functional losses that reduce quality of life.

When evaluating a preventive intervention, it rarely is possible, much less efficient, to track the costs of relevant injuries in a control group and compare those to the costs in a treatment group. Cost

$$\begin{aligned} \text{Incremental cost effectiveness} &= \frac{[\text{cost of more intensive program} - \text{cost of less intensive program}]}{[\text{benefits of more intensive program} - \text{benefits of less intensive program}]} \\ &= \frac{\text{difference in costs}}{\text{difference in benefits}} \end{aligned}$$

Fig. 36.3 Incremental cost–effectiveness ratio of switching to a more intensive program from a less intensive program

tracking requires following the seriously injured for many years and accessing a variety of record systems that rarely are open to the public. A popular and practical alternative in many countries is to access existing national injury cost estimates and adjust them to local prices.

Users of the second CER equation typically take that approach. The costs they subtract are the resource costs. The uncoded portions of the injury savings then are associated with the quality of work life preserved.

Incremental Analysis

When undertaking an economic evaluation, either alternative safety interventions are being compared in terms of their costs and benefits or an intervention is being compared to the alternative of doing nothing (i.e., the “do nothing” option). For example, should we straighten a hazardous curve, put up a sign warning that a sharp curve is coming, or simply leave the road alone?

It is important to distinguish between average and incremental cost effectiveness. An average CER is estimated by dividing the cost of the intervention by a measure of effectiveness without regard to its alternatives. An incremental cost–effectiveness ratio (ICER) is an estimate of the cost per unit of effectiveness of switching from one intervention to another. In estimating an ICER, the numerator and denominator of the ratios represent the differences in costs and outcomes, respectively, between the alternative interventions (Fig. 36.3 shows an incremental cost–effectiveness formula from a “more intensive program” to a “less intensive program”). For example, a graduated or provisional licensing program for drivers aged 15–17 with a midnight curfew costs \$84 per driver because most youths this age are home by midnight anyway (Miller et al. 2012). Switching to a 10 p.m. curfew adds \$164 per driver to the cost because it restricts youth mobility and forces parents to provide transportation. It also adds \$26 in medical cost savings and reduces quality-adjusted life year (QALY) loss by 0.0035, yielding a cost/QALY saved of \$39,430 $((164-26)/0.0035)$. It is not necessary to calculate a CER if one intervention dominates another (i.e., costs less and has better outcomes).

Cost–Utility Analysis

Utility-based outcome measures compare the outcomes of an injury prevention intervention with its alternative by valuing the gains it delivers as measured in health-related quality of life. A cost–utility analysis is a CEA where the quality of life and work losses are quantified using a standardized utility-based measure called a quality-adjusted life year (QALY). A QALY is a health outcome measure that shows how people value health states. The value is measured on a scale where perfect health is valued at 1.0 and death is valued at 0.0. Many people assign health states like an enduring coma values less than 0.0, making them fates worse than death (Miller et al. 1995).

Gains in health-related quality of life are calculated based on two factors: the gain in QALYs and the number of life years over which the gain is sustained. For example, the average quality of life

Fig. 36.4 The EQ-5D classification system

<p>Mobility I have no problems in walking about I have some problems in walking about I am confined to bed</p> <p>Self-Care I have no problems with self-care I have some problems washing and dressing myself I am unable to wash or dress myself</p> <p>Usual activities (e.g., work, study, housework, family or leisure activities) I have no problems with performing my usual activities I have some problems with performing my usual activities I am unable to perform my usual activities</p> <p>Pain/Discomfort I have no pain/discomfort I have moderate pain/discomfort I have extreme pain/discomfort</p> <p>Anxiety/Depression I am not anxious or depressed I am moderately anxious or depressed I am extremely anxious or depressed</p>

loss to a hospital-admitted concussion without skull fracture is 0.764 in the first year post-injury, 0.226 in years 2–5, and 0.068 thereafter (Miller et al. 1995). A person aged 40 in the USA has an expected remaining lifespan of 40 years. So, without discounting future losses, the QALYs lost to a concussion at age 40 would be equal to 4.048 ($0.764 + 0.226 \times 4 + 0.068 \times 35$).

The cost–utility analysis (CUA) measure is computed as

$$\text{CUA} = \frac{\text{intervention cost} - \text{resource cost savings}}{\text{QALYs saved}}. \quad (36.3)$$

QALY Estimation

Several generic quality of life instruments have been developed to use in measuring quality of life across different conditions. These instruments have been tested for validity and reliability. They provide a profile of scores in different health domains.

For example, the EQ-5D simplifies health into five domains: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (Fig. 36.4). Each domain is given a score from 1 to 3, so the health profile would read 11111 for the best scores in all domains and 33333 for the worst. The EQ-5D has 243 possible health profiles (i.e., $3 \times 3 \times 3 \times 3 \times 3$), all of which have been assigned a utility value between 0 and 1 by general population surveys. In addition, a utility loss premium generally is added if any domain is at its worst level.

More than 25 algorithms from around the world exist for converting the five responses to a quality of life score (EuroQol Group 2007). Other widely used quality of life instruments include the Health Utilities Index (Drummond et al. 2005), numerous utility-scored versions of the Short-Form 12 (SF-12, Ware et al. 2002), (Sengupta et al. 2004), the SF-6d (Brazier et al. 2002), the WHOQOL BREF (Skevington et al. 2004), and the Assessment of Quality of Life (AQoL Richardson et al. 2007). In addition, two scales have been developed specifically for injury, the Functional Capacity Index (MacKenzie et al. 1996) and the older Injury Impairment Index (III, Miller et al. 1995).

Average III-based QALY losses by diagnosis group are available for diagnoses coded using the International Classification of Diseases, 9th Edition, Clinical Modification; the Abbreviated Injury Scale (AIS), 1998 Edition, by body part and fracture involvement; and the coding system used in the US National Electronic Injury Surveillance System. Estimates with the III system are widely published and are integrated into crash reduction analyses in the USA, Australia, and parts of Canada. The FCI is matched to the Occupant Injury Codes in the 2005 Edition of AIS.

An alternative approach to using an existing generic quality of life instrument is to measure quality of life directly. This involves defining the relevant health states to be valued for an intervention and then calculating the utility values corresponding to these health states using one of a number of techniques available to elicit health state values (Gold et al. 1996).

Rather than using QALYs to measure health status in cost–utility analysis, some analysts use the disability experienced by a person. This is the basis for measuring DALYs. A DALY is simply 1 minus a QALY. Both the World Bank and World Health Organization at times have incorporated age or income weighting in their calculations of DALYs (Murray 1996), but this practice seems to be falling out of favor.

Cost–Benefit Analysis

Cost–benefit analysis is a form of economic evaluation that values all the costs and benefits of interventions in monetary units. The results of cost–benefit analysis generally are reported using two measures that show a net monetary gain or loss [i.e., the difference between benefits and costs (36.4)] and a BCR that shows rate of return on investment (i.e., the ratio of benefits to costs (36.5)):

$$\text{Net benefit} = \text{monetized benefits} - \text{costs}. \quad (36.4)$$

$$\text{Benefit} - \text{cost ratio (BCR)} = \text{return on investment} = \frac{\text{monetized benefits}}{\text{costs}}. \quad (36.5)$$

A number of methods can be used to place monetary values on the health gains that arise from safety and health interventions. The two main categories are the willingness to pay approach and the human capital approach.

Willingness to Pay Approach

The willingness to pay approach builds on economic theory (e.g., Linnerooth 1979; Arthur 1981). It starts from the observation that safety and health programs do not save lives with certainty. Rather, they make typically small changes in the risk of death, illness, or injury or increase the chance of recovery. Theory suggests that public policy should base decisions about the amount the government should pay to reduce risk on the amount that citizens are willing to pay and actually do pay for risk reductions (Dréze 1964; Schelling 1968). More than 200 studies from around the world have estimated what people pay for small changes in their chance of being killed or injured, how much more they earn if their job is risky, or how much survey respondents state they are willing to pay for risk reductions.

From the willingness to pay estimates, one derives the value of statistical life (VSL). The VSL is the monetary value attached to reducing society’s risk of death by one life. For example, suppose a driver air bag reduces a driver’s risk of death by 1 in 10,000 and 10,000 people each voluntarily spent \$500 to get a car with a driver air bag. In aggregate, they would spend \$5 million (10,000 × \$500) on

Table 36.4 Summary of survey-based estimates of value of statistical life (VSL) (in 2009 US dollars)

	Full sample	Trimmed sample ^a
Mean VSL (standard deviation)	6,894,990 (500,169)	5,043,947 (199,853)
Weighted mean VSL ^b (standard deviation)	8,110,958 (908,787)	5,061,174 (331,903)
Median	2,815,288	2,815,288
Minimum value	4,904	63,504
Maximum value	217,100,000	39,300,000
Number of observations	937	891

Source: Lindhjem et al. (2010)

^aHighest and lowest 2.5% of the values taken out of the sample

^bWeighted by the inverse of the number of observations from each survey

air bags. The expected number of lives saved would be 1.0 ($10,000 \times 1/10,000$). So, the 10,000 purchasers valued a statistical life (an expected value of one life saved) at \$5 million.

Some of the 200 plus VSL studies are of far better quality than others. People complain about the wide range of values, often listing extremes that come from a developing country at the low end or at the high end, a survey that failed to trim billion dollar outliers or at least probe if they resulted from a respondent's lack of algebraic competence. Table 36.4 illustrates that the variance is really not very large and that the most extreme ends of the range come from studies with severe quality problems.

The complainers also fail to recognize that none of the other estimates in a benefit–cost analysis is very certain. The percentage reduction in death risk due to passage of a law, for example, commonly has a standard deviation equal to 25–33% of the mean, with some jurisdictions sometimes even experiencing increases (e.g., Tippetts et al. 2005). Effectiveness estimates for engineering improvements to roads have equally high variance (Elvik et al. 2009). Worse, the percentage reduction in injuries often is simply assumed to follow in lockstep with fatalities, which it does not.

Table 36.5 summarizes the results of ten meta-analyses and older systematic reviews that analyzed the VSL. In general, studies of the labor market have higher values than survey-based studies or studies of markets for safety devices. Some also believe the best labor market studies are subject to fewer validity threats. Thus, a reasonable VSL choice for the USA might be \$5.5 million with sensitivity analysis at \$4 and \$7 million.

Tailoring Values by Country

Seven meta-analyses have examined how to tailor VSL to a specific population. These studies concluded that VSL varies with income. A 1% change in income from a developed country average is associated with a 0.45–1.8% change in VSL. After reviewing the evidence, Hammitt and Robinson (2011) concluded (1) that sensitivity analysis is needed on the value used and (2) that the percentage difference in VSL in low-income countries versus developed countries is almost certainly greater than the percentage difference in per capita income.

Valuing Nonfatal Injury Reduction

Ideally, we would base our values for statistical injuries on the same methods as the VSL. Because injury prevention affects the incidence of hundreds of different diagnoses, that ideal is unrealistic. An alternative that has been used extensively is to estimate the QALY loss associated with the different injuries, then to estimate the willingness to pay for a QALY.

Table 36.5 Value of statistical life estimates: mean and uncertainty range from selected meta-analyses and systematic reviews (in 2009 US dollars) and characteristics of those reviews

Study	Formal meta-analysis	No. of values	Best estimate	Range	Context
Miller (1989)	No	47 World estimates (also identified 20 unsound values)	\$4.4 M	±23%	All
Viscusi (1993)	No	33 World studies	\$8.6 M	±40%	Labor market, survey
Desvousges et al. (1995)	Yes	29 World studies from Viscusi's 33	\$5.5 M		Labor market
Miller (2000)	Yes	68 World estimates	\$5.6 M (USA)	\$5.0–\$6.8 M	All
Mrozek and Taylor (2002)	Yes	203 World estimates, from 33 studies	\$4.8 M ^a	±35%	Labor market
Viscusi and Aldy (2003)	Yes	49 World studies (reviewed more than 60 studies, but some lacked desired variables)	\$7.2 M ^b (US)	\$5.6–\$10.6 M	Labor market
Kochi et al. (2006)	Yes	234 World estimates from 40 studies	\$6.6 M	±44%	Labor market, survey
Bellavance et al. (2009)	Yes	37 Estimates from 34 studies (rejected 15 others that lacked desired data or were flawed)	\$8.3 M	±19%	Labor market
Biausque (2010)	Yes	366 Estimates from 34 studies their authors felt merited inclusion	\$3.5 M	\$2.8–\$3.9 M	Survey on value of adults
Lindhjem et al. (2010)	Yes	1,095 Estimates from 92 studies	\$5.1 M	\$4.7–\$8.2 M	Survey on value of adults

^aThe values selected from Mrozek and Taylor (2002) are based on NIOSH-risk data, which generally are acknowledged to be more complete than contemporaneous Bureau of Labor Statistics risk data

^bModels 3 and 6 from Viscusi and Aldy's (2003) Table 8 were excluded because they have few degrees of freedom and large standard errors

Like all efforts to value intangibles, valuing QALYs has proven controversial. The most common valuation approach has been to simply divide the VSL by the number of QALYs left in an average life (discounted to present value). Others have used surveys to value QALYs directly. The stumbling block is that both the surveys and economic theory suggest the value of a QALY is situational (Hammit 2007; Haninger and Hammit 2007). Although the simple division may be the only practical value for use in economic analysis currently, sensitivity analysis is needed that recognizes that people with few QALYs remaining may value QALYs differently. Some place high values on this scarce commodity. Others feel their time has passed and tell researchers the money should be spent to protect younger people, not them. The literature also suggests society and especially parents value risk reduction for children more highly than for adults (Hammit and Haninger 2010).

Human Capital Approach

The human capital approach entails valuing health benefits by measuring the flow of income that would have been lost if the illness had not been treated or the injury had not been prevented. If the analysis takes the perspective of the health system, human capital cost equals the medical cost savings plus associated savings in claims processing costs. From a government perspective, as the section above on perspective explained, the cost savings are somewhat broader but all are tangible losses that are readily valued in monetary units.

Human capital costs value a person's life as the present value of the amount they are expected to earn over their remaining lifespan, often supplemented by the value of the housework they are expected to do. Although human capital cost estimates first were published by Adam Smith in 1776, analysts never have been able to provide a theoretical basis for using them. From a societal viewpoint, human capital costs are not supported by economic theory.

Moreover, human capital costs are biased. They undervalue the lives of the elderly who no longer are working, especially if they are wealthy enough that they employ a cleaning person or eat out rather than cook. Because of discounting, they value children at lower values than some adults because the children's stream of discounted earnings starts many years in the future. They value the lives of women and minorities who face wage discrimination at lower amounts than nonminority males. From a public policy viewpoint, it also seems inappropriate for the government to value the lives of upper-income citizens at a higher value than the lives of lower-income citizens. One way to reduce the earnings bias problems is by pricing everyone's work losses at the average weekly earnings rate. Even with that approach, by ignoring nonfinancial costs such as loss of quality of life, empirical estimates show that the human capital approach usually fails to capture the majority of the benefits associated with being alive and healthy.

Use of human capital costs in cost-benefit analysis (CBA) persists because they are relatively easy to calculate. CBAs from a societal perspective that use human capital costs are misleading. They are unsupported by theory and quite biased. Unintentional injury prevention measures that would not be cost-beneficial if only human capital costs were considered include passenger air bags, vehicle side impact protection, all-terrain vehicle helmets, community smoke alarm programs, sprinklers in new homes, and the US mattress flammability standard. Hauer (1994) points out that at even a modest discount rate, human capital costs value a death at a lower value than spending an equivalent number of hours of delay time at red lights in the next year. That means road builders who use human capital costs to value safety implicitly have decided "better dead than stuck in traffic." CBAs with benefits valued using human capital costs unfortunately continue to slip through peer review. When they do, informed consumers should ignore them.

Adjusting for Differential Timing

Inflation Adjusting

Costs often occur at different times and may extend over a number of years. In economic evaluation, costs must all be counted in a base year (usually the present year or the year of the incidence data) and must be adjusted to eliminate the effects of inflation. This is done to calculate the real cost of resource use in a constant value of the local currency.

For example, if an intervention prevented one emergency department visit per year and the medical care inflation rate was 3%, the medical spending prevented in the first year would be 3% less than the spending prevented in the second year. Adjusting for inflation lets one add the costs together in a way that makes them equivalent in terms of their purchasing power.

To adjust for inflation, one uses a price index. This may be a general price index or one specific to a cost category like medical care. One multiplies each monetary amount times the price index value for the base year divided by the value for the year of the data. Adjusting from national prices to local prices uses a similar formula if a spatial price index is available (e.g., the ratio of medical care prices in New Jersey to medical care prices in the USA).

Discounting

Discounting is the process of giving greater weight to those costs that arise sooner rather than later. Economists tend to agree that people prefer to have money now as opposed to later. For example, for two reasons, most people would prefer to have \$1,000 today rather than the promise of \$1,000 next year. First, \$1,000 today could be invested at the current interest rate; at a 5% interest rate, the \$1,000 would be worth \$1,050 by this time next year. Second, the choice of \$1,000 today eliminates any uncertainty of not receiving the \$1,000 next year.

To reflect this phenomenon, future costs in economic evaluation are discounted (i.e., given less weight). The standard approach to discounting reduces a stream of costs to an equivalent amount in today's values. That single amount is known as the present value of the future stream of costs (or benefits). The rate at which the present value is computed is known as the discount rate, and thus, the discount rate is, in effect, an "exchange rate" between value today and value in the future.

Assume a program costs \$1,000 this year and \$1,500 next year, and an agreed discount rate of 5%.

Then, the present value (PV) of this stream of costs is:

$$PV = 1,000 + 1,500 / (1 + 0.05) = 2,429 \text{ (i.e., not 2,500).}$$

If the \$1,500 instead were incurred 2 years from now, one would divide by 1.05^2 with a similar formula in future years. Discounting is quite easy in a spreadsheet. Also, discount tables and calculators are available.

While most economists agree that future costs should be discounted, disagreement exists about whether future health outcomes should also be discounted and, if health outcomes are to be discounted, then whether they should be discounted using the same discount rate. Perhaps, the most telling argument in favor of discounting is that uncertainty means almost everyone would rather have a broken leg in 50 years than a broken leg now. The reasons for this include the fact that in the next 50 years, we may find a way to instantly heal a broken leg and also that the person might not even be alive 50 years from now. Indeed, empirical studies show that people generally discount health at a higher interest rate than money.

Most guidelines on economic evaluation recommend discounting of both costs and outcomes using the same discount rate. This discount rate is usually specified in government guidelines.

Confidence Intervals and Sensitivity Analysis

Given the range of assumptions and uncertainty of the estimates inherent in most economic evaluations, it is essential to test the sensitivity of the results of a study to changes in these assumptions. This is true of both cost data and injury outcomes. For example, what impact would different unit prices have on the CER? Or different discount rates? Or a different effectiveness level for the intervention?

An analysis of the impact of a possible range of plausible cost and outcome values, given the uncertainty surrounding the estimate of the true cost and outcome values of an intervention, is called a sensitivity analysis. The sensitivity analysis recomputes the cost–outcome measure with alternative values for selected parameters. The parameters chosen depend on their magnitudes and uncertainty levels. If only one parameter is varied at a time, the sensitivity analysis is one-dimensional. Multidimensional sensitivity analyses test more than one parameter at a time. Sometimes extreme scenarios are tested, in which a range of variables with worst case scenario (e.g., high cost, low effectiveness) or best-case scenario (e.g., low cost, high effectiveness) are used. If the results of the study are not greatly affected over the range of variables, then the study results are considered robust.

Sometimes cost–effectiveness analysts bootstrap estimated uncertainty ranges around the ICERs. Sensitivity analysis still is desirable when bootstrapping in order to evaluate how the choice of discount rate and other analytic assumptions affect the estimates.

Reporting the Results of an Economic Evaluation

The results of an economic evaluation should report all of the assumptions and estimates built into the evaluation and the sources that underlie them. The British Medical Journal developed the following checklist to guide authors and referees on the contents that should be reported in an article describing an economic evaluation (Appendix 1).

Incorporating Economic Evaluation in Decision Making

The cost–effectiveness plane was discussed in section “Framing the Study” in the context of the different scenarios that can arise in relation to the costs and effectiveness of one safety program compared with another. If one intervention dominates the other, that is, if it costs less and has better outcomes, the decision is straightforward. Or if it is dominated by the other intervention (i.e., costs more and has worse outcomes), again the decision is straightforward. However, in many cases, a safety program will be more effective than an alternative option but cost more. How then is a decision made as to whether the safety program represents value for money?

Cost–Effectiveness League Tables

An approach that is often used when evaluating the cost–effectiveness of a new treatment is to compare the BCR or incremental cost–effectiveness per life year (LY) gained or quality adjusted life year (QALY) gained across different treatments. Ranking the treatments from lowest cost per LY gained or lowest cost per QALY gained to highest cost per LY gained or highest cost per QALY gained creates a “league table” that compares the values the treatments provide per dollar invested. Table 36.6 is an example of a league table.

These side-by-side comparisons of economic evaluation results can be useful if the cost–effectiveness analyses have been undertaken specifically to facilitate between-treatment comparisons using standardized methodologies, or if results of economic evaluations of treatments have been adjusted and standardized to make comparisons meaningful. However, this is not always done in preparing cost–effectiveness league tables, with the result that studies in the same table may take different perspectives, use different discount rates, and so on. This often makes CERs in a league table noncomparable. A notable exception is a league table maintained for injury and substance abuse (Miller and Hendrie 2009). The table currently includes 160 interventions with estimates of cost/QALY saved and BCRs computed at a 3% discount rate with consistently computed costs for injury, illness, and other societal ills. A notable feature of this table is the assumption that replications of demonstration programs and randomized trials will achieve 25% less effectiveness than the original programs.

Other concerns regarding the use of league tables are:

1. The use of league tables assumes that the original context of the study in each case is transferable to the specific context within which decisions are currently being made (Drummond et al. 2005).
2. Presenting the results as a single CER in league tables ignores the confidence interval around each CER.
3. Although cost–effectiveness league tables help the decision maker to consider cost–effectiveness results in context, it is not always clear how the information is to be used. For example, if criteria other than cost–effectiveness/return on investment, and net benefit have been used to fund a treatment that is less cost–effective, then how should this impact on funding of treatment?

One way that league tables can inform decision makers is in showing which treatments are clearly not cost–effective (i.e., those with relatively high costs per life year gained or QALY gained). These treatments can then be excluded from consideration for funding as they obviously do not represent “good buys or value for money.”

The Difficulty of Basing Decisions on Cost–Effectiveness

Cost–effectiveness analysis is most useful in guiding the choice between interventions when the decision to take action is already made or the issue is whether to replace an implemented program. With a cost–effectiveness analysis, when using outcome measures that are specific to a particular type of injury (e.g., fall prevented) rather than generic (e.g., life years saved), separating those interventions that represent value for money from those that do not requires some judgment, as no threshold or cutoff values exist. Approaches that can be used to derive these cutoff values include comparing the cost per unit of outcome with other programs, “rules of thumb,” and inferences from past decisions (Gold et al. 1996; Hendrie and Miller 2004).

Table 36.6 League table for child and family injury prevention and control: unit costs, cost savings, and costs/QALY saved (in 2009 dollars, computed at a 3% discount rate)

	Unit cost	Medical cost savings	Productivity and other savings	QALY savings ^a	Benefit–cost ratio	Cost/QALY ^b
<i>1. Road safety</i>						
Pass child safety seat law, ages 0–4	\$58/new user	\$160	\$530	\$1,500	38	<\$0
Child safety seat distribution, ages 0–4	\$52/seat provided	\$160	\$530	\$1,500	42	<\$0
Child seat misuse check-up	\$6/seat in use	\$57	\$220	\$300	81	<\$0
Pass booster seat law, ages 4–7	\$39/new user	\$360	\$790	\$1,300	63	<\$0
Booster seat, distribution, ages 4–7	\$35/seat	\$360	\$790	\$1,300	71	<\$0
Pass/upgrade safety belt law	\$340/new user	\$290	\$1,900	\$3,900	18	<\$0
Enhanced belt law enforcement	\$360/new user	\$290	\$1,900	\$3,900	17	<\$0
Zero alcohol tolerance, drivers under 21	\$39/driver	\$61	\$310	\$590	25	<\$0
Intensive sobriety checkpoints	\$12,000/checkpoint	\$5,400	\$22,000	\$55,000	6.8	<\$0
Alcohol-testing ignition interlock	\$1,200/vehicle	\$300	\$1,800	\$4,800	6.6	<\$0
Provisional licensing + midnight driving curfew	\$84/driver	\$43	\$250	\$390	8.1	<\$0
Change driving curfew to 10 p.m.	\$160/driver	\$26	\$150	\$230	2.5	\$38,000
Mobile speed camera	\$740,000/camera-year	\$840,000	\$4,900,000	\$8,300,000	19	<\$0
Red light camera	\$11,500/camera-year	\$5,700	\$21,000	\$22,000	4.3	<\$0
<i>2. Home and community safety</i>						
Childproof cigarette lighter	\$0.05/lighter	\$0.40	\$1.40	\$1.80	72	<\$0
Pass smoke alarm law	\$49/new user ^c	\$9	\$150	\$660	17	\$0
Smoke Alarm Installation & Fire Education Program	\$310/home ^c	\$17	\$280	\$1,220	4.9	\$22,000
Monitored burglar and fire alarms	\$910/home/year	\$2	\$490	\$490	1.1	\$122,000
Mattress flammability standard	\$26/mattress	\$0.65	\$10	\$60	2.8	\$75,000
Baby walker redesign to prevent stairway falls	\$3.60/walker	\$17	\$20	\$150	46	<\$0
Harlem Hospital Safe Communities Program	\$75/child	\$220	\$800	\$2,800	51	<\$0
Pediatrician injury prevention counseling, ages 0–4	\$11/child	\$8	\$19	\$71	8.9	\$6,000
Nurse–family partnership home visits to firstborn	\$10,700/child	\$1,400	\$23,000	\$27,000	4.8	<\$0
20-Bed domestic violence shelter	\$18,000/bed	\$12,360	\$34,100	\$153,000	11	<\$0
21 Minimum legal drinking age	\$200/youth 18–20	\$43	\$240	\$450	3.6	\$23,000
Poison control center services	\$43/call	\$320	\$0	\$0	7.4	<\$0
Regional Trauma System Services	\$1,850/admit	\$2,200	\$510	\$2,300	2.7	<\$0

Source: selected from the larger table in Miller et al. (2012)

^aTo determine the number of QALYs saved, divide the QALY savings by \$135,182. Total savings are the sum of the medical cost, productivity, and QALY savings. The benefit–cost ratio equals total savings divided by the unit cost

^bCost / QALY = $\frac{\text{QALYs saved}}{\text{intervention cost} - \text{resource cost savings}}$

^cIncludes 3 alarms per home

Cost–Utility Thresholds

The question as to what cost per QALY represents good value for money depends on a comparison of cost per QALY against some benchmark value – the ceiling ratio or cost–effectiveness threshold – that represents society’s willingness to pay for a unit of health improvement. The concept of a single cost effectiveness threshold suggests that any treatment with an ICER below the threshold is funded, whereas any treatment with an ICER above the ceiling ratio is rejected or not funded.

The concept of a single threshold value that represents a cutoff point for funding is simplistic, as factors other than cost–effectiveness impact on resource allocation. Nevertheless, unofficial thresholds can be inferred from past funding decisions by examining the ICER cutoff below which treatments are generally accepted for funding and above which funding is generally rejected. As a general rule of thumb, pharmaceutical, clinical, and public health interventions would not be implemented in North America if they cost more than US \$100,000–\$130,000 per QALY (Clement et al. 2009; Lee et al. 2009). Australia, the Netherlands, and the UK use lower thresholds of \$50,000–\$70,000 per QALY (Harris et al. 2008; Lee et al. 2009). The threshold in New Zealand is believed to be much lower, NZ\$20,000 (Simeons 2009). In low- and middle-income countries, the WHO Commission on Macroeconomics and Health (2001) has suggested that any intervention that costs up to three times GDP per capita should be viewed as cost–effective.

However, as discussed previously, cost–effectiveness is likely to be just one of several criteria influencing resource allocation. For example, the National Institute for Clinical Effectiveness (NICE) in the UK considers other criteria such as wider social costs and benefits including the effects of treatment on productivity and patients’ ability to return to work, the innovative nature of the treatment, and consistency with previous judgments (Morris et al. 2007). It rules out other criteria such as affordability, differential treatment on the grounds of orphan drug status, and the personal characteristics of those achieving health gain (e.g., age, risk factor behavior). In New Zealand, criteria other than cost–effectiveness that are considered are equity, acceptability, and Treaty of Waitangi obligations to serve the Maori population (Morris et al. 2007).

If cost–effectiveness is not the only criterion that is considered in resource allocation decisions, then decision makers will be prepared to sacrifice some efficiency for gains in these other criteria. This suggests that rather than a single cost–effectiveness threshold, the more realistic likelihood is having lower and upper thresholds. Within the range between the lower and upper threshold, cost–effectiveness is being traded off against the other criteria.

The inherent uncertainty in all ICERs is another reason for not establishing a single cutoff cost per QALY. A final complicating factor is that the decision response to a given ICER may depend on whether the decision is to invest in a new approach or disinvest in an established one. Established programs and approaches develop political constituencies stoked both by comfort with and belief in the status quo and by self-interest. It appears that the probability of rejection is lower at every ICER, for example, for currently used medical treatments as opposed to new therapies (Morris et al. 2007).

Cost–Benefit Threshold

Any intervention with a BCR greater than 1.0 is expected to return more than it costs. That does not mean it is worth implementing. The BCR is the best estimate of the return, but it has considerable uncertainty. Approximately half the implementations will perform worse than the mean. Furthermore, budgets are always tight, and many interventions with high BCRs exist. Therefore, it is wise not to implement interventions with a BCR below perhaps 1.6–2.0. An exception to this rule is that the only intervention known to be effective against a problem of major concern in the community or with a special population might be worth implementing at a lower ratio.

Table 36.7 League table for school-based substance abuse prevention programs: cost per pupil; percentage of participants delaying initiation or reducing alcohol, marijuana, cocaine, and tobacco use; and benefit–cost ratio (in 2009 dollars computed at a 3% discount rate)

Program	Cost/pupil	Alcohol	Marijuana	Drugs	Tobacco	BCR
All Stars	\$170	7.0%	6.4%	0.0%	6.0%	36
Keepin' It Real	\$155	10.9%	4.9%	–	2.1%	28
Project Northland	\$490	6.9%	6.6%	3.3%	9.0%	19
Project Toward No Tobacco	\$220	0.0%	0.0%	0.0%	5.5%	19
STARS for Families	\$150	9.2%	–	–	–	8
Project Toward No Drugs	\$220	4.4%	–	3.9%	0.0%	4

Source: selected from the larger table in Miller and Hendrie (2009)

Economics Is Only One Input to Policy

Choices in the real world are complex. They typically require weighing more than just economics. Consider a decision about whether to maintain the status quo or add a locality's first school-based or community-based substance abuse prevention program that will reduce youth impaired driving, violence, and other problem behaviors. Table 36.7 shows the costs, effectiveness, and BCR for a range of proven prevention programs. These programs are multidimensional. They can affect use of alcohol, marijuana, other illicit drugs, and tobacco. In the table, All Stars dominates Project Toward No Tobacco (lower right quadrant). Keepin' It Real costs more than STARS for Families but also is more effective against all substances and offers a higher return on investment (upper right quadrant). Most of the programs, however, do not fit any quadrant. Keepin' It Real probably dominates Project Toward No Drugs, but one program evaluated its effect on marijuana use, while the other evaluated its effect on all illicit drug use. Comparing them accurately requires assuming how these two measures track one another. Project Northland costs about three times as much as All Stars or Keepin' It Real. It is more effective against illicit drugs and tobacco but less effective against alcohol. Which of the three programs to adopt depends on what problem the community is most interested in preventing, whether the community can afford to broadly implement one of the more costly programs, and how comfortable the local school board is with each program. The latter consideration can be decisive, for example, if one program discusses safe sex, and the community is wedded to abstinence messaging.

Conclusion

In summary, economic analysis is an important input to decision making in a resource-constrained world. It helps guide choices about how funds should be allocated both within and between program areas. A good return on investment attracts potential funders and provides a powerful defense when budget cuts have to be made.

The usefulness of studies is dependent on the quality of the evidence used, as well as the methods and analytical techniques adopted. Economic evaluations of injury prevention and control programs that have been conducted according to methodologically sound principles present useful information to decision makers about the costs, efficiency, and affordability of alternative courses of action. Together with concerns of equity and political feasibility, efficiency considerations help answer questions about how to allocate injury prevention and control resources in a way that maximizes the return on investment (Hendrie and Miller 2004).

Appendix 1 Guidelines for Authors and Peer Reviewers of Economic Submissions to the British Medical Journal

STUDY DESIGN

(a) Study question

- The economic importance of the research question should be outlined.
- The hypothesis being tested, or question being addressed, in the economic evaluation should be clearly stated.
- The viewpoint(s) – for example, health care system, society – for the analysis should be clearly stated and justified.

(b) Selection of alternatives

- The rationale for choice of the alternative programs or interventions for comparison should be given.
- The alternative interventions should be described in sufficient detail to enable the reader to assess the relevance to his or her setting – that is, who did what, to whom, where, and how often.

(c) Form of evaluation

- The form(s) of evaluation used – for example, CEA, CUA, CBA – should be stated.
- A clear justification should be given for the form(s) of evaluation chosen in relation to the question(s) being addressed.

DATA COLLECTION

(d) Effectiveness data

- If the economic evaluation is based on a single effectiveness study (for example, a clinical trial), details of the design and results of the study should be given (for example, selection of study population, method of allocation of subjects, whether analysed by intention to treat or evaluable cohort, effect size with confidence intervals).
- If the economic evaluation is based on an overview of a number of effectiveness studies, details should be given of the method of synthesis or meta-analysis of evidence (for example, search strategy, criteria for inclusion of studies in the overview).

(e) Benefit measurement and valuation

- The primary outcome measure(s) for the economic evaluation should be clearly stated (for example, cases detected, life years, quality adjusted life years (QALYs), willingness to pay).
- If health benefits have been valued, details should be given of the methods used (for example, time trade off, standard gamble, contingent valuation) and the subjects from whom valuations were obtained (for example, patients, members of the general public, health care professionals).
- If changes in productivity (indirect benefits) are included, they should be reported separately and their relevance to the study question discussed.

(f) Costing

- Quantities of resources should be reported separately from the prices (unit costs) of those resources.
- Methods for the estimation of both quantities and prices (unit costs) should be given.
- The currency and price date should be recorded and details of any adjustment for inflation, or currency conversion, given.

(g) Modeling

- Details should be given of any modeling used in the economic study--for example, decision tree model, epidemiology model, regression model.
- Justification should be given of the choice of the model and the key parameters.

ANALYSIS AND INTERPRETATION OF RESULTS

(h) Adjustments for timing of costs and benefits

- The time horizon over which costs and benefits are considered should be given.
- The discount rate(s) should be given and the choice of rate(s) justified.
- If costs or benefits are not discounted an explanation should be given.

(i) Allowance for uncertainty

- When stochastic data are reported, details should be given of the statistical tests performed and the confidence intervals around the main variables.
- When a sensitivity analysis is performed, details should be given of the approach used (for example, multivariate, univariate, threshold analysis) and justification given for the choice of variables for sensitivity analysis and the ranges over which they are varied.

(j) Presentation of results

- An incremental analysis--for example, incremental cost per life year gained-- should be reported, comparing the relevant alternatives.
- Major outcomes--for example, impact on quality of life-- should be presented in a disaggregated as well as aggregated form.
- Any comparisons with other health care interventions--for example, in terms of relative cost effectiveness--should be made only when close similarity in study methods and settings can be demonstrated.
- The answer to the original study question should be given; any conclusions should follow clearly from the data reported and should be accompanied by appropriate qualifications or reservations.

References

- Arthur, W. B. (1981). The economics of risks to life. *The American Economic Review*, 71(1), 54–64.
- Bellavance, F., Dionne, G., & Lebeau, M. (2009). The value of a statistical life: a meta-analysis with a mixed effects regression model. *Journal of Health Economics*, 28(2), 444–464.
- Biausque, V. (2010). *The value of statistical life: a meta-analysis*. Working Party on National Environmental Policies, OECD, Geneva.
- Brazier, J., Roberts, J., & Devierill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, 21(2), 271–292.
- Clement, F. M., Harris, A., Li, J. J., Yong, K., Lee, K. M., & Manns, B. J. (2009). Using effectiveness and cost-effectiveness to make drug coverage decisions: a comparison of Britain, Australia, and Canada. *Journal of the American Medical Association*, 302(10), 1437–1443.
- Desvousges, W., Johnson, F. R., & Banzhaf, S. H. (1995). *Assessing environmental externality costs for electricity generation*. Triangle Economic Research report to Northern States Power Company, Research Triangle Park, NC.
- Dréze, J. H. (1964). Some postwar contributions of French economists to theory and public policy: with special emphasis on problems of resource allocation. *The American Economic Review*, 54(4), 2–64.
- Drummond, M. F., Sculpher, M., Torrance, G. W., O'Brien, B., & Stoddart, G. L. (2005). *Methods for the economic evaluation of health care programmes* (3rd ed.). Oxford: Oxford University Press.
- Elvik, R., Hoye, A., Vaa, T., Erke, A., & Sorensen, M. (2009). *Handbook of road safety measures* (2nd ed.). Bingley, UK: Emerald.
- EuroQol Group. (2007). *EQ-5D value sets: inventory, comparative review and user guide*. The Netherlands: Springer.
- Gluck, H. A., Doshi, J. A., Sonnad, S. S., & Polsky, D. (2007). *Economic evaluation in clinical trials*. New York: Oxford University Press.
- Gold, M. R., Franks, P., & Erickson, P. (1996). Assessing the health of the nation: the predictive validity of a preference-based instrument and self-rated health. *Medical Care*, 34(2), 163–177.
- Goldstein, J. A., Winston, F. K., Kallan, M. J., Branas, C. C., & Schwartz, J. S. (2008). Medicaid-based child restraint system disbursement and education and the vaccines for children program: comparative cost-effectiveness. *Ambulatory Pediatrics*, 8(1), 58–65.
- Hammit, J. K. (2007). Valuing changes in mortality risk: lives saved versus life years saved. *Review of Environmental Economics and Policy*, 1(2), 228–240.

- Hammit, J. K., & Haninger, K. (2010). Valuing fatal risks to children and adults: effects of disease, latency, and risk aversion. *The Journal of Risk and Uncertainty*, 40, 57–83.
- Hammit, J., & Robinson, L. (2011). The income elasticity of the value per statistical life: transferring estimates between high and low income populations. *Journal of Benefit-Cost Analysis*. doi:10.2202/2152-2812.1009.
- Haninger, K., & Hammit, J. K. (2007). *Willingness to pay for quality-adjusted life years: empirical inconsistency between cost-effectiveness analysis and economic welfare theory*. Working paper, Harvard Center for Risk Analysis, Cambridge, MA.
- Harris, A. H., Hill, S. R., Chin, G., Li, J. J., & Walkom, E. (2008). The role of value for money in public insurance coverage decisions for drugs in Australia: a retrospective analysis 1994–2004. *Medical Decision Making*, 28(5), 713–722.
- Hatziaandreu, E. J., Sacks, J. J., Brown, R., Taylor, W. R., Rosenburg, M. L., & Graham, J. D. (1995). The cost effectiveness of three programs to increase use of bicycle helmets among children. *Public Health Reports*, 110(3), 251–259.
- Hauer, E. (1994). Can one estimate the value of life or is it better to be dead than stuck in traffic? *Transportation Research Series A*, 28A(2), 109–118.
- Hendrie, D., & Miller, T. R. (2004). Economic evaluation of injury prevention and control programs. In R. McClure, M. Stevenson, & S. McEvoy (Eds.), *The scientific basis of injury prevention and control*. Sydney: IP Communications.
- Kochi, I., Hubbell, B., & Kramer, R. (2006). An empirical Bayes approach to combining and comparing estimates of the value of a statistical life for environmental policy analysis. *Environmental & Resource Economics*, 34(3), 385–406.
- Lawrence, B. A., Max, W., & Miller, T. R. (2003). *Costs of injuries resulting from motorcycle crashes: a literature review* (Report No. DOT HS 809 242). Washington, DC: National Highway Traffic Safety Administration.
- Lee, C., Chertow, G., & Zenios, S. (2009). An empiric estimate of the value of life: updating the renal dialysis cost-effectiveness standard. *Value in Health*, 12(1), 80–87.
- Lestina, D., Miller, T. R., Langston, E. A., Knoblauch, R., & Nitzburg, M. (2002). Benefits and costs of ultraviolet fluorescent lighting. *Traffic Injury Prevention*, 3(3), 209–215.
- Lindhjem, H., Analyse, V., & Navrud, S. (2010). *Meta-analysis of stated preference VSL studies: further model sensitivity and benefit transfer issues*. Working Party on National Environmental Policies, OECD, Geneva.
- Linnerooth, J. (1979). The value of human life: a review of the models. *Economic Inquiry*, 17(1), 52–74.
- MacKenzie, E. J., Damiano, A., Miller, T. R., & Luchter, S. (1996). The development of the functional capacity index. *The Journal of Trauma*, 41(5), 799–807.
- Meyer, G., Wegscheider, K., Kersten, J. F., Icks, A., & Mühlhauser, I. (2005). Increased use of hip protectors in nursing homes: economic analysis of a cluster randomized controlled trial. *Journal of the American Geriatrics Society*, 53(12), 2153–2158.
- Miller, T. R. (1989). *65 MPH: Winners and losers*. Washington, DC: The Urban Institute.
- Miller, T. R. (1994). *Costs of safety belt and motorcycle helmet nonuse - Testimony to Subcommittee on Surface Transportation*. Washington, DC: House Committee on Public Works and Transportation.
- Miller, T. R. (2000). Variations between countries in values of statistical life. *Journal of Transport Economics & Policy*, 34(2), 169–188.
- Miller, T. R., & Blewden, M. (2001). Costs of alcohol-related crashes: New Zealand estimates and suggested measures for use internationally. *Accident; Analysis and Prevention*, 33(6), 783–791.
- Miller, T. R., & Hendrie, D. (2009). *Substance abuse prevention dollars and cents: a cost-benefit analysis*. DHHS Pub. (SMA) 07-4298. Substance Abuse and Mental Health Services Administration, Rockville, MD.
- Miller, T. R., Pindus, N. M., Douglass, J. B., & Rossman, S. B. (1995). *Databook on nonfatal injury: incidence, costs, and consequences*. Washington, DC: The Urban Institute Press.
- Miller, T. R., Finkelstein, E., Zaloshnja, E., & Hendrie, D. (2012). The cost of child and adolescent injuries and the savings from prevention. In K. Liller (ed.), *Injury prevention for children and adolescents: research, practice, and advocacy* (2nd ed.). Washington, DC: American Public Health Association.
- Morris, S., Devlin, N., & Parkin, D. (2007). *Economic analysis in health care*. Chichester: Wiley.
- Mrozek, J. R., & Taylor, L. O. (2002). What determines the value of life? A meta-analysis. *Journal of Policy Analysis and Management*, 21(2), 253–270.
- Murray, C. J. L. (1996). Rethinking DALYs. In C. J. L. Murray & A. D. Lopez (Eds.), *The global burden of disease* (The global burden of disease and injury series, Vol. 1). Cambridge, MA: Harvard University Press.
- Richardson, J., Peacock, S., Iezzi, A., Day, N., & Hawthorne, G. (2007). *Construction and validation of the Assessment of Quality of Life (AQoL) Mark II instrument*. Research paper 24. Centre for Health Economics, Monash University, Melbourne.
- Schelling, T. (1968). The life you save may be your own. In S. Chase (Ed.), *Problems in public expenditure analysis*. Washington, DC: Brookings Institution.
- Sengupta, N., Nichol, M. B., Wu, J., & Globe, D. (2004). Mapping the SF-12 to the HU13 and VAS in a managed care population. *Medical Care*, 42(9), 927–937.
- Simeons, S. (2009). Health economic assessment: a methodological primer. *International Journal of Environmental Research and Public Health*, 6(12), 2950–2966.

- Skevington, S. M., Lofty, M., O'Connell, K. A., & Group, W. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. *Quality of Life Research*, *13*(2), 299–310.
- Tippetts, A. S., Voas, R. B., Fell, J. C., & Nichols, J. L. (2005). A meta-analysis of .08 BAC laws in 19 jurisdictions in the United States. *Accident; Analysis and Prevention*, *37*(1), 149–161.
- Trumbull, W. N. (1990). Who has standing in cost-benefit analysis? *Journal of Policy Analysis and Management*, *9*(2), 201–218.
- Viscusi, W. K. (1993). The value of risks to life and health. *Journal of Economic Literature*, *31*(4), 1912–1946.
- Viscusi, W. K., & Aldy, J. E. (2003). The value of statistical life: a critical review of market estimates throughout the world. *The Journal of Risk and Uncertainty*, *27*(1), 5–76.
- Ware, J. E., Kosinski, M., Turner-Bowker, D. M., & Gandek, B. (2002). *How to score version 2 of the SF-12 health survey (with a supplement documenting version 1)*. Lincoln, RI: Quality Metric.
- WHO Commission on Macroeconomics and Health. (2001). *Macroeconomics and health: investing in health for economic development*. Geneva: World Health Organization.
- Zaloshnja, E., Miller, T., Jones, P., Litovitz, T., Coben, J., Steiner, C., et al. (2008). The impact of poison control centers on poisoning-related visits to EDs – United States, 2003. *The American Journal of Emergency Medicine*, *26*(5), 310–315.

Index

A

Abbreviated Injury Scale (AIS), 62, 269, 283–288
Abrasion, 96–97
Absolute threshold (AL), 143
Accidental death, 92
ACSCOT field decision scheme
 anatomic criteria, 302
 mechanism and risk factor criteria, 302–303
 physiologic criteria, 301–302
Adequate stimulus intensity, 142–143
Adverse drug reactions, 592
Adverse event reporting, 588
Affordances, 148–149
Agent factor, 205–206
Age–period–cohort modeling
 concepts and definitions, 407–409
 graphical data, 409–411
 historical uses, 411–412
 implications, 419–420
 median polish procedure and methods, 415–419
 multi-phased method, 414–415
 statistical approaches, 412–414
Air bag systems, 123–124
Alanine aminotransferase (ALT), 344
Alcohol research, 186–199
Alcohol use, 258–260
Alkaline phosphatase (ALP), 344
Ambulatory medical care, 67–68
Analytic coding, 227
Anterograde amnesia, 62
ApoE, 340
Articulated body model (ATB), 129
Aspartate aminotransferase (AST), 344

B

Barell matrix, 31
 injury profiling
 AIS, 273
 background, 269–270
 ICD-9-CM matrix, 270–273
 for traumatic brain injury, 70
Behavioral approach
 child-resistant packaging and unintentional
 poisoning risk, 558–559

ecological models, 552–553
host
 drinking and driving, 556
 fire escape behavior, 554–555
 vector/vehicle, 557
 water heaters and scald burn risk, 557–558
injury influence, 553–554
media coverage and house fires, 562–563
pediatric injury, 561–562
roles for
 education and behavior change, 551
 3Es, 550–551
 traffic calming and pedestrian injury, 559–561
Behavioral determinants
 archival and secondary data, 261–262
 biochemical measures, 260–261
 National highway traffic safety association
 (NHTSA), 261
 naturalistic driving studies, 260
 protecting others, 255–257
 self protecting, 254–255
 self-reports
 alcohol use, 258–260
 cognitive interviewing, 258
 validation studies, 257
Behavioral economics, 197
Behavioral risk factor surveillance system (BRFSS), 261
Belt restraint systems, 121
Bilirubin, 344
Binary recursive partitioning, 327
Biochemical measures, 260–261
Biomarkers, traumatic injury
 biomarker signature, 337
 brain
 genetic polymorphism, 340–342
 metabolites, TBI, 339–340
 mRNA and miRNA, 342–344
 protein biomarkers, TBI, 337–339
 gut
 d-lactate, 345
 intestinal-type fatty acid binding protein
 (I-FABP), 345
 polymorphonuclear elastase, 346
 procalcitonin (PCT), 345
 heart, 346

- Biomarkers, traumatic injury (*cont.*)
- kidney
 - clusterin, 344
 - kidney injury molecule-1 (KIM-1), 344
 - Liver-type fatty acid binding protein (L-FABP), 344
 - neutrophil gelatinase-associated lipocalin (NGAL), 344
 - trefoil factor 3 (TEF3), 344
 - liver, 345
 - lung, 346–347
 - top-down approach, 336
- Biomechanical analysis, 172
- Biomechanical models, 158–161
- Birth cohort, 214
- Bloodstream infections, 593–594
- Blunt force injury
- abrasion, 96–97
 - contusion (bruise), 97
 - description, 96
 - laceration, 98
 - skeletal fracture, 98–99
- Broad surveillance programs, 46
- C**
- Case-control design, 212–213
- Case-crossover design, 213–214
- Catecholamines, 340
- Census of Fatal Occupational Injury (CFOI), 7
- Centers for Disease Control and Prevention (CDC), 24
- CFOI. *See* Census of Fatal Occupational Injury (CFOI)
- Chemokines, 339
- Chop wound, 101–102
- Cleaved tau (C-Tau), 337
- Clinical prediction rules
- characteristics, 329
 - definition, 315
 - development of, 317
 - grades of, 316
 - implementation and impact of, 330
 - need for, 317–318
 - prediction rule derivation
 - outcome assessment, 323–328
 - outcome definition, 322–323
 - population, 320
 - predictor variables, 320–321
 - prospective data, advantages of, 319
 - variables, 321–322
 - prediction rule sensibility, 318–319
 - statistical technique, 323–324
 - validation of, 328–330
- Clusterin, 344
- Cluster randomized trials, 191
- Coding, 227
- Coefficient of friction (COF), 162, 164
- Cognition errors, 150–153
- Cohort design, 214–215
- Combat-related trauma, 79
- Community-level social factors, 245
- Confounders, 162
- Contact wounds, 102–103
- Contusion (bruise), 97
- Coroners, 6
- Coroner system, 89
- Cost-benefit analysis
- described, 645–646
 - human capital approach, 656–657
 - nonfatal injury reduction, 656
 - pay approach, willingness, 653–654
 - tailoring values, by country, 654–656
 - threshold, 663
- Cost-effectiveness analysis (CEA)
- described, 645
 - difficulty, 662
 - incremental analysis, 651
 - injury cost savings, 650–651
 - league tables, 659–662
- Cost information, for injury
- equilibrium approach, 373–374
 - human capital framework, 370–372
 - illness calculation, 375–376
 - willingness, pay approach, 372–373
- Cost-minimization analysis, 644
- Costs estimation
- concept, 646–647
 - impact, 647
- Cost-utility analysis (CUA)
- described, 645
 - QALY estimation, 652–653
 - thresholds, 662–663
- Crash injury mechanisms
- delta v and principal direction of force, 127
 - described, 113
 - occupant kinematics, 127–129
 - scene inspection, 129
- Crash prevention engineering
- active safety systems, 541–542
 - driver and occupant factors, 541
 - factors, 540–541
 - timeline, 539
 - types, in USA, 539–540
- Crash survivability, 130–131
- Creatinine kinase-muscle/brain isozyme (CK-MB), 346
- Cross-sectional surveys, 28
- Cytochrome P450, 341–342
- D**
- Decision errors, 152–153
- Descriptive coding, 227
- Differential threshold (DL), 143
- Diffusion of interventions, 629–632
- Distant wounds, 104–105
- d-lactate, 345
- DPSEEA framework, 239
- Drinking and driving, 556
- Dynamic cohort, 214

E

- Economic evaluation, interventions
 - analysis, types, 644
 - confidence intervals and sensitivity analysis, 658
 - cost–benefit analysis
 - described, 645–646
 - human capital approach, 656–657
 - nonfatal injury reduction, 656
 - pay approach, willingness, 653–654
 - tailoring values, by country, 654–656
 - threshold, 663
 - cost–effectiveness analysis (CEA)
 - described, 645
 - difficulty, 662
 - incremental analysis, 651
 - injury cost savings, 650–651
 - league tables, 659–662
 - cost–effectiveness analysis (CEA)
 - described, 645
 - QALY estimation, 652–653
 - thresholds, 662–663
 - decision making, 658–659
 - differential timing adjustment
 - discounting, 657–658
 - inflation adjustment, 657
 - outcome measurement, 649
 - program costs vs. cost savings, 649
 - resources, measurement and value, 649
 - resource use identification, 648
 - steps in, costing, 647–648
 - study, objectives, 644
- Effectiveness vs. efficacy, 622
- Electronic death registration, 18
- Electronic medical records/electronic health records, 39
- Electronic micro-data, 37
- EMS and trauma systems
 - effectiveness research, 574–577
 - goal, 570
 - injury research
 - data sources, 571–573
 - hospital discharge data, 574
 - trauma registries, 574
- Environmental approach
 - causation vs. prevention, 519–520
 - costs, unintended consequences, and trade-offs, 524–525
 - environmental design, 521–524
 - epidemiology, 520–521
- Environmental determinants
 - physical environment, key aspects of
 - global environmental change, 243
 - housing and home environment, 241–242
 - travel environment, 242
 - work environment, 242–243
 - rationale for, 247–249
 - research methods, 246–249
 - social and physical environments
 - defining, 235–236
 - ecological models, 238–239
 - Haddon matrix, 237, 238
 - of health and injury, 239–240
 - life course influences, 241
 - systems approaches and injury prevention, 237–238
 - social environment, key aspects of
 - family and community-level, 245
 - international disparities, 245
 - policies and legislation, 246
 - socioeconomic status and ethnicity, 244
- Environment factor, 206–207
- Epidemiologic methods
 - definition, 203
 - epidemiologic causation, 208–210
 - epidemiologic designs
 - case–control design, 212–213
 - case–crossover design, 213–214
 - classification of, 211
 - cohort design, 214–215
 - natural experiment, 215–216
 - epidemiologic triad
 - agent factor, 205–206
 - environment factor, 206–207
 - extension of, 207–208
 - host factor, 206
 - history, 203–204
 - injury as disease, 210–211
 - specialties, 204
- Epidemiologic triad
 - agent factor, 205–206
 - environment factor, 206–207
 - extension of, 207–208
 - host factor, 206
- Ergonomics
 - biomechanical analysis, 172
 - definition, 139
 - design features
 - adequate stimulus intensity, 142–143
 - cognition errors, 150–153
 - hazard equivocation or masking, 147–149
 - hazard feature detection, 143–146
 - hazard recognition, 142
 - motor performance demands, 153
 - design impact, 139–141
 - metabolic energy expenditure analysis, 172–177
 - operational analysis, 171–172
 - overexertion injuries
 - exertion stress, strain models, 156–166
 - force, work, and power, 154–156
 - physical work capacity and workload analysis, 166–169
 - thermal stress, 170–171
 - thermal stress analysis, 178–179
- Ethnicity, 244

- Exertion stress-strain models
 biomechanical models, 158–161
 confounders, 162
 falls, 162–165
 hernias, 161–162
 overuse syndromes, 166
 vehicles and impact injuries, 165
 whole-body vibration, 165
- Experimental methods
 measurement
 behavioral economics, 197
 neurocognitive performance, 195–196
 safety-related performance, simulators
 for, 192–195
 sleep quality, 197–198
 subjective states, 199
 study designs
 between-group trials, 188–189
 cluster randomized trials, 191
 crossover trials, 189–190
 mixed between-groups/within-subjects trials, 190
 quasi-experiments, 191–192
- F**
 Falls, 162–165
 Family environment, 245
 FARS. *See* Fatality Analysis Reporting System (FARS)
 Fatality Analysis Reporting System
 (FARS), 7, 12, 15, 68
 Firearm injury
 gunshot wounds
 contact wounds, 102–103
 distant wounds, 104–105
 intermediate wounds, 103–104
 shotgun wounds, 105–106
 Fireworks, 614
 F2-isoprostane, 339
 Fixed cohort, 214
 Follow-back surveys, 17
 Forensic pathology
 cause of death, 91
 coroner system, 89
 definition, 90
 injuries, common types of
 abrasion, 96–97
 blunt force injury, 96–99
 contusion (bruise), 97
 firearm injury, 102–106
 laceration, 98
 sharp force injury, 99–102
 skeletal fracture, 98–99
 manner of death
 accidental death, 92
 classification of, 91
 homicide, 93
 natural death, 92
 suicide, 92–93
 undetermined cause of death, 94
 medical examiner system, 90
 postmortem toxicological analysis, 95–96
 in public health and safety, 106–108
- G**
 Gamma glutamyl transferase (GGT), 344
 Geographical information systems (GIS), 247
 analysis result, 457
 crash analysis and prevention
 database management, 446–447
 network analysis, 449–450
 spatial analysis, 447–449
 statistical spatial analysis, 450–452
 visualization and mapping, 447
 hazardous road locations
 calculation, 455
 crash location validation, 452–454
 road network segmentation, 454–455
 spatial pattern modeling, 456–457
 threshold value determination, 455–456
 Geographically weighted regression (GWR). *See* Spatial
 regression
 Glasgow Coma Scale (GCS), 61–62, 289
 Glial fibrillary acidic protein (GFAP), 337–338
 Global environmental change, 243
 Glycolytic intermediates, 340
 Gunshot wounds
 contact wounds, 102–103
 distant wounds, 104–105
 intermediate wounds, 103–104
- H**
 Haddon matrix, 208, 237, 238
 Harborview assessment for risk of mortality score
 (HARM), 290
 Harm susceptibility model, 592–593
 Harm susceptibility ratio (HSR), 592–593
 Hazard equivocation or masking, 147–149
 Hazard feature detection, 143–146
 Hazard identification, patient safety
 adverse event, 588–589
 direct observation, 587
 failure mode and effect analysis, 589–590
 malpractice claims, 588
 medical record, 585–587
 provider surveys/focus groups/patient
 interviews, 588
 simulation, 589
 Hazard recognition, 142
 Health care provider-based injury data, 25
 Heart-type fatty acid binding protein (H-FABP), 346
 Helmet sensors, 80
 Hernias, 161–162
 Home-related injuries, 612–615
 Homicide, 93
 Hospital inpatient data, 26
 Host factor, 206
 Housing and home environment, 241–242
 Human motor performance, 153

4-Hydroxynonenal (4-HNE), 339

Hypothetical regression analysis, 324–325

I

ICD-9-CM matrix

background, 270

Barell+ matrix, 272

conceptual framework, 270–271

ICD-10 matrix, 272–273

injury profiling, 271

matrix use review, 272

ICD codes, 268–269

Incised wound, 99–100

Infodemiology, 51

Injury indicators, 13

Injury mechanisms

acceleration and blunt force impact, human tolerance to, 117

analysis, 134–135

in crashes

delta v and principal direction of force, 127

occupant kinematics, 127–129

scene inspection, 129

description, 111–112

injury identification, 114–116

injury severity, 116–117

regional tolerance to impact, 126

restraint, seating, and support, 120–124

tolerance, 124–126

traumatic injuries, classification of, 113–114

vehicle

body fluids and tissues, 132

crash survivability, 130–131

deformations caused by occupant

contact, 131–132

steering wheel, seats, and restraints, 132–134

survival space, 130

whole-body acceleration tolerance, 117–120

Injury morbidity surveillance

classification of

clinical modifications, 30

external cause of injury codes, 30–31

ICD, 30

nature of injury codes, 31

confidentiality concerns, 39–40

data linkage, 39

data presentation and dissemination

analytic issues, 35

injury indicators, 35

rates and population coverage, 36

sample weights and variance estimation, 35–36

standard publications, micro-data and online resources, 37

trends, 36–37

data sources

cross-sectional surveys, 28

emergency department data, 25–26

Fatality Analysis Reporting System, 7, 12, 15, 68

health care provider-based data, 25

hospital inpatient data, 26

longitudinal surveys, 28–29

NASS-GES, 29

population-based data, 27–28

trauma registries, 26–27

electronic medical records/electronic health records, 39

factors affecting case definitions of

external cause vs. diagnosis of injury, 32–33

identifying injury events, 32

injury incidence, 32

injury severity, 33–34

primary vs. multiple diagnoses, 33

unit of analysis, 31

narrative text, 38

supplements to surveillance systems, 38

surveillance system evaluation, 38

timeliness of data release, 39

Injury mortality surveillance

advances in less resourced environments, 18–19

classification of

ICD-coded data, matrices for, 10

International Classification of Diseases, 8–10

place and activity at time of injury, 10

data sources

from coroners and medical examiners, 6

multiple, systems based on, 7

supplementary, 7–8

vital records, 4–6

dissemination

analytic issues, 13–14

injury indicators, 13

standard publications and web-based, 14

limitations, 4

linking data sources, 18

narrative text, 18

operational definitions

late deaths, 12

unit of analysis, 11–12

using ICD, 11

purpose, 3–4

systems evaluation and enhancements

supplements, 16–17

vital statistics mortality data,

quality of, 15–16

timely vital data, 17–18

vital statistics data quality, 18

Injury prevention law

effectiveness, 499–500

legal validity, 497–499

political considerations, 500–502

Injury profiling

abbreviated injury scale, 269

Barell matrix

background, 269–270

matrix for AIS, 273

original ICD-9-CM matrix, 270–273

coding and classification, 267–268

ICD codes, 268–269

injury severity assessment, 276–277

Injury profiling (*cont.*)
 multiple injury profiles
 background, 273
 conceptual framework, 273–275
 injury epidemiology, 276

Injury severity, 116–117
 assessment, 276–277
 scaling
 acronyms and full names of, 281
 AIS scale, 283–288
 anatomic profile, 289
 dimensions, 280–283
 future development, 290–291
 Glasgow Coma Scale, 289
 Harborview assessment for risk of mortality score, 290
 injury severity score, 288
 Low- and middle-income countries, 599–618
 KABCOU, 283
 maximum AIS, 288
 revised trauma score, 289
 survival risk ratios, 289–290
 trauma and injury severity score, 289

Injury Severity Score (ISS), 288

Intercellular adhesion molecule-1 (ICAM-1), 346

Interleukin-6 (IL-6), 338

International classification of diseases (ICD)
 external cause of injury, 9
 ICECI, 8–9
 matrices used in, 10
 nature of injury, 10
 operational definitions of injury, 11
 place and activity at time of injury, 10

International classification of external causes of injury (ICECI), 8–9

International disparities, 245

Inter-observer reliability, 326

Inter-rater reliability, 326

Interventions
 implementation and evaluation
 appreciation requirement, 623–625
 effectiveness vs. efficacy, 622–623
 innovation, diffusion, 629–632
 mapping, 625–629
 RE-AIM framework, 632–634
 reason for, 619–621
 translation of, finding, 635–636

low-income countries
 effectiveness maximize, 607–608
 factors in, 605–607
 level of effectiveness, 600–602
 transferred interventions, 608–615
 transferring interventions, challenges, 602–604

Intestinal-type fatty acid binding protein (I-FABP), 345

K
 KABCOU, 117, 283
 Kidney injury molecule-1 (KIM-1), 344
 Krebs von den Lungen-6 (KL-6), 346

L

Laceration, 98
 Late deaths, 12
 Legal approach
 injury prevention law
 effectiveness, 499–500
 legal validity, 497–499
 political considerations, 500–502
 law usage, in injury reduction, 495–496
 tort law, 502–504
 Legislation, 246
 Life course epidemiology, 241
 Liver-type fatty acid binding protein (L-FABP), 344
 Localized muscle fatigue (LMF), 169
 Longitudinal surveys, 28–29
 Low- and middle-income countries, 599–618

M

MADYMO, 129
 Malpractice claims, 588
 Medical examiner (ME), 6
 Medical examiner system, 90
 Member checking, 228
 Metabolic energy expenditure analysis, 172–177
 Metabolic energy prediction models, 168–169
 Motor-vehicle-related fatalities, 68
 Multilevel modeling
 advantages, 440–441
 in injury mortality, 429–431
 logit command, 437–440
 shrinkage coefficient, 433–436
 single-level modeling, injury mortality, 426–429
 software used, 431–433
 Multiple Cause of Death Data (MCDD), 66
 Myelin basic protein (MBP), 338

N

N-acetylaspartate (NAA), 340
 Nagi's disablement framework, 360–361
 Narrative text, 18
 injury morbidity surveillance, 38
 National Automotive Sampling System-General Estimates System (NASS-GES), 29
 National Center for Health Statistics (NCHS), 37
 National Electronic Injury Surveillance System-All Injury Program (NEISS-AIP), 25–26, 67
 National Fire Incident Reporting System (NFIRS), 29
 National Health and Nutrition Examination Survey (NHANES), 17
 National Health Interview Survey (NHIS), 17
 National Highway Traffic Safety Administration (NHTSA), 260
 National Hospital Ambulatory Medical Care Survey (NHAMCS), 25, 67
 National Hospital Discharge Survey (NHDS), 66
 National Study of the Costs and Outcomes of Trauma (NSCOT), 577

National Trauma Data Bank, 572–574
 National Violent Death Reporting System (NVDRS), 6, 7, 46–47
 National Vital Statistics System (NVSS), 66
 Nationwide Inpatient Sample (NIS), 66
 Natural death, 92
 Neurobehavioral evaluation system 3 (NES3), 196
 Neurocognitive performance measurement, 195–196
 Neuroimaging, 80
 Neuron-specific enolase (NSE), 337
 Neutrophil gelatinase-associated lipocalin (NGAL), 344
 Nonstatic forces, 154–155
 NTDB National Trauma Data Bank, 66

O

Occupant kinematics, 127–129
 Outcome measurement
 activity and participation domains, 362–363
 causal relationships, ICF, 362
 evidence-based public health (EBPH), 355
 future of, 363–365
 goals, 356–357
 international classification of function, 360, 361
 measurement validity, 357–359
 Nagi's disablement framework, 360–361
 Wood's disablement framework, 361
 Overexertion injuries
 exertion stress-strain models
 biomechanical models, 158–161
 confounders, 162
 falls, 162–165
 hernias, 161–162
 overuse syndromes, 166
 vehicles and impact injuries, 165
 whole-body vibration, 165
 force, work, and power
 nonstatic forces, 154–155
 segment rotations create additional forces, 155–156
 physical work capacity and workload analysis
 localized muscle fatigue, 169
 metabolic energy prediction models, 168–169
 systemic fatigue, 166–168
 thermal stress, 170–171
 Over-triage, 303
 Overuse syndromes, 166

P

Passive protection devices, 254
 Patient safety, 583–598
 Pedestrian injuries, 560–561
 Physical environment, key aspects of
 global environmental change, 243
 housing and home environment, 241–242
 travel environment, 242
 work environment, 242–243
 Poison prevention, 558–559, 614

Policies, 246
 Polymorphonuclear elastase, 346
 Population-based data, 27–28
 Postmortem toxicological analysis, 95–96
 Prediction rule derivation
 outcome assessment, 323–328
 outcome definition, 322–323
 population, 320
 predictor variables, 320–321
 prospective data, advantages of, 319
 variables, 321–322
 Primary trauma triage, 299–300
 Primary vs. multiple diagnoses, 33
 Procalcitonin (PCT), 345
 Proportional reporting ratio (PRR), 592
 Prothrombin time, albumin, and globulins, 344
 Psychomotor vigilance task (PVT), 196
 Public policy
 consumer nonrationality, 510–511
 enforcement, 513–514
 externalities, 509–510
 imperfect information, 508–509
 monitoring, 513
 public goods, 511–512
 public health approach, 514
 rationale for, 507
 regulations, 512–513
 residential smoke detectors, 508
 rules, 513
 ski boots and bindings, 508
 suicide, 514–515

Q

Qualitative research methods
 aim, 222–223
 concept mapping, 230
 description, 222
 five ideas, 231
 in injury research, 223
 qualitative data analysis software, 228
 quality of, 229
 rules of analysis, 228–229
 tools and rules in
 data analysis, 229
 data collection, 224–226
 sampling qualitative data, 226–227
 traditions of, 229–230
 Quasi-experiments, 191–192
 Quasi-static exertion, 160

R

RE-AIM Framework, 632–634
 Receiver operating characteristic curves (ROCs), 145–146
 Receptor for advanced glycation end-products (RAGE), 346
 Recognition threshold (RL), 143
 Regression analysis, 324

Restraint systems

- air bag systems, 123–124
- belt restraint systems, 121
- design, 122–123
- multiple functions of, 122
- principles, 122

Revised trauma score (RTS), 289

Root cause analysis (RCA), 590–591

S

Safety engineering. *See also* Technological approach
development, 534–537

- crash/event investigation, 531–532
- physical/computational modeling, 532–534

Safety-related performance measurement,
192–195

Scald burn risk, 557–558

Seat examination, 133

Secondary trauma triage, 300–301

Serum biomarkers, 80

Sharp force injury

- chop wound, 101–102
- described, 99
- incised wound, 99–100
- stab wound, 100–101

Shotgun wounds, 105–106

Signal detection theory (SDT), 143–145

Skeletal fracture, 98–99

Sleep quality measurement, 197–198

Social environment, key aspects of

- family and community-level, 245
- international disparities, 245
- policies and legislation, 246
- socioeconomic status and ethnicity, 244

Social network analysis

- attributes
 - individual, 477
 - network, 477–478
- complete network, 479
- data analyses, 478
- described, 476–477
- dyadic network, 478
- egocentric network, 478
- as methodological tool, 475–476
- triadic network, 478
- types, 476
- youth violence, public health issue
 - adolescent development, 479–480
 - methodology, 482–488
 - preliminary findings, 481–482
 - risk and protective factors, 480–481

Socioeconomic status, 244

Spatial regression

- analysis, 470–471
- definition, 464
- GWR, injury research, 471–472
- linear regression *vs.* GWR, 468–469
- modeling, 466–468

Special populations

- capturing, 48–49
- community stakeholders, forging
and maintaining, 56–57
- definition, 47
- feasibility, funds, and framework, 49–51
- knowledge gain of, 57
- public health importance, 58
- statistical aspects
 - elegant simplicity, 56
 - growing options, 55–56
 - positives and negatives, 56
- technology
 - elegant simplicity, 54–55
 - growing options, 52–53
 - infodemiology, 51
 - national computer-based surveillance systems, 51
 - positives and negatives, 53–54
 - “Twitter” surveys, 51
 - understanding culture of, 57–58
 - variables, 48

Sprinkler systems, 613

Stab wound, 100–101

State-based hospital discharge data (HDD), 67

Statistical considerations

- quantifying probability theoretically, 384–388
- quantifying relationship, 388–393
- role, 381–383

Statistical validation of regression models, 325

Steering wheel, 132–133

Suicide, 92–93

Surfactant protein D (SP-D), 346

Systemic fatigue, 166–168

Systems approach, patient safety

- adverse event analysis
 - analysis method, 590–591
 - risk prioritization methods, 591–593
- challenges, 584
- conceptual framework, 584–585
- hazard identification
 - adverse event, 588–589
 - direct observation, 587
 - failure mode and effect analysis, 589–590
 - malpractice claims, 588
 - medical record, 585–587
 - provider surveys/focus groups/patient
interviews, 588
 - simulation, 589
- injury prevention and mitigation, 593–594

T

Task-human-environment-machine (THEM)
system, 139–141

Technological approach

- concepts, 530
- crash prevention engineering:
 - active safety systems, 541–542
 - driver and occupant factors, 541

- factors, 540–541
 - timeline, 539
 - types, in USA, 539–540
- safety engineering
 - countermeasure development, 534–537
 - crash/event investigation, 531–532
 - physical/computational modeling, 532–534
- Terminal threshold (TL), 143
- Thematic analysis, 228
- Thermal stress, 170–171, 178–179
- Topic coding, 227
- Tort law, 502–504
- Toxicological Analysis, 94–96
- Traffic calming, 559
- Transdermal scopolamine, 189–190
- Transforming growth factor beta (TGF- β), 338–339
- Trauma and injury severity score (TRISS), 289
- Trauma center care, 297
- Trauma registries, 26–27, 572
- Traumatic brain injury (TBI)
 - administrative data sets, 68–70
 - administrative data systems
 - ambulatory medical care, 67–68
 - Barell matrix, 70
 - cost, 75
 - data sources, 64, 66
 - definitions, 63–65
 - disability, 72–73
 - emergency department data, 67
 - ICD-10-CM, 64, 65
 - ICD-9-CM code-based definition, 64, 65
 - ICD-9 diagnosis codes, 63–64
 - incidence, 71
 - morbidity, 66–67
 - mortality, 66
 - prevalence, 71–72
 - Barell matrix for, 70
 - biomarker signature, 337
 - clinical case definitions
 - Abbreviated Injury Scale (AIS), 62
 - anterograde amnesia, 62
 - clinical criteria, 63
 - Glasgow Coma Scale (GCS), 61–62
 - description, 336
 - epidemiologic measures, 70–71
 - future directions in, 80
 - genetic polymorphism
 - ApoE, 340
 - cytochrome P450, 341–342
 - symptoms, 341
 - in treatment, 341
 - helmet sensors, 80
 - lifetime prevalence of, 71–72
 - metabolites
 - catecholamines, 340
 - F2-isoprostane, 339
 - glycolytic intermediates, 340
 - 4-hydroxynonenal (4-HNE), 339
 - N-acetylaspartate (NAA), 340
 - in military personnel and veterans
 - administrative health-care data, 77–79
 - clinical case definition, 76–77
 - motor-vehicle-related fatalities, 68
 - mRNA and miRNA, 342–344
 - neuroimaging, 80
 - outcomes
 - disability, 72–75
 - economic cost, 75
 - late mortality, 75
 - long-term adverse health, 72
 - protein biomarkers
 - chemokines, 339
 - cleaved tau (C-Tau), 337
 - glial fibrillary acidic protein (GFAP), 337–338
 - interleukin-1 (IL-1), 338
 - interleukin-6 (IL-6), 338
 - myelin basic protein (MBP), 338
 - neuron-specific enolase (NSE), 337
 - transforming growth factor beta (TGF- β), 338–339
 - ubiquitin carboxyl-terminal esterase-L1 (UCHL1), 338
 - serum biomarkers, 80
- Trauma triage
 - ACSCOT field decision scheme
 - anatomic criteria, 302
 - mechanism and risk factor criteria, 302–303
 - physiologic criteria, 301–302
 - cost implications, 308
 - field, accuracy of, 304
 - field provider cognitive reasoning, 308
 - field trauma, 295–297
 - finite trauma resources, 297–298
 - future directions, 309
 - history, 295
 - limitations
 - data quality, field triage criteria, 305
 - full clinical decision rule methodology, 306
 - out-of-hospital injury population, 305
 - study design, 304–305
 - target variability, 305–306
 - timing, 306
 - 2006 National trauma triage protocol, 296
 - over-triage, 303
 - populations
 - children, 307
 - elders, 307
 - rural patients, 307–308
 - primary, 299–300
 - secondary, 300–301
 - target population, 298–299
 - trauma center care, 297
 - under- and over-triage, balance between, 304
 - under-triage, 303
- Travel environment, 242
- Trefoil factor 3 (TEF3), 344
- Troponins, 346

U

Ubiquitin carboxyl-terminal esterase-L1 (UCHL1), 338
Under-triage, 303
Undetermined cause of death, 94

V

Vehicle, injury mechanisms
 body fluids and tissues, 132
 crash survivability, 130–131
 deformations caused by occupant
 |contact, 131–132
 steering wheel, seats, and restraints, 132–134
 survival space, 130
Video data analysis
 conceptualization, 398
 data analysis, 402–403
 data collection, 402
 instrument development, 399–401
 methodological considerations, 396–398

resources, 398–399

sports injury research, 395–396
strengths and limitations, 403–404

Vital records, 4–6

Vital statistics mortality data, 15–16, 18

W

Web-based injury statistics query and reporting system
 (WISQARS), 14

Weber's law or fraction, 143

Wet-bulb-globe-thermometer (WBGT), 170–171

Whole-body acceleration tolerance, 117–120

Whole-body vibration, 165

Wood's disablement framework, 361

Work environment, 242–243

Y

Youth violence, 479–489