Integrated Series in Information Systems 27

Series Editors: Ramesh Sharda · Stefan Voß

Daniel Zeng
Hsinchun Chen
Carlos Castillo-Chavez
William B. Lober
Mark Thurmond *Editors*

Infectious Disease Informatics and Biosurveillance

Springer

Integrated Series in Information Systems

Volume 27

For other titles published in this series; go to
http://www.springer.com/series/6157

INFECTIOUS DISEASE INFORMATICS AND BIOSURVEILLANCE

Editors
Daniel Zeng
Hsinchun Chen
Carlos Castillo-Chavez
William B. Lober
Mark Thurmond

Springer

Editors
Daniel Zeng
Chinese Academy of Sciences
Institute of Automation
Beijing, China
Department of MIS, Eller College of Management
University of Arizona, Tucson Arizona, USA
zengdaniel@gmail.com

Hsinchun Chen
Eller College of Management
University of Arizona
E. Helen St. 1130
85721 Tucson Arizona
430Z McClelland Hall
USA
hchen@eller.arizona.edu

Carlos Castillo-Chavez
Department of Mathematics and Statistics
Arizona State University
Tempe Arizona
USA
chavez@math.asu.edu

William B. Lober
Health Sciences Center
University of Washington
NE. Pacific St. 1959
98195 Seattle Washington
USA
lober@u.washington.edu

Mark Thurmond
Department of Medicine and Epidermiology
School of Veterinary Medicine
University of California, Davis
Shields Avenue 1
95616 Davis California
USA
mcthurmond@ucdavis.edu

ISSN 1571-0270
ISBN 978-1-4419-6891-3 e-ISBN 978-1-4419-6892-0
DOI 10.1007/978-1-4419-6892-0
Springer New York Dordrecht Heidelberg London

Printed on acid-free paper

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

PREFACE

Information systems are central to the development of effective comprehensive approaches aimed at the prevention, detection, mitigation, and management of human and animal infectious disease outbreaks. Infectious disease informatics (IDI) is a subfield of biomedical informatics concerned with the development of methodologies and technologies needed for collecting, sharing, reporting, analyzing, and visualizing infectious disease data and for providing data-driven decision-making support for infectious disease prevention, detection, mitigation, and management. The growth and vitality of IDI are central to our national security. Biosurveillance is an important partner of IDI applications and focuses primarily on the early detection of new outbreaks of infectious diseases and on the early identification of elevated or new diseases' risks.

IDI and biosurveillance research directly benefits public health and animal health agencies in their multiple activities in fighting and managing infectious diseases. IDI and biosurveillance research provides quantitative methods and computational tools that are instrumental in the decision-making process carried out by government agencies with responsibilities in infectious diseases within national and international contexts. IDI also has important applications in law enforcement and national security concerning, among other issues, the prevention of and timely response to the deliberate release of biological agents. As a result of the increasing threats to our national security, a large amount of animal and public health infectious disease data are being collected by various laboratories, health care providers, and government agencies at local, state, national, and international levels. In fact, many agencies charged with collecting these data have developed information access, analysis, and reporting systems of varying degrees of sophistication. Researchers from a wide range of backgrounds including but not limited to epidemiology, statistics, applied mathematics, computer science and machine learning/data mining, have contributed to the development of technologies that facilitate real-time data collection and access. They have also developed algorithms needed to analyze or mine the collected data.

This book on IDI and biosurveillance compiles a high-quality collection of academic work in various sub-areas of IDI and biosurveillance to provide an integrated and timely view of the current state-of-the-art. It also identifies technical and policy challenges and opportunities with the goal of promoting cross-disciplinary research that takes advantage of novel methodology and lessons learned from innovative applications. This book fills a systemic gap in the literature by emphasizing informatics-driven perspectives (e.g., information system design, data standards, computational aspects of biosurveillance algorithms, and system evaluation) rather than just statistical modeling and analytical work. Finally, this book attempts to reach policy makers and practitioners through the clear and effective communication of recent research findings in the context of case studies in IDI and biosurveillance, providing "hands-on" in-depth opportunities to practitioners to increase their understanding of value, applicability, and limitations of technical solutions.

SCOPE AND ORGANIZATION

This volume collects the state-of-the-art research and modern perspectives of distinguished individuals and research groups on cutting-edge IDI technical and policy research and its application in biosurveillance. The contributed chapters are grouped into three units.

Unit I provides an overview of recent biosurveillance research while highlighting the relevant legal and policy structures in the context of ongoing IDI and biosurveillance activities. It also identifies IDI data sources and addresses information collection, sharing, and dissemination issues, as well as ethical considerations.

Unit II consists of chapters that survey various types of surveillance methods used to analyze IDI data in the context of public health and bioterrorism. Specific computational techniques covered include: text mining, time series analysis, multiple data streams methods, ensembles of surveillance methods, spatial analysis and visualization, social network analysis, and agent-based simulation.

Unit III examines IT and decision support for public health event-response and bio-defense. Included are discussions of practical lessons learned in developing public health and biosurveillance systems, technology adoption, and syndromic surveillance for large events.

These three units include the following chapters:

Unit I: Informatics Infrastructure and Surveillance Data Sources

- Real-time Public Health Biosurveillance: The chapter surveys recent public health biosurveillance efforts, highlights related legal and policy considerations, and shares insights on various constraints biosurveillance system designers need to consider.
- Designing Ethical Practice in Biosurveillance: The chapter draws upon experience and lessons learned through Project Argus and presents ethical and legal dimensions of biosurveillance systems design and operations.

- Using Emergency Department Data for Biosurveillance: The chapter presents benefits and challenges of using Emergency Department data for IDI and biosurveillance. Detailed examples from a well-known biosurveillance system, NC DETECT, are presented.
- Clinical Laboratory Data for Biosurveillance: The chapter provides an overview of the types of data used for IDI and biosurveillance, and discusses in detail clinical laboratory data as a data source for biosurveillance and related data sharing and analysis issues.
- Biosurveillance based on Test Orders from Veterinary Diagnostic Labs: The chapter discusses the use of tests orders made to veterinary diagnostic laboratories as a biosurveillance data source. It also shares insights concerning outbreak detection and biosurveillance involving zoonotic pathogens.

Unit II: Surveillance Analytics

- Markov Switching Models for Outbreak Detection: The chapter presents an outbreak detection model using syndrome count-based time series. This model is rooted in Markov Switching models and possesses many desirable computational properties.
- Detection of Events in Multiple Streams of Surveillance Data: The chapter reviews analytic approaches that can be used to simultaneously monitor multiple data streams. Both multivariate methods and more recent methods that do not assume joint models of multiple data streams are presented.
- Algorithm Combination for Improved Performance in Biosurveillance: The chapter introduces a new outbreak detection scheme that is based on ensembles of existing algorithms. The IDI application of this scheme is demonstrated through monitoring daily counts of pre-diagnostic data.
- Modeling in Space and Time: The chapter presents an open-source IDI and biosurveillance software system, the Spatial-temporal Epidemiological Modeler (STEM), as a collaborative platform to define and visualize simulations of infectious disease spreading.
- Surveillance of Infectious Diseases Using Spatial and Temporal Clustering Methods: This chapter surveys common temporal, spatial, and spatio-temporal clustering methods and discusses how such methods can be used for outbreak detection, disease mapping, predictive modeling.
- Age-adjustment in National Biosurveillance Systems: The chapter presents population surveillance as a subarea of biosurveillance, with a particular emphasis on age and age-adjustment. Both data sources available for population surveillance and related analytical tools are discussed.

- Modeling in Immunization and Biosurveillance Research: The chapter presents mathematical modeling techniques suitable for applications concerning vaccine-preventable diseases. Issues concerning vaccination modeling and the interface between biosurveillance and public health response to vaccine-preventable diseases are also discussed.
- Natural Language Processing for Biosurveillance: The chapter presents various types of national language processing techniques that have been applied to outbreak detection and characterization. Four common classes of textual data associated with healthcare visits are presented along with the applicable data processing techniques.
- Knowledge Mapping for Bioterrorism-related Literature: The chapter introduces major knowledge mapping techniques, focusing on text mining and citation network analysis methods. A case study on bioterrorism-related literature is presented.
- Social Network Analysis for Contact Tracing: The chapter illustrates how social network analysis techniques can contribute to epidemiological investigations and public health policy evaluation. A case study using the 2003 Taiwan SARS outbreak is presented.

Unit III: Decision Support and Case Studies

- Multi-Agent Modeling of Biological and Chemical Threats: The chapter presents a city-level dynamic-network model based on multi-agent systems technology, BioWar, as a computational tool to assess public health and biosecurity policies. A case study on using BioWar to assess the impact of school closures and quarantine when facing pandemic influenza is presented.
- Integrated Health Alerting and Notification: The chapter discusses a detailed case study concerning design and operation of a state-wide integrated health alerting and notification system.
- Design and Performance of a Public Health Preparedness Informatics Framework: The chapter discusses a model informatics framework aimed at supporting public health emergency preparedness and presents an evaluative study assessing this framework during a full-scale exercise simulating an influenza outbreak.
- System Evaluation and User Technology Adoption: The chapter highlights the importance of conducting system evaluation and user studies with the objective of promoting advanced IDI systems in field adoption. Two empirical studies are presented along with detailed discussions on evaluation and adoption research design, and measurement instruments.
- Syndromic Surveillance for the G8 Hokkaido Toyako Summit Meeting: The chapter reports a detailed international case study on conducting syndromic surveillance for a major event.

AUDIENCE

The goal of this book is to provide an accessible interdisciplinary IDI and biosurveillance volume that serves as a reference or as a stand-alone textbook or as a supplemental text. Upper-level undergraduates and graduate-level students from a variety of disciplines including but not limited to public health, veterinary medicine, biostatistics, information systems, computer science, and public administration and policy, will benefit from learning the concepts, techniques, and practices of IDI and biosurveillance.

Researchers, including both IDI researchers and public health/IT/public policy researchers who have an interest in IDI, will find in this book a comprehensive source of reviews of the recent advances in the field. This book is intended to help further define the field as a reference book and promote community development across disciplines and between academia and the practitioners, given the dynamic nature of current IDI and biosurveillance research.

This book provides an up-to-date review of current IDI and biosurveillance research and practice, critical evaluation of current approaches, and discussion of real-world case studies and lessons learned. The information and perspective presented should prove their utility to epidemiologists in public health and veterinary health departments and private-sector practitioners in healthcare and health IT.

We would like to acknowledge support from the following U.S. National Science Foundation grants: NSF Information Technology Research Grant IIS-0428241, "ITR-(NHS)-(DMC): A National Center of Excellence for Infectious Disease Informatics;" NSF Grant IIS-0839990, "Transnational Public Health Informatics Research: US-China Collaboration;" and NSF IIS Grant IIS-0748308, "US/China Digital Government Collaboration: US-China Infectious Disease Informatics and BioSurveillance Workshop." We also acknowledge funding support from the U.S. Department of Homeland Security through Grant 2008-ST-061-BS0002 ("DHS Center of Excellence in Border Security and Immigration") and from the National Center for Medical Intelligence.

Our work has been supported by a number of research grants in China as well: the National Natural Science Foundation of China through grants 90924302 and 60621001; the Chinese Academy of Sciences through grants 2F07C01 and 2F08N03; the Ministry of Science and Technology through grants 2006CB705500 and 2006AA010106; and the Ministry of Health through grants 2009ZX10004-315 and 2008ZX10005-013.

TABLE OF CONTENTS

Unit I: Informatics Infrastructure and Data Sources

Chapter 4: Clinical Laboratory Data for Biosurveillance

EILEEN KOSKI

Chapter 5: Biosurveillance Based on Test Orders from Veterinary Diagnostic Labs

LOREN SHAFFER

Unit II: Surveillance Analytics

Chapter 7: Detection of Events In Multiple Streams of Surveillance Data: Multivariate, Multi-stream and Multi-dimensional Approaches

ARTUR DUBRAWSKI

Chapter 8: Algorithm Combination for Improved Performance in Biosurveillance: Univariate Monitoring

INBAL YAHAV, THOMAS LOTZE AND GALIT SHMUELI

Chapter 14: Knowledge Mapping for Bioterrorism-Related Literature
YAN DANG, YULEI ZHANG, HSINCHUN CHEN AND CATHERINE A. LARSON

LIST OF CONTRIBUTORS

KATHLEEN M. CARLEY, Center for Computational Analysis of Social and Organizational Systems, School of Computer Science, ISR – Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213, USA

NEAL ALTMAN CASOS, Center for Computational Analysis of Social and Organizational Systems, School of Computer Science, ISR – Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213, USA

TA-CHIEN CHAN, College of Public Health, Institute of Epidemiology, National Taiwan University, 17 Xu-Zhou Road, Taipei (100),Taiwan

WENDY W. CHAPMAN, Department of Biomedical Informatics, University of Pittsburgh, 200 Meyran Avenue, Pittsburgh, PA 15260, USA

HSINCHUN CHEN, Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA

YI-DA CHEN, Department of Management Information Systems, The University of Arizona, Tucson, AZ 85721, USA

STEVEN A. COHEN, Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, USA
Tufts University Initative for the Forecasting and Modeling of Infectious Diseases (InForMID), Tufts University, 136 Harrison Ave, Boston, MA 02111, USA, steven_a.cohen@tufts.edu

JEFF COLLMANN, O'Neill Institute for National and Global Health Law, School of Nursing and Health Studies, Georgetown University Medical Center, Associate Professor and Director, Disease Prevention and Health Outcomes, 3700 Reservoir Rd, NW, Box 571107, Washington, DC 20057–1107, USA, collmanj@georgetown.edu

YAN DANG, Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA

JOHN N. DOWLING, Department of Biomedical Informatics, University of Pittsburgh, 200 Meyran Avenue, Pittsburgh, PA 15260, USA

ARTUR DUBRAWSKI, Auton Lab, Carnegie Mellon University, Pittsburgh, PA 15213, USA

DANIEL A. FORD, Research Staff Member, Healthcare Informatics Research, Department of Computer Science, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA

ZHANHAI GAO, School of Public Health and Community Medicine, Faculty of Medicine, UNSW, Australia

IVAN J. GOTHAM, Bureau of Healthcom Network Systems Management, New York State Department of Health, Empire State Plaza, Room 148, Albany, NY 12237, USA
School of Public Health, State University at Albany, Albany, NY 12222, USA

ADI V. GUNDLAPALLI, Division of Epidemiology, School of Medicine, University of Utah, 30 North 1900 East, AC230A, Salt Lake City, UT 84132, USA

MARY E. HENNESSEY, New York State Department of Health, Empire State Plaza, Albany, NY 12237, USA

TAKU HORIE, Chiimiya-Horie Clinic, Japan

PAUL JEN-HWA HU, Department of Operations and Information Systems, David Eccles School of Business, University of Utah, Salt Lake City, UT 84132, USA

AMY I. ISING, Carolina Center for Health Informatics, Department of Emergency Medicine, University of North Carolina at Chapel Hill, 100 Market Street, Chapel Hill, NC, 27516, USA, isina@med.unc.edu

GERALDINE S. JOHNSON, New York State Department of Health, Empire State Plaza, Albany, NY 12237, USA

JAMES H. KAUFMAN, Research Staff Member, Healthcare Informatics Research, Department of Computer Science, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA

KIYOSHI KIKUCHI, Shimane Prefectural Central Hospital, Japan.

CHWAN-CHUEN KING, College of Public Health, Institute of Epidemiology, National Taiwan University, 17 Xu-Zhou Road, Taipei (100), Taiwan, chwanchuen@gmail.com

KAZUO KODAMA, Kodama Clinic, Japan

EILEEN KOSKI, M.Phil., 875 West 181st Street, Apt. 6M, New York, NY 10033, USA, EileenKoski@gmail.com

CATHERINE A. LARSON, Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA

LINH H. LE, Bureau of Healthcom Network Systems Management, New York State Department of Health, Empire State Plaza, Room 148, Albany, NY 12237, USA

THOMAS LOTZE, Applied Mathematics & Scientific Computation Program, University of Maryland, College Park, MD 20742, USA

HSIN-MIN LU, Department of Information Management, College of Management, National Taiwan University, Taipei, Taiwan. luim@ntu.edu.tw

C. RAINA MACINTYRE, School of Public Health and Community Medicine, Faculty of Medicine, UNSW, Australia

ERIC MALLOY, Center for Computational Analysis of Social and Organizational Systems, School of Computer Science, ISR – Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213, USA

YOSSI MESIKA, Research Staff Member, Healthcare and Life Sciences, IBM Haifa Research Lab, Haifa University Campus, Mount Carmel, Haifa, 31905, Israel

ELENA N. NAUMOVA, Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, USA
Tufts University Initative for the Forecasting and Modeling of Infectious Diseases (InForMID), Tufts University, 136 Harrison Ave, Boston, MA 02111, USA, elena.naumova@tufts.edu

JEAN O'CONNOR, Office of Critical Information Integration and Exchange, National Center for Zoonotic, Enteric, and Vector-borne Diseases, Centers for Disease Prevention and Control, MS-D68, 1600 Clifton Road, Atlanta, GA 30333, USA, jgo6@cdc.gov

YASUSHI OHKUSA, Infectious Disease Surveillance Center, National Institute of Infectious Disease, Japan

NOBUHIKO OKABE, Infectious Disease Surveillance Center, National Institute of Infectious Disease, Japan

STEPHANIE E OSTROWSKI, New York State Department of Health, Empire State Plaza, Albany, NY 12237, USA

MICHAEL J. PRIMEAU, New York State Department of Health, Empire State Plaza, Albany, NY 12237, USA

ADAM ROBINSON, The MITRE Corporation, Portfolio Division Manager, 7515 Colshire Drive, McLean, VA 22102, USA, arobinson@mitre.org

HENRY ROLKA, Office of Critical Information Integration and Exchange, National Center for Zoonotic, Enteric, and Vector-borne Diseases, Centers for Disease Prevention and Control, MS-D68, 1600 Clifton Road, Atlanta, GA 30333, USA, HRolka@cdc.gov

LORETTA A. SANTILLI, New York State Department of Health, Empire State Plaza, Albany, NY 12237, USA

MATTHEW SCHOLER, Carolina Center for Health Informatics, Department of Emergency Medicine, University of North Carolina at Chapel Hill, 100 Market Street, Chapel Hill, NC, 27516, USA, mscholer@med.unc.edu

LOREN SHAFFER, The Ohio State University, College of Veterinary Medicine, 1920 Coffey Road, Columbus, OH 43210

GALIT SHMUELI, Department of Decision, Operations & Information Technologies, Robert H Smith School of Business, University of Maryland, College Park, MD 20742, USA

DEBRA L. SOTTOLANO, Bureau of Healthcom Network Systems Management, New York State Department of Health, Empire State Plaza, Room 148, Albany, NY 12237, USA

BRETT R. SOUTH, Division of Epidemiology, School of Medicine, University of Utah, 30 North 1900 East, AC230A, Salt Lake City, UT 84132, USA

TAMIE SUGAWARA, The Infectious Disease Surveillance Center, the National Institute of Infectious Diseases, Japan

HIROAKI SUGIURA, Sugiura Clinic, Japan

KIYOSU TANIGUCHI, Infectious Disease Surveillance Center, National Institute of Infectious Diseases, Japan

DEBBIE A. TRAVERS, Carolina Center for Health Informatics, School of Nursing, University of North Carolina at Chapel Hill, 100 Market Street, Chapel Hill, NC 27516, USA, dtravers@email.unc.edu

ANNA E. WALLER, Carolina Center for Health Informatics, Department of Emergency Medicine, University of North Carolina at Chapel Hill, 100 Market Street, Chapel Hill, NC 27516, USA, awaller@med.unc.edu

ROCHELLE WATKINS, Faculty of Health Science, Curtin University of Technology, Australia

JAMES G. WOOD, School of Public Health and Community Medicine, Faculty of Medicine, UNSW, Australia

INBAL YAHAV, Department of Decision, Operations & Information Technologies, Robert H Smith School of Business, University of Maryland, College Park, MD 20742, USA

DANIEL ZENG, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; Management Information Systems Department, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA

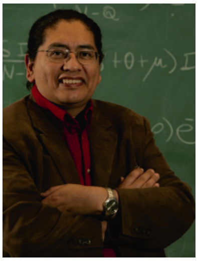
YULEI ZHANG, Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA

EDITORS' BIOGRAPHIES

Daniel Zeng received the M.S. and Ph.D. degrees in industrial administration from Carnegie Mellon University and the B.S. degree in economics and operations research from the University of Science and Technology of China, Hefei, China. He is a Research Professor at the Institute of Automation in the Chinese Academy of Sciences and a Professor and Honeywell Fellow in the Department of Management Information Systems at the University of Arizona. Zeng's research interests include intelligence and security informatics, spatial-temporal data analysis, infectious disease informatics, social computing, recommender systems, software agents, and applied operations research and game theory with application in e-commerce and online advertising systems. He has published one monograph and more than 170 peer-reviewed articles. He has also co-edited 14 books and conference proceedings, and chaired 25 technical conferences or workshops including the IEEE International Conference on Intelligence and Security Informatics (ISI), the Biosurveillance and Biosecurity Workshop (BioSecure), and the International Workshop on Social Computing (SOCO). He serves on editorial boards of 15 Information Technology and Information Systems related journals. He is also active in information systems and public health informatics professional activities and is Vice President for Technical Activities for the IEEE Intelligent Transportation Systems Society and Chair of INFORMS College on Artificial Intelligence. His research has been mainly funded by the U.S. National Science Foundation, the National Natural Science Foundation of China, the Chinese Academy of Sciences, the U.S. Department of Homeland Security, the Ministry of Science and Technology of China, and the Ministry of Health of China.

Hsinchun Chen is the McClelland Professor of Management Information Systems at the University of Arizona, where is also the (founding) director of the University of Arizona Artificial Intelligence Lab and the Hoffman E-Commerce Lab. He received his B.S. degree from National Chiao-Tung University in Taiwan, MBA degree from SUNY Buffalo, and Ph.D. degree in Information Systems from New York University. He is the

Editor-in-Chief (EIC) of the new journal ACM Transactions on Management Information Systems (ACM TMIS) as well as of the new Springer journal, Security Informatics; he also serves as the Associate EIC of IEEE Intelligent Systems. He serves on ten editorial boards including: ACM Transactions on Information Systems; IEEE Transactions on Systems, Man, and Cybernetics; Journal of the American Society for Information Science and Technology; Decision Support Systems; and the International Journal on Digital Libraries. He is the author/editor of 20 books, 25 book chapters, 200 SCI journal articles, and 130 refereed conference articles covering Web computing, search engines, digital library, intelligence analysis, biomedical informatics, data/text/web mining, and knowledge management. He has served as the conference/program co-chair for many of the past International Conferences of Asian Digital Libraries (ICADL) and is also the (founding) conference co-chair of the IEEE International Conferences on Intelligence and Security Informatics (ISI) 2003 through the present. He is the founder of Knowledge Computing Corporation (KCC), a university spin-off company and market leader in law enforcement and intelligence information sharing and data mining that was acquired by a major private equity firm in 2009. He has been an advisor for major NSF, DOJ, NLM, DOD, DHS, and other international research programs in digital libraries, digital government, medical informatics, and national security research, and has served as a Scientific Counselor/Advisor to the National Library of Medicine (USA), Academia Sinica (Taiwan), and the National Library of China (China). His work is funded primarily through the National Science Foundation as well as other public agencies. He is a Fellow of IEEE and AAAS. In 2006, he was awarded the IEEE Computer Society 2006 Technical Achievement Award, and in 2008 garnered the INFORMS Design Science Award.

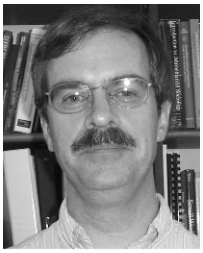


Carlos Castillo-Chavez is a Regents and a Joaquin Bustoz Jr. Professor in the School of Human Evolution and Social Change and the School of Mathematical and Statistical Sciences at Arizona State University (ASU). He is the founding director of the Mathematical, Computational and Modeling Sciences Center. Castillo-Chavez has co-authored nearly two hundred publications, co-authored a textbook in Mathematical Biology in 2001, a volume (with Tom Banks) on the use of mathematical models in homeland security (SIAM's Frontiers in Applied Mathematics Series, 2003), and volumes "Mathematical Studies on Human Disease Dynamics: Emerging Paradigms and Challenges" (American Mathematical Society, 2006) and Mathematical and Statistical Estimation Approaches in Epidemiology (Springer-Verlag, 2009) highlighting his interests in the applications of mathematics in emerging and re-emerging diseases. Castillo-Chavez is a member of the Santa Fe Institute's external faculty, adjunct professor at Cornell University and a fellow of the (i) American Association for the Advancement of Science (AAAS), (ii) the Society for Industrial and Applied Mathematics, and (iii) the American College of Epidemiology. He is the recipient of two White House Awards (1992, 1997), the 2007 AAAS Mentor award, and the 12th recipient of the American Mathematical Society Distinguished Public Service Award. He is a member of the President's Committee on the National Medal of Science and the National Research Council's Board of Higher Education and Workforce. He is currently a member of three scientific mathematical sciences advisory boards at The National Institute for Mathematical and Biological Synthesis (NIMBioS), the Statistical and Applied Mathematics Sciences Institute (SAMSI), and Banff International Research Station (BIRS). NIH, NSF, NSA, and the Sloan Foundation have funded his research.

Bill Lober is an Associate Professor of Health Informatics and Global Health in the University of Washington (UW) Schools of Medicine, Nursing, and Public Health. He is Director of the Clinical Informatics Research Group, Director of Informatics for the International Training and Education Center on HIV (I-TECH), and Associate Director of the UW Center for Public Health Informatics. He graduated from the UCSF/UC Berkeley Joint Medical Program, completed his residency in Emergency Medicine at University of Arizona, and was board certified in EM. He also completed a National Library of Medicine fellowship in Medical Informatics. In addition to his clinical and post-doctoral training, he has a BSEE in Electrical Engineering from Tufts University, and 10 years of industry experience in hardware and software engineering. His research focuses on the development, integration, and evaluation of information systems to support individual and population health. He directs global health informatics projects encompassing facility level and national information systems in Haiti, Kenya, Cote d'Ivoire, Mozambique, Namibia, and Vietnam. His public health research includes surveillance, case reporting, and health information exchange applications to support public health practice. Clinical informatics research includes interoperability frameworks, enterprise architecture, system implementation in resource poor settings, observational cohort studies, and patient-reported outcomes. He was the 2005 Organizing Chair of the Syndromic Surveillance Conference, and a founding Co-Editor of Advances in Disease Surveillance. He has served on the International Society of Disease Surveillance board since 2005, and has had conference and committee roles in AMIA, CSTE, Washington State Department of Health, and other organizations.

Mark Thurmond received the DVM and Masters in Preventive Veterinary Medicine from the School of Veterinary Medicine, University of California, and a PhD in Dairy Science-Epidemiology from the University of Florida. Over the past 40 years, he has been involved in livestock and clinical programs in private, public, and academic practices that served to create and operate surveillance systems for infectious diseases of cattle. His teaching and research dealt heavily with infectious disease epidemiology of livestock, including systems for disease detection. He is currently Professor Emeritus at the University of California at Davis, where he remains engaged in infectious disease surveillance systems as Co-Director of the Center for Animal Disease Modeling and Surveillance.

CONTRIBUTORS' BIOGRAPHIES



Neal Altman is a Senior Programmer on the staff of the Center for Computational Analysis of Social and Organizational Systems and was a team leader and developer of BioWar. He received a MS in Human–Computer Interaction from Carnegie Mellon University and an MA in Anthropology from the University of Pittsburgh (e-mail: na@cmu.edu).

Kathleen M. Carley received her Ph.D. in Sociology from Harvard in 1984, and an S.B. in Political Science and a second S.B. in Economics from Massachusetts Institute of Technology in 1978. She is the Director of the Center for Computational Analysis of Social and Organizational Systems and a Professor of Computation, Organizations and Society in the School of Computer Science at Carnegie Mellon University, Pittsburgh, PA. Her research combines cognitive science, social networks and computer science to address complex social and organizational problems. Her specific research areas are dynamic network analysis, computational social and organization theory, adaptation and evolution, text mining, and the impact of telecommunication technologies and policy on communication, information diffusion, disease contagion and response within and among groups particularly in disaster or crisis situations. Dr. Carley is a member of the North American Association for Computational Social and Organizational Science, the American Sociological Association, the IEEE and the Association for Computing Machinery (e-mail: Karley@cmu.edu).

Dr. Ta-Chien Chan has had research experiences mainly in health informatics and health geographic information system (GIS). He graduated with a master degree (M.S.) from the Institute of biomedical informatics, National Yang-Ming University in 2006. He received his Ph.D. in infectious disease epidemiology from National Taiwan University in 2010. He worked part-time at the

National Health Research Institute (NHRI) and Centers for Disease Control in Taiwan for data analysis in environment health, GIS applications to public health and surveillance of infectious diseases. His research interests include geographical information system (GIS) applications in public health, spatio-temporal clustering algorithms in disease surveillance, climate change and infectious diseases, influenza epidemiology and surveillance.

Yi-Da Chen received the BBA degree in Management Information Systems (MIS) from National Sun Yat-Sen University, Taiwan in 1998 and the MBA degree in MIS from National Chengchi University, Taiwan in 2000. He joined the first twelve-inch fabrication of Taiwan Semiconductor Manufacturing Company as an equipment automation engineer from 2002 to 2004. He is currently a doctoral student in the MIS department at the University of Arizona. His research interests include text mining, infectious disease informatics, and online communities.

Dr. Steven A. Cohen is an Assistant Professor in the Department of Public Health and Community Medicine at the Tufts School of Medicine. His research as a public health demographer focuses on the impacts of population aging on the dynamics of infection in older adults focuses and assessing age-related changes in disease patterns. He has collaborated on a variety of interdisciplinary research projects on how population dynamics contribute to population health and the pathways through which they occur. He received his DrPH in Population Studies from the Johns Hopkins Bloomberg School of Public Health earned an MPH is Biostatistics and Epidemiology from Tufts School of Medicine (e-mail: steven_a.cohen@tufts.edu).

Dr. Jeff Collmann, Center Director and Associate Professor, Disease Prevention and Health Outcomes, O'Neill Institute for National and Global Health law, School of Nursing and Health Studies, Georgetown University, obtained his Ph.D in Social Anthropology from the University of Adelaide, Adelaide, South Australia. His research focuses on understanding the effect of bureaucracy and other complex forms of organization on everyday life. The results of his research on social change among Australian Aborigines have been published in numerous articles and as a book, Fringedwellers and Welfare: the Aboriginal response to bureaucracy. He completed a Postdoctoral Fellowship in Clinical Medical Ethics, Department of Philosophy, University of Tennessee. He joined the ISIS Center, Department of Radiology, Georgetown University in January 1992 where he developed a national reputation in the area of assuring organizational compliance with health information security regulations. He developed and helped implement the HIPAA security program for the Military Health System. Dr. Collmann helped found Project Argus by contributing to social disruption theory and developing the project doctrine. He teaches courses at Georgetown University in the anthropology of biodefense, infectious disease and Australian culture. He joined the O'Neill Institute in February 2008 (e-mail: collmanj@georgetown.edu).

Jean O'Connor is the Associate Director of Policy for the Office of Critical Information Integration and Exchange (OCIIX). Before joining OCIIX, Jean served as the Director of International Research for the Campaign for Tobacco-Free Kids, Deputy Director of the Center for Health Policy and Legislative Analysis at the MayaTech Corporation, and Health Policy and Legislative Attorney for Georgia Governor Roy Barnes. As a public health policy research and advocate, she worked extensively on a number of different public health issue areas including preparedness, federalism, tobacco, access to healthcare, and the relationship between public health and the criminal justice systems. She has been recognized for her achievements in developing state and federal legislative strategies, communicating with the media and state policy makers on emerging public health issues, and conducting comparative public health policy analysis.

Jean holds a Doctor of Law and a Master of Public Health in Health Policy from Emory University and a Doctor of Public Health from the University of North Carolina at Chapel Hill. She is completing her Doctorate in Public Health at the University of North Carolina at Chapel Hill in 2009.

Her training includes legal clerkships in the Office of the General Counsel at CDC, CDC's Office of Technology Transfer, and Blue Cross Blue Shield of Georgia. She is an Adjunct Assistant Professor of Health Policy at Emory University's Rollins School of Public Health. Her work has been published in journals such as the Annals of Emergency Medicine, and the Journal of Law, Medicine and Ethics (e-mail: jgo6@cdc.gov).

Yan Dang received her B.S. and M.S. degrees in Computer Science from Shanghai Jiao Tong University, P. R. China. She is currently working toward her Ph.D. degree in MIS Department, University of Arizona. Her research interests include human computer interaction and text mining.

Dr. Artur Dubrawski is the Director of the Auton Lab and a scientist in the Carnegie Mellon University School of Computer Science. He researches practical autonomy of intelligent systems. His work involves inventing new analytic algorithms and data structures, and pursuing their applications in scientific, government and industrial domains. Dr. Dubrawski gratefully acknowledges receipt of funding from the National Science Foundation, Centers for Disease Control and Prevention, U.S. Department of Agriculture, Food and Drug Administration, U.S. Department of Defense, U.S. Department of Homeland Security, and from several industrial partners. Artur Dubrawski teaches graduate courses on data mining and business intelligence at the Carnegie Mellon University Heinz College. He is also an adjunct professor at the University of Pittsburgh Department of Bio-Medical Informatics.

Dr. Zhanhai Gao received the PhD degree in applied mathematics (mathematical modelling) from the University of New South Wales (UNSW), Sydney, Australia, and the BSc and MSc degrees in mathematics from Northwest University, Xian, China. He joined the School of Public Health and Community Medicine UNSW in April 2008 after completing a PhD in applied mathematics (modelling

HIV and Hepatitis C virus epidemics in Australia) and subsequent post-doctoral research in modelling for parasite biology & infectious disease in livestock as well as mathematical modelling for vaccine preventable disease. His research interests include mathematical modelling and computing for infectious diseases, biostatistical analysis, mathematical biology, clinical trial data management and dynamical systems.

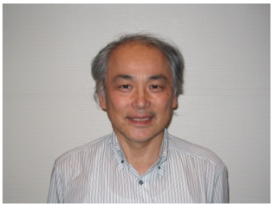
Dr. Ivan J. Gotham, Ph.D., is a Senior Research Scientist at the NYS Health department and Bureau Director of Healthcom Information Systems and Informatics. He is also Assistant Professor, Department of Biometry and Epidemiology (School of Public Health) at the State University of New York at Albany. He has served as Principal and Co-Principal investigator on federal informatics grant projects, authored peer reviewed publications on informatics topics and works closely with the NYS Office of Health Information Technology Transformation in supporting the state's plan to establish a Statewide Health Information Network for Health Information Exchange.

Taku Horie is a doctor at Chiimiya-Horie Clinic. He has been a vice-chairman of the medical association of Izumo City from 2007. He graduated from Iwate Medical University in 1976. He was trained as an anesthesiologist and a physician.

Amy Ising, MSIS, Adjunct Assistant Research Professor in the Department of Emergency Medicine at the University of North Carolina at Chapel Hill, is the Program Director for the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT). Ms. Ising oversees the maintenance and development of NC DETECT and has over ten years' experience handling the extraction, standardization, auditing and reporting of emergency department data for public health surveillance. She is a co-investigator on a wide variety of health informatics and biosurveillance-related research projects. Ms. Ising received a B.A. with Distinction from the University of Virginia, and a M.S. in Information Science from the University of North Carolina at Chapel Hill.

Kiyoshi Kikuchi is Executive Director (patient safety) and Vice President, senior pediatrician in Shimane Prefectural Central Hospital, Japan. He graduated from Kyoto University School of Medicine in 1977, and had worked at Departments of Pediatrics, Kurashiki Central Hospital between 1977 and 1980, Kyoto University between 1980 and 1986, and Shimane Medical University between 1986 and 1995. He moved to Shimane Prefectural Central Hospital in 1995. He received Ph.D. in Medicine from Kyoto University in 1989. He had studied molecular biology of IGF1 at Washington University School of Medicine (St. Louis, USA) between 1989 and 1992.

Dr. Chwan-Chuen King, a professor at the Institute of Epidemiology, College of Public Health, National Taiwan University, has taught several courses of infectious diseases, including epidemiology of infectious diseases, current problems and responses to emerging and reemerging infectious diseases, epidemiology of arboviral infections, pathogenesis of viral diseases, and vaccines. Her research interests include epidemiology and pathogenesis of dengue fever and dengue hemorrhagic fever, surveillance of infectious diseases, epidemiologic changes and interspecies transmission of influenza viruses, and epidemiology and public health policies. She has been an advisor for the Centers for Disease Control in Taiwan on elimination of poliomyelitis, epidemiology and health policies of severe acute respiratory syndrome (SARS), and pandemic influenza preparedness. In the past, she had been a consultant for Liberal Education Program at the Minister of Education in Taiwan and a standing committee member for Science Monthly Journal published in Taiwan. Currently, she is also a free writer on issues of infectious disease and public health, science and policies, and higher education.



Kazuo Kodama is a physician in Kodama Clinic. He has been a regular director of the medical association of the Shimane Prefecture since 2004. He graduated from the Kansai Medical School in 1975 and was trained as a digestive organs physician.

Eileen Koski is Director of Consulting Analytics at Medco Health Solutions. She had extensive experience designing and developing medical database and reporting systems while at Columbia University, as well as being Director of Operations on a large multi-center NIH-sponsored clinical trial. As Director of Informatics Research at Quest Diagnostics, her work spanned three areas: graphical representation of medical data, particularly report design and GIS; population based data analysis including the use of data mining and anomaly detection; and Health Information Technology (HIT) standards development and implementation. Eileen currently serves as Chair of the Biosurveillance Tools and Workforce Development sub-groups of the Health R&D Joint Advisory Working Group, and has previously served as chair of the ACLA HIT Standards Committee, co-chair of the ANSI HITSP Population Health Technical Committee, and a member of other committees related to HIT standards and biosurveillance. She has a bachelor's degree in Biology and an M.Phil. in Sociomedical Sciences from Columbia University (email: EileenKoski@gmail.com).

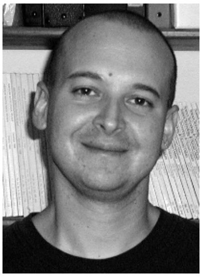
Catherine A. Larson is the Associate Director of the Artificial Intelligence Lab and an associate research Scientist in the Management Information Systems Department at The University of Arizona. She received the B.A., with majors in Spanish and Anthropology and a minor in Portuguese, and the M.S. in Library and Information Science from the University of Illinois at Urbana-Champaign. She has held librarian positions at the University of Illinois, University of Iowa, and University of Arizona. Her research interests include system evaluation and user studies, digital library, and the organization of information.

Linh Le is a research scientist with New York State Department of Health (NYSDOH), where he is coordinating informatics research and development projects for Bureau of Healthcom Network Systems Management. He is the original technical architect of New York State (NYS) Integrated Health Alerting and Notification System (IHAN). His research interests include decision support system, interoperability standards for health information exchange, data visualization, predictive modeling of healthcare resources utilization, computer modeling of adverse health effects and geographical distribution of environmental contaminants.

Thomas Lotze is a PhD candidate in the University of Maryland's Applied Mathematics and Scientific Computation program. His dissertation research takes a statistical approach to anomaly detection in time series, with a focus on providing early warning of disease outbreaks. His other research interests include machine learning, information visualization, and network analysis. He received his Bachelor's degree in Computer Science from Harvard University.

Hsin-Min Lu received the bachelor's degree in business administration and MA degree in economics from the National Taiwan University, and the PhD degree in information systems from the University of Arizona. He is an Assistant Professor in the Department of Information Management at the National Taiwan University. His research interests include data mining, text mining, and applied econometrics.

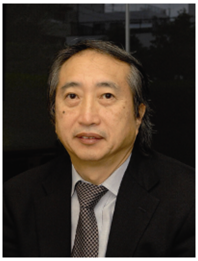
Professor Raina MacIntyre is Head of the School of Public Health and Community Medicine at UNSW and professor of infectious diseases epidemiology. She runs a highly strategic research program spanning epidemiology, vaccinology, mathematical modelling, public health and clinical trials in infectious diseases. She is best known for research in the detailed understanding of the transmission dynamics and prevention of infectious diseases, particularly respiratory pathogens such as influenza, tuberculosis and other vaccine-preventable infections. She has a particular interest in adult vaccination with a focus on the elderly.



Eric Malloy is a Simulation Specialist on the staff of the Center for Computational Analysis of Social and Organizational Systems. He created the parallel execution architecture used by BioWar as well participating in the simulation code development (e-mail: emalloy@cmu.edu).

Elena N. Naumova is a biostatistician interested in the development of analytical tools for time series and longitudinal data analysis applied to disease surveillance, exposure assessment, and studies of growth. Her particular research emphasis is the creation and application of statistical tools to evaluate the influence of an extreme and/or intermediate event on spatial and temporal patterns. Her research activities span a broad range of research programs in infectious disease, environmental epidemiology, molecular biology, immunogenetics, nutrition and growth (e-mail: elena.naumova@tufts.edu).

Dr. Yasushi Ohkusa is the senior scientist of Infectious Disease Surveillance Center, National Institute of Infectious Disease, Japan. He received Ph.D. for medicine in 2005 from Tsukuba University and for economics in 2001 from Osaka University. He researches syndromic surveillance, the mathematical modeling and cost-effectiveness analysis for vaccine.



Nobuhiko Okabe, MD, PhD, is the Director of Infectious Disease Surveillance Center, National Institute of Infectious Disease, Japan. He graduated from Jikei University School of Medicne in 1971. He studied pediatrics as post graduate training, and worked at several hospitals as a pediatrician. For medical research, he studied infectious diseases. He studied his specialty at Vanderbilt Univ. School of Medicine, USA, as a research associate of pediatric infectious diseases in 1978-1970. He served at the World Health Organization Western Pacific Regional Office in Philippines as a Regional Advisor for Communicable Diseases in 1990-1994. He returned to Japan and worked at Jikei Univ. Schoolo of Medicine as an associate professor of Pediatrics and moved to National Institute of Infectious Diseases at 1997.

Adam D. Robinson, Esq., Portfolio Division Manager for the Preparedness, Protection and Response Division of The MITRE Corporation's Homeland Security Center. Bringing more than 25 years of government and commercial experience in large-scale, enterprise solution deployment, Mr. Robinson has worked with national and international organizations providing system architecture, design and implementation leadership including operational and policy guidance. In addition to his enterprise systems engineering capabilities, Mr. Robinson is admitted to the District of Columbia Bar and Virginia State Bar. He holds a Bachelor's degree in Mathematics from the University of Vermont and a Juris Doctor from the Catholic University of America Columbus School of Law (e-mail: arobinson@mitre.org).

Henry Rolka is the Associate Director for Information Exchange in the Centers for Disease Control's Office of Critical Information Integration and Exchange. He has previously served the CDC as a Senior Advisor to the National Center for Public Health Informatics, Branch Chief for the Statistical Analysis Branch of the National Immunization Program, one of the original members of the CDC Surveillance Systems Integration Project as well as a statistical consultant for the Epidemic Intelligence Service Officer's Program. He was instrumental in the early implementation of the BioSense Program in its first phase, served as the CDC representative to the Defense Advanced Research Project Agency's BioAlirt Program as well as the design review for the Department of Homeland Security's National Biosurveillance Integration System and is currently the Chair of CDC's Statistical Advisory Group.

Henry completed a Master's degree in Statistics at Kansas State University where his research included the application of non-linear models and the use of simulation techniques to evaluate them. He also has a graduate degree in Health Services Administration as well as a background as a Registered Nurse and in Human Intelligence in the US Army. His history with surveillance and countermeasure administration dates back to the 1970s when he managed a project to evaluate the use of a new vaccine in an endemic population in rural Africa (e-mail: HRolka@CDC.Gov).

Matthew Scholer, MD, PhD, FACEP is an assistant professor of emergency medicine and a practicing academic emergency physician. Dr. Scholer has been involved in public health informatics since joining the faculty of emergency medicine at UNC Chapel Hill in 2003. As chair of the NC DETECT Syndrome Definition Workgroup since 2004, he leads an interdisciplinary group of researchers and informaticians, including state and university epidemiologists, in the development of syndrome definitions for use in the NC DETECT automated statewide infectious disease and bioterrorism surveillance program. Dr. Scholer received his MD/PhD from New York Medical College and completed his residency program in Emergency Medicine at UNC Chapel Hill. His PhD degree is in Pharmacology.

Loren Shaffer received his Masters of Public Health and Ph.D. in Veterinary Epidemiology from The Ohio State University. Dr. Shaffer is a former U.S. Army Preventive Medicine Officer. He began working in civilian public health in 2003. His work efforts have focused on public health preparedness and the development of syndromic surveillance systems for early outbreak detection at the local and state public health level. Dr. Shaffer's academic interests include pursuing a better understanding of the animal-human interface that contributes to outbreaks of certain disease and utilizing existing veterinary data sources to improve on outbreak detection. (e-mail: loreneshaffer @gmail.com)

Galit Shmueli is Associate Professor of Statistics in the department of Decision, Operations & Information Technologies at the Smith School of Business, University of Maryland. She is co-director of the eMarkets research lab, and affiliated with the Center for Health and Information Decision Systems and the Center for Electronic Markets & Enterprises. Dr. Shmueli's research focuses on statistical and data mining methods for modern data structures. Her main fields of application are biosurveillance and information systems (in particular, electronic commerce).

Debra L. Sottolano, Ph.D. is the NYS Health Alert Network Coordinator and the Manager of the Informatics Unit within the Bureau of Healthcom Network Systems Management, New York State Department of Health (NYSDOH). Responsibilities of the Informatics Unit include development and training of foundation communications and data visualization tools that serve public health preparedness and response applications on the health department's secure internet portal, Commerce. Dr. Sottolano works closely with local health departments, health providers, laboratories, and other NYSDOH partners in health data exchange in developing and implementing these tools and applications. As the NYS Health Alert Network Coordinator, Dr. Sottolano interacts on the national level with CDC and other states in developing best practices and helping to inform policy decisions regarding public health preparedness and response communications. Dr. Sottolano earned her Ph.D. in Organizational Studies at the University at Albany, and also holds an M.B.A. in Marketing, and B.S. in Biology and Chemistry from the University at Albany.

Tamie Sugawara is a researcher of the Infectious Disease Surveillance Center, the National Institute of Infectious Diseases of Japan. She received her Ph.D. degree from University of Tsukuba in 2006. She is a GIS specialist for simulation or syndromic surveillance.

Hiroaki Sugiura is a physician and the vice-director of Sugiura Clinic. He graduated from Shimane Medical University in 1991. After physician training at several hospitals, he has been working at Sugiura Clinic since 2003.

Kiyosu Taniguchi is currently the Chief of Division 1 of Infectious Disease Surveillance Center, National Institute of Infectious Diseases, Japan. He graduated from the Faculty of Medicine at National Mie University in 1984 and trained and worked as pediatrician at several hospitals for 15 years. In 1996 he joined current Institute and worked for WHO/CDS/CSR 2000-2002. Since 2003 he has been working at the current position.

Debbie Travers, PhD, RN, FAEN, Assistant Professor of Health Care Systems in the School of Nursing and of Emergency Medicine in the School of Medicine at the University of North Carolina at Chapel Hill, has been involved in informatics for over 15 years, first in clinical informatics and now as a researcher and teacher. She has also practiced as an emergency nurse since 1981. She has been involved as Principal and Co-Investigator on several health informatics related projects and was the original Principal Investigator of the NC DETECT system. Dr. Travers received a Bachelor's of Science in Nursing from the University of Kansas, a Master's of Science in Nursing from UNC-CH and a PhD in Information and Library Science from UNC-CH.

Anna Waller, ScD, Research Associate Professor in the Department of Emergency Medicine at the University of North Carolina at Chapel Hill and Adjunct Faculty in the Department of Health Behavior and Health Education in the UNC School of Public Health, is the Director of the Carolina Center for Health informatics. She is also Principle Investigator and Science Director for the North

Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT) and either Principal or Co-Investigator on several other health informatics related research projects. She has been involved in public health surveillance work for more than 25 years and in the use of emergency department visit data for public health research since 1994. Dr. Waller received her BA from UNC-CH and her Doctorate of Science from the Johns Hopkins University School of Public Health.

Rochelle Watkins is a senior research fellow at the Australian Biosecurity Cooperative Research Centre at Curtin University of Technology. Initially trained as a physiotherapist, she completed a PhD in social epidemiology, and conducts applied research in the area of infectious disease epidemiology and disease surveillance. Her most recent work has explored the use of geographic information systems in disease surveillance and evaluation.

Dr. James Wood trained in mathematics at the University of Queensland, receiving his PhD in 2004 in the field of mathematical physics. After postdoctoral work at the national centre for immunisation research & surveillance (NCIRS) he joined the school of public health and community medicine at UNSW in 2008 as a senior lecturer. His major research interest is in modelling the prevention and control of infectious diseases and is currently working on models and studies of tuberculosis epidemiology, the use of vaccines to prevent influenza transmission and evaluating control of vaccine preventable diseases in Australia.

Inbal Yahav is a Visiting Assistant Professor in the department of Decision, Operations & Information Technologies at the Smith School of Business, University of Maryland. Her main research interest is the interface between operations research and statistical data modeling. Inbal has presented her work at multiple conferences and have published papers in books and journals. She received her B.A. in Computer Science and my

M.Sc in Industrial Engineering from the Israel Institute of Technology and her PhD in Operations Research and Data Mining from the University of Maryland, College Park.

Yulei Zhang received his B.S. degree in Computer Science and M.S. degree in Bioinformatics from Shanghai Jiao Tong University, P. R. China. He is currently working toward his Ph.D. degree in MIS Department, University of Arizona. His research interests include text mining and human computer interaction.

UNIT I: INFORMATICS INFRASTRUCTURE AND DATA SOURCES

Chapter 1

REAL-TIME PUBLIC HEALTH BIOSURVEILLANCE

Systems and Policy Considerations

HENRY ROLKA* and JEAN O'CONNOR

CHAPTER OVERVIEW

Biosurveillance includes all efforts by public health officials to capture and interpret information about hazards and threats of public health significance, including naturally occurring outbreaks of diseases, reportable conditions and syndromes, occurrences of zoonotic diseases, environmental exposures, natural disasters, and conditions caused by acts of terrorism. Partnerships, such as those between governments and the private sector, non-governmental organizations, healthcare providers, public utilities, and the veterinary medicine community, are a critical component of a comprehensive approach to biosurveillance. Partnerships allow for the collection of data from diverse sources and allow for the inclusion of a range of stakeholders. However, establishing biosurveillance systems with the capacity to quickly collect, analyze and exchange diverse types of data across stakeholder groups is not without practical, technological, political, legal, and ethical challenges. This chapter first explores recent biosurveillance efforts, then examines the legal and policy structure and considerations associated with biosurveillance efforts, and finally provides a framework for considering the needs of stakeholders in the development of new biosurveillance systems.

Keywords: Biosurveillance; Public health; Analytics; Stakeholders; Policy; Law

* *Office of Critical Information Integration and Exchange, National Center for Zoonotic, Enteric, and Vector-borne Diseases, Centers for Disease Prevention and Control, MS-D68, 1600 Clifton Road, Atlanta, GA 30333, USA,* HRolka@cdc.gov

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_1,

1. INTRODUCTION

Although there is no single definition of biosurveillance, it can best be described as the routine collection and integration of timely health-related information to achieve early detection, characterization, and awareness of exposures and acute human health events of public health significance [1]. Biosurveillance includes all efforts by public health officials to capture and interpret information about hazards and threats of public health significance, including naturally occurring outbreaks of diseases, reportable conditions and syndromes, occurrences of zoonotic diseases, environmental exposures, natural disasters, and conditions caused by acts of terrorism. There are significant efforts underway to automate some aspects of biosurveillance; however, much of it is conducted manually. Domestically, biosurveillance is conducted at the federal, state, local, and tribal levels. The vast majority of data is collected at the state and local levels, which are reserved the primary responsibility to regulate the public's health under the tenth amendment of the U.S. Constitution [2]. In the international setting, biosurveillance is conducted at the country level in collaboration with partners such as the World Health Organization [3].

Partnerships, such as those between governments and the private sector, non-governmental organizations, healthcare providers, public utilities, and the veterinary medicine community, are a critical component of a comprehensive approach to biosurveillance. Partnerships allow for the collection of data from diverse sources and allow for the inclusion of a range of stakeholders. Traditional public health surveillance, which is both a subset of, and different from, biosurveillance, relies on the collection of cases of reportable diseases and health conditions, such as HIV/AIDS and tuberculosis, and from vital records, such as birth and death certificates [4]. This type of surveillance is often dependent on active reporting by healthcare providers and the processing of records, which can result in significant time delays between an event occurring and awareness that a public health event, such as an infectious disease outbreak, has occurred [5]. These delays may mean that public health interventions, such as social distancing measures, cannot be effectively implemented to slow the spread of disease. However, if designed and used appropriately, comprehensive and real-time, or near real-time, biosurveillance systems may help to detect the conditions that precede the public health event and can help to detect the event as it is occurring, allowing public health officials to take steps to protect the public. Biosurveillance systems may also provide important data that can be analyzed to better understand the cause, route of transmission, dose-response relationship, and other characteristics of diseases or other health conditions. And, the development and improvement of biosurveillance systems can improve collaboration and promote learning and communication across the diverse stakeholder groups through discussions on what data elements are needed, when and by whom.

Establishing biosurveillance systems with the capacity to quickly collect, analyze and exchange diverse types of data across stakeholder groups is not without practical, technological, political, legal, and ethical challenges. This chapter first explores recent biosurveillance efforts, then examines the legal and policy structure and considerations associated with biosurveillance efforts, and finally provides a framework for considering the needs of stakeholders in the practical operational use of biosurveillance systems.

2. BACKGROUND AND RECENT HISTORY

2.1 Public Health Surveillance

Prior to 2001, domestic surveillance for disease and health conditions was primarily focused on gathering the information necessary to identify, qualify, quantify, and analyze disease statistics, and to enable state and local health departments to report case information to the Centers for Disease Control and Prevention (CDC). Between 1995 and 2001, the Internet had just started to transform public health surveillance, increasing the requirement for common standards and interoperability through increasing societal expectations for real-time information of all kinds. Federal programs and funding for surveillance activities emphasized the harnessing of this potential. The National Electronic Disease Surveillance System (NEDSS) was initiated to promote the transfer of public health, laboratory, and clinical data to state and local health departments efficiently and securely using common data standards, which still serve as the backbone of surveillance systems. There was also some interest in integrating data across sources using the Internet. CDC's Surveillance System Integration Project, initiated in 1998, was developed to tie together many of the current separate systems used for public health surveillance into a comprehensive solution that facilitates the efficient collection, analysis, and use of data and the sharing of software across disease-specific program areas (Table 1-1).

2.2 Impact of the Fall of 2001 on Biosurveillance

The attacks on September 11, 2001 and the anthrax release in the fall of 2001, which heightened awareness of population vulnerability and of the unmet need for early and timely detection of human health threats, accelerated

Table 1-1. Table of acronyms.

BioAlirt	Bio-event Advanced Leading Indicator Recognition Technology
CDC	Centers for Disease Control and Prevention
DARPA	Defense Advanced Research Projects Agency
DHHS	Department of Health and Human Services
DHS	Department of Homeland Security
DoD	Department of Defense
EPA	Environmental Protection Agency
ESSENCE	Early Notification of Community-Based Epidemics
GEIS	Global Emerging Infections System
GIS	Geographic Information System
HIPAA	Health Insurance Portability and Accountability Act of 1996
HL-7	Health Level Seven (electronic messaging format)
HSPD	Homeland Security Presidential Directive
ICD-9	The International Classification of Diseases, 9th Revision
IHR	International Health Regulations
JHU/APL	Johns Hopkins University Applied Physics Laboratory
NBIS	National Biosurveillance Integration System
NEDSS	National Electronic Disease Surveillance System
OTC	Over-the-Counter
PAHPA	Pandemic and All-Hazards Preparedness Act
PHI	Protected Health Information
RODS	Real-Time Outbreak and Disease Surveillance System Laboratory
VA	Department of Veterans' Affairs

efforts to improve the exchange of surveillance data and identify new methods of detecting public health events. Federal funding for new approaches to disease surveillance increased markedly [6]. The concept of using novel data sources previously unexploited in public health surveillance to improve detection rapidly became the topic of research, development and implementation. And, around this time, there was a shift in public health surveillance towards pre-diagnostic identification of population illness, or "syndromic surveillance". (for an overview of the taxonomy of surveillance, see Figure 1-1.)

In late 2001, the Department of Defense (DoD) Defense Advanced Research Projects Agency (DARPA) initiated the Bio-event Advanced Leading Indicator Recognition Technology (BioAlirt) Program [7]. The purpose of BioAlirt was to develop information sources and technologies and to generally advance the science for more quickly recognizing disease caused by the deliberate public release of a weaponized pathogen. The use of novel data sources such as over-the-counter (OTC) sales, pre-diagnostic clinical data, animal health data and absenteeism indicators were explored by BioAlirt (see Figure 1-2). BioAlirt was among the first major programs to address technical challenges of using such data, including evaluation of early detection algorithms, as well as issues such as selecting data sources, complementary sources for optimal detection value in various biowarfare and bioterror scenarios and privacy and confidentiality considerations.

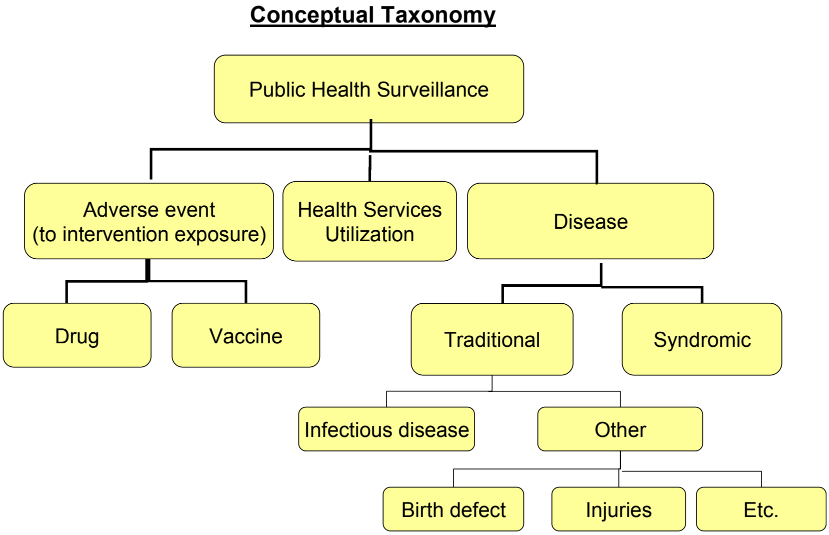


Figure 1-1. An overview of the taxonomy of surveillance.

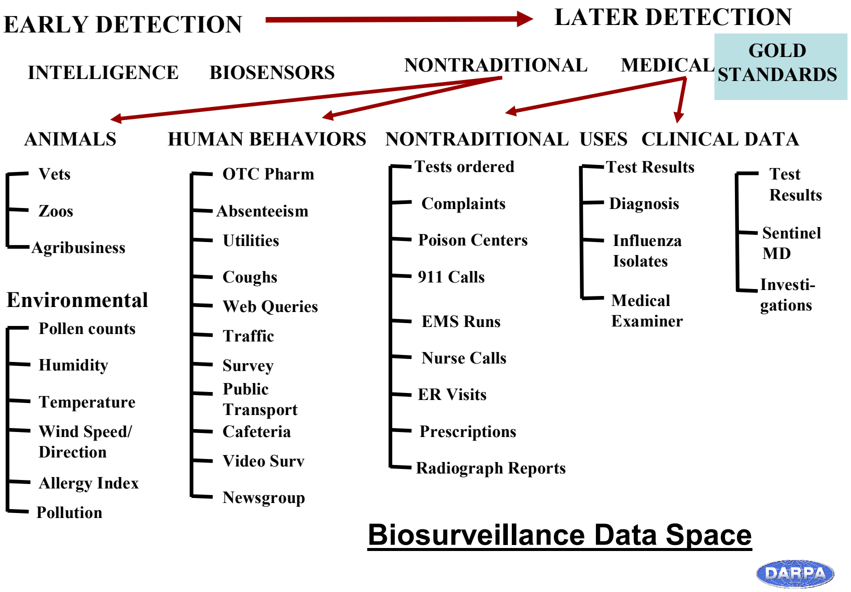


Figure 1-2. Novel data sources explored by BioAlirt.

Although funding for BioAlirt was discontinued after 3 years, the program involved several groups that have been instrumental in the development of biosurveillance systems. Researchers at the Johns Hopkins University Applied Physics Laboratory (JHU/APL) who were key in the development of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE) and the adaptation of ESSENCE for use by local public health departments [8] participated in BioAlirt. ESSENCE, used in the DoD Global Emerging Infections System (GEIS), was originally designed and implemented in 1999 and is integral in the relationships between other federal biosurveillance programs [9]. Its initial implementation was intended as a component to help protect the Washington, D.C., capital district region and the national defense infrastructure from a biological population health threat. Another remarkably influential consortium of novel biosurveillance professionals that participated in the BioAlirt Program were from the University of Pittsburgh and Carnegie Mellon University. This group affiliated through the Real-Time Outbreak and Disease Surveillance System (RODS) Laboratory. Similar to ESSENCE, the RODS Laboratory concepts have been adapted for the development of electronic detection systems in multiple civilian public health communities. The RODS Laboratory staff also established the National Retail Data Monitor (NRDM) which utilizes over-the-counter (OTC) pharmaceutical sales records in near real-time to enable detection of changes in sales volumes that could indicate attempts to self-medicate in the early phases of an outbreak of disease [10].

2.3 BioSense, BioWatch and the National Biosurveillance Integration System

The Public Health Security and Bioterrorism Preparedness and Response Act of 2002 [11] required enhanced biosurveillance capabilities, resulting initiatives to be composed of three key programs: BioWatch, BioSense, and the National Biosurveillance Integration System (NBIS) [12]. A fourth program, Project BioShield, was also developed but was designed for encouraging the development of needed countermeasures, not as an information system [13]. BioWatch, an early warning environmental monitoring system, uses collectors to obtain air samples from multiple locations throughout the U.S. It is a cooperative program among the Environmental Protection Agency, the Department of Homeland Security and the Department of Health and Human Services.

BioSense is a CDC Program to support enhanced biosurveillance through the acquisition of near real-time health indicator data. The program has evolved considerably since its inception in 2003. Data in BioSense include ICD-9 codes, chief complaints, and laboratory tests ordered to serve as indications of likely disease based on clinical impressions prior to laboratory confirmation [14]. BioSense is intended to provide a consolidated standards-based means for collaboratively monitoring the health status of the nation across all levels of public health and to coordinate access to traditional and nontraditional health seeking information by local, state and national public health officials. This monitoring of signs and symptoms, rather than laboratory-confirmed cases, has been generally referred to as "syndromic surveillance" and is similar in concept to the way the ESSENCE and RODS implementations were designed to function at the various sites where they were implemented. ESSENCE was designed to serve the DoD and RODS was advanced as a platform on which to build early outbreak detection methodologies. BioSense was specifically initiated to serve all levels of public health and is a multi-jurisdictional data-sharing surveillance application with nationwide coverage and available to all state and local public health departments. Other applications implemented at local or regional levels may use other data sources and user interfaces; however, without a means for centralized coordination they may not easily be used to compare different localities across the United States.

When BioSense was initiated, the first sources of data included already nationalized sources such as DoD and Department of Veterans' Affairs (VA) ambulatory clinical diagnoses and procedures and Laboratory Corporation of America (Labcorp) laboratory-test orders [15]. The VA, DoD and Labcorp data are batch processed once per day. In 2005, CDC began connecting directly to hospitals for more rapidly cycled (15–20 min) data transport processed as Health Level Seven (HL-7) messages. Because the data elements and analytic approaches are consistent across the country, BioSense data analysts can compare multiple locations by using the same types of data. An application interface summarizes and presents analytical results and data visualizations by source, day, syndrome and location through maps, graphs, and tables. BioSense organizes incoming data into 11 syndromes that are indicative of the clinical presentations of critical biologic terrorism-associated conditions [16]. These syndromic categories and their related codes are classified on the basis of definitions identified by multi-organizational working groups. The real-time civilian hospital data are categorized into 78 syndrome groupings primarily to be more analytically specific.

The BioSense user interface presents information designed to increase users' abilities to detect data anomalies. A data anomaly is a change in distribution or frequency in data compared with geographic or temporal context. Change is quantified by using probabilistic scores of statistical algorithms. An adaptation of a cumulative sum and a generalized linear modeling approach are two methods used to identify anomalies. Changes are also identified by noticeable departures from visually presented patterns of the data. A third anomaly-detection method used in BioSense is the spatial-temporal scan statistic; specifically SatScan [17]. This is an algorithm that combines spatial and temporal activity characteristics and provides indication of the time and geographical region for which the realized number of events is most unusually large. Although algorithms are applied to help interpret data patterns, the BioSense application is not fully automated and does not determine inferential conclusions. Human analysis informed by specific knowledge and experience with the data and application is required to enable reasonable interpretation. Data analysts using the BioSense data and analytic components support implementation and use of the information provided by the system, gather and provide feedback to improve system features by troubleshooting developmental problems, and generate ideas for application enhancement. Analysts also conduct inquiries and document data anomalies. This process has the potential to enable more rapid resolution of health threat indicators through an awareness of operational nuances and cross referencing with other data sources. (See the Analytic Requirements section of this chapter for further discussion of analytics needs.)

The National Biosurveillance Information System, which is currently in development, is a very different type of system from BioSense. The concept for NBIS came out of Homeland Security Presidential Directive 10 (HSPD-10), which describes a national bioawareness system (see *Policy Considerations in Biosurveillance* for additional discussion of HSPD-10). The purpose of NBIS is to create situation awareness using a wide range of real-time information that makes up a common operating picture to facilitate decision-making at the Homeland Security Operations Center and across partner agencies. The Department of Homeland Security is leading this national interagency effort. For NBIS to be successful, the following objectives must be met: (1) the development of a robust information management system capable of handling large quantities of structured and unstructured data; (2) the establishment of a corps of skilled subject matter experts responsible for analyzing the data and providing situational awareness; and (3) the maintenance of a culture of cooperation among interagency partners. Achieving these objectives will not be simple. Although the establishment of a culture of cooperation has been acknowledged by NBIS leadership [18], there is concern the program will fall short of its target [19].

3. POLICY CONSIDERATIONS IN BIOSURVEILLANCE

Policies related to biosurveillance can be loosely categorized into three types: (1) laws that structure which governments or agencies conduct biosurveillance activities, (2) budgets governing the financing of biosurveillance systems, and (3) laws and policies related to who may access and use biosurveillance data.

3.1 Federalism

To understand these laws and policies, especially those that structure which levels of government and which government agencies conduct biosurveillance activities, it is helpful to understand the Constitutional framework that governs public health practice in the United States. Under the Constitution, all powers not specifically delegated to the federal government are reserved to the states [2]. One such power reserved to the states is the police power, which includes the power to promote and protect the public's health [20]. The federal government also may regulate the public's health under the specific powers reserved to it, such as the power to regulate interstate commerce and the power to direct how federal funds, including those given to states under grants and cooperative agreements, may be spent [21]. This sharing of power between the states and the federal government is known as federalism.

Federalism has important practical applications for biosurveillance. States and local governments, where states have further delegated responsibility, carry out most public health practice activities in the United States, including establishing laws related to disease reporting, conducting epidemiologic investigations, regulating the provision of healthcare, and monitoring of the health of the population. And, in general, states have discretion about whether to share public health information with the federal government [22]. However, the federal government, under its power to lay and collect taxes, also appropriates funds with certain conditions for use by states in conducting surveillance activities [21]. Conditions on funds may effectively mean that, if the funding so requires, a state must share data with, or must report the data to, the federal government in compliance with certain data standards or using particular processes [23]. The federal government's vast resources also mean that technical expertise on best practices for surveillance resides primarily within federal agencies. The result is a partnership with the states, where the states provide data and information and the federal government provides resources and specialized technical expertise.

The federal government also plays an important role in identifying opportunities for the improvement and coordination of surveillance activities among all levels of government and across federal agencies. This role has become increasingly important since the terror and bioterror attacks in the fall of 2001 made clear the need for near real-time public health event detection to mitigate the consequences of an event. These events and this need triggered significant biosurveillance policy developments, including the release of Homeland Security Presidential Directives (HSPD) 10 and 21, the publication of the report of the National Commission on Terrorist Attacks Upon the United States, and the passage of the federal Pandemic and All-Hazards Preparedness Act.

Issued in 2002, HSPD-10 outlined the Executive branch's Biodefense Strategy for the twenty-first century, reflecting policy on the spectrum of biodefense, from event detection to response [24]. HSPD-10 called for the Department of Homeland Security to develop a national bioawareness system built on existing federal, state, local, and international surveillance systems that would permit early warning of a biological event. Recognizing the complexity of such an undertaking, the Pandemic and All-Hazards Preparedness Act (PAHPA) [25] was passed in 2006. Under PAHPA, the responsibility for development of the system was transferred to the Secretary of Health and Human Services, in collaboration with state, local, and tribal public health officials. And, the nature of the system to be developed was expanded and redefined as an "interoperable network of systems to share data and information to enhance early detection of rapid response to, and management of, potentially catastrophic infectious disease outbreaks and other public health emergencies that originate domestically or abroad."

Like HSPD-10 and mirroring the language in PAHPA, HSPD-21 calls on HHS to build the network of systems using existing federal, state, and local surveillance systems and to provide incentives to establish local surveillance systems where systems do not currently exist [26]. HSPD-21, which was issued in 2007, provides for collaboration across HHS and other federal agencies to establish a plan to develop the network, work which was underway as this chapter was being written. HSPD-21 also sets forth other guiding principles for the development of the network, including the need for the network to be flexible, timely, and comprehensive; the need for the network to protect individually identifiable data; and the need for the systems in the network to incorporate data into a nationally shared understanding of current biothreats and events, a concept known as the "biological common operating picture."

The development of HSPD-10 and PAHPA, followed by HSPD-21, reflects the development in thinking about the importance of a shared awareness of situations and events by leaders and decision-makers across government. This shared understanding through the exchange of data is viewed as critical to protecting the public's health from all hazards. In the final report of the National Commission on Terrorist Attacks Upon the United States, also sometimes referred to in the popular press as the 9/11 Report, the Commission reflected on the need to draw on all relevant sources of information to protect the public [27].

This need for a shared understanding of ongoing events in order to facilitate decision-making and rapid intervention is also reflected in international policies. The International Health Regulations (IHR) is an international agreement that requires countries that are parties to the IHR develop the surveillance capacity to detect, assess and report to the World Health Organization certain public health events and conditions [28]. Federalism, however, raises interesting questions related to U.S. compliance with the IHR in that disease reporting is, as described previously in this chapter, primarily within the domain of the states, and the federal government relies largely on voluntary sharing of information by the states [29]. The legal and policy environment for achieving shared awareness of public health situations across the states, the federal government, and with international entities is complex.

3.2 Privacy and Data Use

Equally important and equally complex to the laws related to biosurveillance infrastructure are the laws and policies related to who can access and use biosurveillance data. Domestically, these laws and policies are voluminous. Federalism means that the states and the federal government have parallel, overlapping, and sometimes conflicting laws and policies related to the use, and sharing of biosurveillance data and information. These laws and policies often intersect with other important interests that are not specific to public health, such as the interest in preserving individual privacy and autonomy. The U.S. Constitution protects, to some degree, at least two types of privacy – an individual's interest in being free from government interference in certain life decisions and an individual's interest in controlling their personal information. This latter type of privacy, commonly known as informational privacy, is most relevant to biosurveillance. However, like most legal rights, the right to informational privacy is not absolute. Disclosures of personal information, where such disclosures serve a purpose in the public's interest, may be permissible.

One important law related to the privacy of human health-related information is the Privacy Rule of the federal Health Insurance Portability and Accountability Act of 1996 (HIPAA) [30]. HIPAA established a comprehensive, national minimum standard restricting the use and disclosure of individually-identifiable health-related data or information, protected health information (PHI). Entities required to comply with the rule, known as "covered entities," include all healthcare providers, insurers, some government programs, and their business associates that conduct electronic transactions. The Rule establishes a presumption of non-disclosure and requires covered entities to engage in a number of practices to protect PHI, including establishing systematic safeguards to protect PHI and accounting for each disclosure of PHI. Notably, public health authorities are not required to comply with the Rule and disclosure by a covered entity for public health purposes is allowable under the Rule. However, much of the data collected by public health authorities for biosurveillance purposes is obtained from covered entities. This has important practical implications for biosurveillance, in that it places a burden on the disclosing entity to ensure that the purpose of sharing the information, even with public health authorities, is within the allowable exceptions under the law. Also, because HIPAA establishes a minimum standard for the protection of PHI but does not preempt more stringent state privacy and confidentiality laws, including those that provide special protections for specific diseases (i.e., HIV/AIDS-related information), state laws can be an important consideration in the design of biosurveillance activities.

3.3 Other Policy Considerations

There are a range of other legal and policy considerations in biosurveillance. The Constitution protects individuals from having their property taken by the government without just compensation [31], including protection from the diminution of value through regulation. This can be important where biosurveillance data may be used to make decisions about the need to close a facility, stop the sale of a product, or impact the perceptions of the financial health of a company. The Constitution also protects individuals from being subject to unreasonable search and seizure of property [32], a consideration which may apply where biosurveillance data involves or potentially involves the collection of information that could be considered a trade secret, such as a product ingredient. Laws related to intellectual property and data ownership may also apply. Also, where data or information is obtained through particular mechanisms, laws related to intelligence classifications could prevent sharing or disclosure.

4. ACHIEVING INTEGRATED, REAL-TIME BIOSURVEILLANCE

Real-time biosurveillance requires continuous monitoring of data streams, the systems that maintain the data flow, and cross referencing useful information products from multiple sources. Much progress has been made toward these objectives but much work also remains in meeting the stakeholder, analytic, and policy requirements for these systems.

4.1 Stakeholder Perspectives and Information Requirements

One critical aspect of achieving integrated, real-time biosurveillance systems is the identification of stakeholder perspectives and information requirements. Stakeholder needs for any given biosurveillance system will vary considerably but are likely to vary at least by setting level, temporal aspects of the data, and need for detail or granularity level of the data (see Figure 1-3).

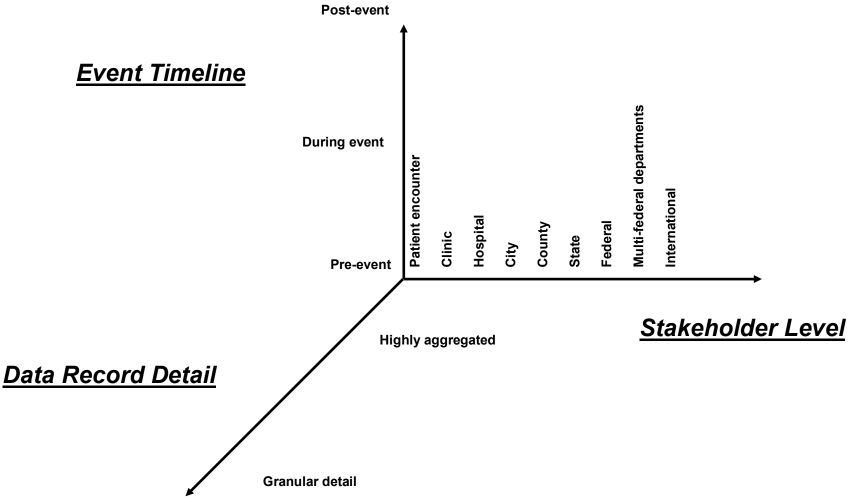


Figure 1-3. System stakeholder information requirements.

The location setting of a biosurveillance stakeholder generally determines interest and particular information of relevance to that setting. These settings, generally, include: (1) patient care in clinical settings, (2) hospital-wide level, (3) multi-hospital groups, (4) city/county levels, (5) states, (6) nationwide, (7) multi-departmental federal level, and (8) global or international setting. The temporal aspect of the biosurveillance interest also is related to stakeholder needs. For example, the nature of an applied biosurveillance approach will be different depending on whether someone is considering routine biosurveillance when there is no event of public health significance taking place (pre-event), an outbreak is beginning or in progress, or after an outbreak to monitor effectiveness of a public health response (post-event). This continuum will drive different information requirements and a different set of priorities. And, the biosurveillance activity will differ depending on the sensitivity of the information needed by the stakeholder. For example, biosurveillance at the clinical level may involve reporting of a diagnostic test for an identified person's individual laboratory specimen (granular data at the clinical setting across the event timeline for the information provided). However, for a syndromic assessment of a population in a multi-jurisdictional geographic area using an automated algorithm, detecting spatio-temporal change in a stream of aggregated data would be more meaningful (outward from the city or county stakeholder level closer toward the origin with de-identified data perhaps aggregated to the zip code level).

4.2 Analytic Requirements

Integrated biosurveillance activities acquire, analyze and interpret data and information from many sources across domains. Achieving the goals and objectives of integrated real-time biosurveillance system investments, as defined by current policy directives, will also require attention to the analytic requirements of such systems. Data acquisition and data transactional preprocessing (extraction, transformation and loading) requires information technology (IT) and informatics skills. The *analytic data preparation, data analytics*, and *interpretation activities* however, require programming, modeling, statistical reasoning, and subject matter expertise as well as a high level of communication skills [33]. The skill sets needed to integrate and analyze data from multiple sources differ from traditional disease surveillance program skills; the skills needed may be independent of the substantive expertise needed to develop and implement specific surveillance systems, which require extensive knowledge of the issue being studied. Instead, the required skills involve extracting data and information using combinations of deductive and inductive reasoning from various and diverse data/ information sources and to communicate conclusions with supported degrees of uncertainty [34].

It is important to consider carefully the role for professional expertise in data analysis and analytic data integration in the public health environment. For example, CDC currently and traditionally has utilized a "dispersed model" for addressing the data analysis and statistical requirements of the public health mission. Analytic data managers and statisticians are usually assigned to program areas and primarily function in the role of "team member" in a specific content area. This enables necessary close working relationships between domain experts and analytic process experts. However, attention by these "hands-on" analysts over time on discrete individual subject areas exclusively can be counterproductive when trying to determine how specific analytic methodologies can fit into larger integrated systems. In an organizational model where centralized human analytic assets serve as consultants to programs and projects, the social pressures for group think are minimized by maintaining a *centralized* professional association support structure.

Another analytics concern in biosurveillance is the distinction between data and information. In traditional public health surveillance, the term data is often used to refer to record-level detail. Public health still relies strongly on data collected in surveys or record-by-record in surveillance settings. These data are analyzed using time-tried statistical methodologies for exploration and inference in public health. There is also, however, now a wealth of real-time public health information available for potential surveillance value in the form of unstructured or text data. Data or information in such form includes chief complaints at the record level and news or intelligence-like reports that come from the news media and systems like EpiX, ProMed and HealthMap. Analyzing and understanding data may be very different from analyzing and understanding unstructured information. The need to combine or fuse data and information adds a layer of complexity to the analytics.

A final and related analytics concern is the distinction between anomalies and events. Anomalies refer to unexpected changes in the pattern of data in a geographic or temporal context. Events are public health threats of actual or potential importance noted in an information source. Event information may come from news releases, web discussions, and sentinel information exchange forums, such as CDC's EpiX, CDCInfo or the Emergency Operations Watch Desk. In order to be informed for interpreting anomalies, it is necessary for analysts to maintain a current knowledge of ongoing events. In CDC biosurveillance operations, addressing observed data anomalies in BioSense, for example, involves first ruling out data transmission issues and gauging the relative urgency for the identified anomaly. Then, the following questions from the available data are answered: (1) How widespread is the anomalous pattern? (2) Are similar patterns found in adjacent regions? (3) For how many days has the anomaly lasted? (4) Has the geographic spread changed with time? (5) Does the pattern have a day-of-week or cyclical nature? (6) Did a similar pattern exist during the same period last year? And, (7) Does the anomaly affect primarily one sex or age group? Finally, corroborating data and/or information relating to an anomaly or an event of public health relevance is sought [16]. Conducting these steps within each of many biosurveillance domain information systems and reporting out discovery and non-discovery assessments on a regular real-time cycle to a fusion center is a logical next step in this process for the development of a comprehensive biological common operating picture and the maintenance of current public health situational awareness.

4.3 Policy Requirements

In addition to attention to stakeholder and analytic needs, achieving integrated, real-time biosurveillance calls for consideration of what laws and policies impact the system and whether such laws and policies are adequate. As discussed earlier in this chapter, in determining what laws and policies apply, there are least three types to consider: (1) laws that structure which government entity is collecting the data, (2) budgets governing the financing of the system, and (3) laws and policies related to who may access and use the data. The adequacy of laws and policies for integrated, real-time biosurveillance should be examined from the perspective of all of the stakeholders involved in the particular system or systems. Generally, adequacy can include consideration of the extent to which the laws and policies identified facilitate or inhibit the following: conducting the biosurveillance activity, use of appropriate technology, availability of analytic and personnel resources, access to needed financial resources, sharing data and information, protection of individual privacy and civil liberty interests, community participation, public trust in government, and promotion of the public's health. Where laws, policies, or stakeholder interests are in conflict or inhibit biosurveillance activities, although policy change may be necessary, special consideration should first be given to the purpose, ethics, and legality of the biosurveillance system.

5. CONCLUSION AND DISCUSSION

Biosurveillance systems and policy considerations are closely linked, with one affecting the other. A confluence of factors including the events of the fall of 2001, developments in technology, and an emphasis at the federal level on preparedness policy has generated new interest in pre-event monitoring and event detection in public health, creating substantial opportunities to advance efforts to achieve integrated, real-time biosurveillance. Real-time biosurveillance will allow public health leaders at all levels of government to achieve situation awareness through a common operating picture. However, achieving this vision will involve a monumental effort not dissimilar in scope to efforts like the Manhattan Project or the moon landing.

In achieving this vision, biosurveillance experts must keep in perspective that biosurveillance, like traditional public health surveillance, is about people, not technology. Technology is only a tool to meet stakeholder, analytic and policy objectives. There are many stakeholders and stakeholder settings in real-time biosurveillance. It will improve understanding and communication among partners to consider the "stakeholder space" and policy considerations described in this chapter. Biosurveillance objectives also can only be met when the analytic operational requirements are more fully recognized and addressed. A persistent and destructively limiting factor for past biosurveillance development has been neglect for recruiting, training, development and maintenance of a large analytic data knowledge, analytic programming and statistical analysis human resource pool that includes leadership from the statistical and related information science (GIS, mathematical modeling and simulation) community. The public health and biosurveillance communities must recognize this need in order to achieve sound information products.

This goal and other goals – policy and programmatic – for biosurveillance will best be achieved through studying lessons learned from current and past biosurveillance program efforts and through partnerships across governments, the healthcare sector, and information science professionals.

The findings and conclusions in this chapter are those of the authors and do not necessarily represent the views or position of the Centers for Disease Control and Prevention.

QUESTIONS FOR DISCUSSION

1. How does biosurveillance differ from traditional public health surveillance?
2. Who are the stakeholder groups in biosurveillance activities and systems and how should their needs be taken into consideration in the development of new systems?
3. In designing a biosurveillance system, what factors would you take into consideration and why?
4. Using examples of existing surveillance systems, describe how surveillance systems can be improved to achieve integrated, real-time biosurveillance.
5. How do policies and laws impact the development of surveillance systems and the way in which the needs of stakeholders are met?
6. Explain the tension between privacy and data use and sharing in biosurveillance and why it matters.
7. What are the analytics needs for integrated, real-time biosurveillance?
8. How does biosurveillance support public health situation awareness and the development of a common operating picture?

REFERENCES

1. Centers for Disease Control and Prevention. National Biosurveillance Strategy for Human Health, version 2.0, February 2010. Atlanta, Georgia. Available from: http://www.cdc.gov/osels/pdf/NBSHH_V2_FINAL.PDF
2. U.S. Const. amend. X.

3. WHO. The world health report 2007 – A safer future: global public health security in the 21st century, 2007. [http://www.who.int/whr/2007/en/index.html].
4. Thacker SB, Berkelman RL (1988). Public Health Surveillance in the United States. Epidemiol Rev, 10:164–90.
5. Jajosky RA, Groseclose SL (2004). Evaluation of Reporting Timeliness of Public Health Surveillance Systems for Infectious Diseases. BMC Public Health, 4:29.
6. Prevent proliferation of weapons of mass destruction [website]. [http://www.fas. org/ota/reports/9341.pdf].
7. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW (2005). Algorithms for Rapid Outbreak Detection: A Research Synthesis. J Biomed Inform, 38:99–113.
8. Lombardo J (2007). Disease Surveillance: A Public Health Informatics Approach. New Jersey: John Wiley and Sons.
9. Embrey EP, "Statement by Deputy Assistant Secretary of Defense for Force Health Protection and Readiness on The Role of Department of Defense Biosurveillance Efforts – Including ESSENCE – in Support of the National Biosurveillance Integration System (NBIS)", Before the House Committee on Homeland Security, May 11, 2006.
10. Wagner MM, Espino JU, Tsui F-C, Gesteland P, Chapman WW, Ivanov O, Moore AW, Wong WK, Dowling J, Hutman J. Syndrome and Outbreak Detection from Chief Complaint Data: Experience of the Real-Time Outbreak and Disease Surveillance Project. Presented at the National Syndromic Surveillance Conference, New York Academy of Medicine, New York, October 20–24, 2003.
11. Public Health Security and BioTerrorism Preparedness and Response Act of 2002, Pub. L. No. 107–188, 116 Stat. 594.
12. Cecchine G, Moore M (2006). Infectious Disease and National Security: Strategic Information Needs, Prepared for the Office of the Secretary of Defense, RAND National Defense Research Institute. [http://rand.org/pubs/technical_reports/2006/ RAND_TR405.pdf].
13. Marburger J (2003). Keynote Address on National Preparedness by Director of the Office of Science and Technology Policy, Executive Office of the President. [http://www.ostp.gov/pdf/10_20_03_jhm_biosecurity_2003.pdf].
14. National Center for Health Statistics. International Classification of Diseases Ninth Revision, Clinical Modification, Sixth Edition. [Accessible at: http://www.cdc.gov/nchs/datawh/ftpserv/ftpicd9/ftpicd9.htm].
15. Bradley CA, Rolka H, Walker D, Loonsk J. (2005). BioSense: Implementation of a National Early Event Detection and Situational Awareness System. MMWR Morb Mortal Wkly Rep, 26(54 Suppl):11–9.
16. Sokolow LZ, Grady N, Rolka H, Walker D, McMurray P, English-Bullard R, Loonsk J (2005). Practice and Experience Deciphering Data Anomalies in BioSense. MMWR Morb Mortal Wkly Rep, 26(54 Suppl):133–9.
17. Kulldorff M (2001). Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic. J R Stat Soc A, 164(1):61–72.
18. Smith K (2006). Creating a Nation-wide, Integrated Bio-surveillance Network. Statement for the Record by Acting Deputy Chief Medical Officer, Department of Homeland Security to U.S. House of Representatives Committee on Homeland Security, Subcommittee on Prevention of Nuclear and Biological Attack. [http://chs.clientapp2.com/hearings/viewhearing.aspx?id=30].

19. Skinner RL (2007). Better management needed for the National Bio-Surveillance Integration System Program, U.S. Department of Homeland Security, Office of the Inspector General. [http://www.dhs.gov/xoig/assets/mgmtrpts/OIG_07-61_Jul07.pdf.].
20. Jacobson v. Massachusetts, 197 U.S. 11 (1905).
21. U.S. Const. art I, § 8.
22. Neslund VS, Goodman RA, Hadler JL (2007). Frontline Public Health: Surveillance Field Epidemiology. In: Goodman R, Hoffman RE, Lopez W, Matthews GW, Rothstein MA, Foster KL, editors. Law in Public Health Practice, 2nd ed. New York: Oxford University Press. pp. 222–37.
23. See, for example, South Dakota v. Dole, 483 U.S. 203 (1987).
24. Homeland Security Presidential Directive 10: Biodefense for the 21st Century (April 28, 2004) [http://homelandsecurity.tamu.edu/framework/keyplans/hspd/hspd10].
25. Pandemic and All Hazards Preparedness Act, Pub. L. No. 109–417, 120 Stat. 2831.
26. Homeland Security Presidential Directive 21: Public Health and Medical Preparedness (Oct 17, 2007) [http://www.dhs.gov/xabout/laws/gc_1219263961449.shtm].
27. National Commission on Terrorist Attacks against the United States (2004). 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks against the United States, July 22, 2004. [http://www.gpoaccess.gov/911/Index.html].
28. International Health Regulations 2005. WHO: Geneva, 2006.
29. Wilson K, McDougall C, Upshur R (2006). The New International Health Regulations and the Federalism Dilemma. PLoS Med, 3(1):e1. [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1315361].
30. C.F.R. § 160.202.
31. U.S. Const. amend. V.
32. U.S. Const. amend. IV.
33. Davenport TH, Harris JG (2007). Competing on Analytics. Boston, MA: Harvard Business School Press, pp. 153–73.
34. Schum DA (1994). The Evidential Foundations of Probabilistic Reasoning. New York: John Wiley and Sons, Inc.

SUGGESTED READING

1. Homeland Security Presidential Directive 10: Biodefense for the 21st Century (April 28, 2004) [Available at: http://homelandsecurity.tamu.edu/framework/keyplans/hspd/ hspd10].
2. International Health Regulations 2005, WHO: Geneva, 2006.
3. Lombardo JS. Disease Surveillance: A Public Health Informatics Approach.
4. National Commission on Terrorist Attacks against the United States (2004). 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Against the United States, July 22, 2004. [Available at: http://www.gpoaccess.gov/911/Index.html].
5. Pandemic and All-Hazards Preparedness Act, Pub. L. No. 109-417, 120 Stat. 2831.
6. Homeland Security Presidential Directive 21: Public Health and Medical Preparedness (Oct 17, 2007) [Available at: http://www.dhs.gov/xabout/laws/gc_1219263961449.shtm].
7. Wagner J. Handbook of Biosurveillance.

8. Bettencourt, L.M. A., R. M. Ribeiro, G. Chowell, T. Lant, C. Castillo-Chavez. "*Towards real time epidemiology: data assimilation, modeling and anomaly detection of health surveillance data streams*. Zeng Gotham D, Komatsu K, Lynch C. (Eds.) Intelligence and security informatics: Biosurveillance. Proceedings of the 2nd NSF Workshop, Biosurveillance, 2007. Lecture Notes in Computer Science. New Brunswick, NJ: Springer-Verlag Berlin. Pp. 79–90, 2007.

ONLINE RESOURCES

1. BioSense Website (http://www.cdc.gov/BioSense/)
2. Centers for Disease Control and Prevention Website (http://www.cdc.gov)
3. Health Level-7 Messaging Website (http://www.hl7.org/)
4. International Health Regulations Website (http://www.who.int/csr/ihr/wha_58_3/en/index.html)

Chapter 2

DESIGNING ETHICAL PRACTICE IN BIOSURVEILLANCE

The Project Argus Doctrine

JEFF COLLMANN[1,*] and ADAM ROBINSON[2]

CHAPTER OVERVIEW

Biosurveillance entails the collection and analysis of information needed to provide early warning of outbreaks of infectious disease, both naturally occurring and intentionally introduced. Data derived from repositories containing various types of sensitive information may be required for this purpose, including individually identifiable, copyrighted, and proprietary information. The Project Argus Biosurveillance Doctrine was developed to ensure that ethical and legal principles guide the collection and handling of such information. Project Argus does not, however, use individually identifiable information or any material derived from individually identifiable information for any phase of the project. Further, Project Argus is not used for purposes of law enforcement, counterterrorism, or public health surveillance. This chapter details why and how the doctrine was developed and summarizes its guiding principles and key elements.

Keywords: Biosurveillance; Sensitive information; Information protection; Privacy

[1*] *O'Neill Institute for National and Global Health Law, Disease Prevention and Health Outcomes, School of Nursing and Health Studies, Georgetown University Medical Center, Box 571107, 3700 Reservoir Rd, NW, Washington, DC 20057–1107, USA, collmanj @georgetown.edu*

[2] *The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_2,

1. INTRODUCTION

Biosurveillance entails the collection and analysis of information needed to provide early warning of outbreaks of infectious disease, both naturally occurring and intentionally introduced. Data derived from repositories containing various types of sensitive information may be required for this purpose, including individually identifiable, copyrighted, and proprietary information. Project Argus searches open media in all countries of the globe except the United States to find direct and indirect indications that local communities have identified and begun to respond to an emerging infectious disease such as SARS or pandemic influenza [1]. The ultimate goal is to provide early warning of such events so that countermeasures can be taken to limit the spread and mitigate the consequences of the disease. When originally planning Project Argus, we developed a "Biosurveillance Doctrine" to ensure that ethical principles would guide the collection and handling of such information. Specifically, our efforts included five steps:

- Development and analysis of scenarios for managing sensitive project information
- Examination of relevant laws, regulations, good practice, and case studies in the acquisition, analysis, and archiving of this information
- Development of administrative, physical, and technical policies and procedures for safely managing project information
- Development of technical design requirements for the Project Argus biosurveillance system and
- Development of a doctrine management process

Through these efforts, the system incorporated requirements for the ethical handling of sensitive information from the start, rather than retrofitting it later. Because the initial phase of the project focused only on the acquisition, archiving, analysis, and presentation of biosurveillance information, we limited our initial efforts to ensure ethical practice of/for these activities as reported in this chapter. Project Argus did not, when implemented, use individually identifiable information or any material derived from individually identifiable information for any phase of the project. Further, Project Argus is not used for purposes of law enforcement, counterterrorism, or public health surveillance. This chapter details why and how the doctrine was developed and summarizes its key components.

2. BACKGROUND

Project Argus developed the technical and doctrinal requirements for an integrated, multisource information system designed to perform global biosurveillance for epidemics; biological accidents; and bioattacks on humans, animals, and plants [1].

The ethical issues surrounding the development and maintenance of such a system have been a key consideration from the outset of Project Argus. No single public law or set of regulations governs the handling and protection of the broad range of sources, types, security classifications, and potential uses of the information to be collected and analyzed. Therefore, the Project Argus doctrine team was formed to develop the necessary guidance. The resulting biosurveillance doctrine sets forth explicit principles, management structures, policies, procedures, and technical design requirements intended to ensure the ethical handling and use of sensitive information by project participants. In this respect, a strong moral, organizational, and technical divide exists between Project Argus and initiatives that have drawn the censure of Congress, the media, and the American public for their failure to ensure such protections. After describing the methods used to develop the doctrine, we provide a high-level view of its guiding principles and key elements.

It should be noted that the version of the doctrine presented here applies only to the acquisition, archiving, analysis, and presentation of biosurveillance information in Project Argus. The doctrine team analyzed a broad set of sensitive information, including individually identifiable, copyrighted and public information. We include our analysis of and approach for handling this broad set of sensitive information for the sake of completeness and as a guide to others. Project Argus does not use any individually identifiable information in any phase of the project.

3. OVERVIEW: INFORMATION PROTECTION

Project Argus collects, archives, and interprets various types of information, including confidential or sensitive information that requires special handling. The leaders and sponsors of Project Argus required development of the Biosurveillance Doctrine to familiarize all project members, contractors, and partners with relevant laws, regulations, ethical principles, and good industrial practices governing use of sensitive information and ensure their compliance with their precepts. In addition to examining relevant cases such as the controversy about the Terrorism Information Awareness (TIA) program (see below), the Biosurveillance Doctrine team investigated issues associated with using specific types of sensitive information, including individually identifiable, proprietary, and copyrighted information. Certain guidance, such as the Principles of Fair Use of copyrighted materials, bears directly on the type of information that Project Argus acquires. Other guidance, such as the Security and Privacy Standards of the Health Insurance Portability and

Accountability Act (HIPAA) of 1996 and the European Privacy Directive, directly or indirectly affects how Project Argus shares information with potential partners. For example, no Project Argus investigators qualify as "covered entities" under HIPAA. Project Argus doctrine must, nonetheless, refer to HIPAA because it may potentially collaborate with health-care providers who should share patient information only with HIPAA-compliant partners. HIPAA and the European Privacy Directive also embody versions of good privacy and security practice. By aligning its practice with principles expressed in these regulatory regimes, Project Argus demonstrates good faith in protecting sensitive information obtained from its partners or through its own initiatives.

Good information security practice requires establishing administrative, physical, and technical controls to protect the confidentiality, integrity and availability of all project data. We imagined that research data from Project Argus might reside in various locations, including the ISIS Center at Georgetown University and MITRE (partners in the development of the Argus information system). Relevant information security policies from all such hosts and other project participants appear as appendices to the Project Argus Biosurveillance Doctrine as required. The ISIS Center houses several R&D projects that manage confidential information, including individually identified health information. The ISIS Center has established a risk-based information security program with controls to protect information of several types including public, commercially sensitive, research, and individually identifiable information. MITRE has rigorous controls reflecting its identity as a major Federally Funded Research and Development Center serving sensitive sectors of the U.S. government. Project Argus benefits from the general organizational controls and tailors specific controls to meet its own needs when appropriate. Memoranda of Understanding among participating organizations document their mutual obligations in protecting shared project information of any kind whenever necessary.

The Biosurveillance Doctrine Team's analysis of these general issues yielded some implications for Project Argus. Summaries of these implications follow to help the reader better understand the specific policies and procedures proposed for Project Argus.

3.1 Fair Information Practice Principles

Fair Information Practices represent an international consensus on appropriate handling of personal information. Various versions of these principles appear in the European Privacy Directive, the U.S. Privacy Act, and guidelines issued by the Organization for Economic Cooperation and

Development and the Canadian Standards Association. Key provisions include the following.

- Notice: At or before the time of collection, individuals shall be informed of the personal information to be collected, the purpose of the collection, and to whom the information may be disclosed.
- Consent: To the maximum extent possible, individuals shall consent to the collection of their personal information at or before the time of collection.
- Accuracy: Personal information shall be sufficiently accurate, complete, and current to serve the intended purpose.
- Security: Personal information shall be protected by safeguards appropriate to the sensitivity of the information.
- Access: Individuals shall have the opportunity to review the personal information held about them and records of its disclosure.
- Redress: Individuals shall have the opportunity to request correction of their personal information and to challenge compliance with stated practices.
- Limitation: Collection, use, disclosure, and retention of personal information shall be limited to that which is necessary for the intended purpose.

3.2 Proprietary Information

Copyrighted Information. Project Argus seeks, acquires, archives, and analyzes copyrighted materials, primarily through its web search technology (known as Apollo). The ISIS Center, the home base for Project Argus, qualifies as a non-profit, educational, research-oriented institution. Furthermore, Project Argus is a pilot study of limited scope. If its methods do not prove useful, the project will be discontinued. For materials not easily acquired retroactively due to their ephemeral nature, such as dynamic web pages, Project Argus downloads a copy of the web page and creates an archive of Hypertext Markup Language (html) files for future reference. To stay within the bounds of fair use, as defined in copyright law (17 U.S.C. §107), Project Argus acknowledges its use of copyrighted materials first by purchasing materials of interest when necessary, either directly from the publisher or through an aggregator. Project Argus excludes all website sections not relevant to its research requirements and labels all archived articles as "For Research Purposes Only." The archives will only exist for the life of the project. Project Argus does not distribute, republish, or disseminate the archived articles for any commercial or non-commercial purpose under any conditions. For these reasons, the limited use of copyrighted materials in Project Argus constitutes fair use.

Confidential Business Information. Project Argus may handle confidential business information in many forms. For example, Project Argus could potentially contract with commercial companies to provide aggregated data of various types. In such cases, Project Argus drafts contracts that reflect its own information protection and use policies as well as comply with federal law and policy. In all instances, Project Argus only uses the data for the defined purposes of the project and does not share the data with parties external to Project Argus.

3.3 Individually Identifiable Information

- **Protected Health Information (PHI).** The ISIS Center, the home base for Project Argus, does not qualify as a covered entity under HIPAA because it does not provide or pay for medical treatment of individuals. Under certain circumstances, Project Argus may receive PHI from covered entities such as Georgetown University Hospital or the Washington Hospital Center as part of conducting research in biosurveillance. The HIPAA Privacy Rule would require submission of a Human Subjects Review application of some type to the Georgetown University Medical Center's Institutional Review Board (IRB). Government sponsors might also require review of Project Argus' use and disclosure of PHI by a relevant IRB. Depending on the actual circumstances, the application may seek an expedited or full review. This has not yet occurred in the project but may occur in later phases.
- **Telephone call detail.** Although never used in Project Argus, we investigated methods for preparing aggregate telephone call data between regions of interest based on individually identifiable telephone call information. Each call on a telephone network generates a call detail record (CDR) that stores the telephone number of the phone that made the call (the originating number), the dialed number, the telephone number that received the call (the terminating number), the time at which the call was placed, and the duration of the call [2]. Telephone companies routinely use these statistical analyses to monitor network reliability and detect international fraud. Companies may also perform analysis of aggregated CDR on a contractual basis to third parties.

One may compile aggregated data of call volumes from one specified region to another. It is not highly granular information that an analyst could use to deduce the identities of individual callers in a designated area. When telecommunications carriers are required to provide call-identifying information, it is by court order and is limited to specific individuals and forms of communication. Because the purpose of Project Argus is to detect indications of societal disruption, only aggregated regional call data is of use. Project Argus has never used telephone data of any kind in its work; but, evaluated these measures for the sake of completeness and scholarly relevance.

- **Individual financial information.** None of the data streams initially proposed for Project Argus analysis was financial in nature. There could come a time, however, when aggregated financial data, such as the number of automated teller machine transactions in a given period for a given region, could prove useful for detecting societal disruption. Privacy regulations such as the ones in the Gramm–Leach–Bliley Act could serve as a model in the future but are not needed at this time.
- **Intelligence on U.S. Persons.** The mission of Project Argus includes detecting social disruption, not tracking individuals. Thus, Project Argus has and will not develop, pilot, or evaluate means for identifying individuals for law enforcement, crime prevention, or public health surveillance purposes on U.S. or foreign persons. Project Argus implemented policies and procedures designed to minimize the incidental, or unintentional, collection and to dispose of information about U.S. persons.
- **European Privacy Directive.** As the foregoing discussions of specific types of individually identifiable information imply, the United States implements a sectoral approach to privacy. By contrast, the European Privacy Directive handles all individually identifiable information with a single, comprehensive approach and restricts the flow of information to countries that do not provide substantially equivalent protections. The United States and the European Union (EU) have adopted “Safe Harbor” provisions with which U.S. entities must comply in order to transact business involving personal information between the United States and EU member states. The Biosurveillance Doctrine Team examined the European Privacy Directive and the Safe Harbor provisions because they represent a major instance of the Fair Information Practice Principles, described below, with which Project Argus aspires to comply.
- **Aggregation of Information from Disparate Databases.** Project Argus recognizes the theoretical possibility that it might generate individually identifiable data in the course of combining otherwise de-identified data from disparate databases. In general, Project Argus does not seek to create individually identifiable data from any sources. Project Argus investigators, furthermore, discard any such data that appears incidentally as a function of intended or unintended project procedures.

4. METHODS

The Project Argus doctrine team drew general inspiration from the work of William Odom, who recommends organizing the intelligence community to reflect the phases of the intelligence cycle: topic selection and data collection, analysis, use, and evaluation [3]. The doctrine described in this chapter focused only on the acquisition, archiving, analysis, and presentation of biosurveillance information. Eventually, the doctrine will provide end-to-end protection of information through all phases of biosurveillance (see Figure 2-1).

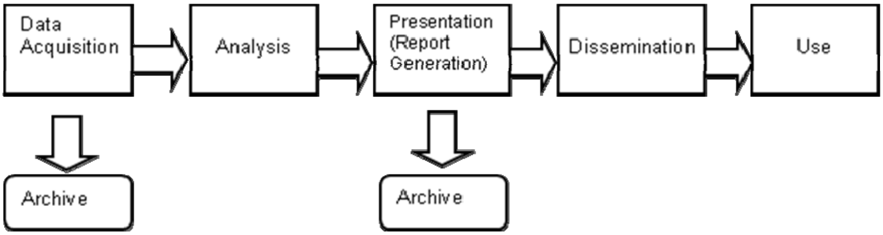


Figure 2-1. End-to-end protection of biosurveillance information.

The doctrine team conducted five types of activities to carry out its charge. The first was an analysis of some typical scenarios that Project Argus team members may confront when managing the sensitive but unclassified information originally imagined for analysis in the project. One scenario was developed for each of five information types: telecommunications information, information from open-source media, remote-sensing information, changes in website content, and air transportation information. The results of the scenario analyses informed our second activity, an examination of laws, regulations, good practice, and case studies in the acquisition, analysis, and archiving of this information, such as the Privacy and Security Rules of the Health Insurance Portability and Accountability Act (HIPAA) of 1996 [4–5]. We developed the doctrine in three steps based on these efforts: we authored administrative, physical, and technical policies and procedures for acquiring, analyzing, archiving, and protecting project information; we created technical design requirements for the system; and we developed a doctrine management process. Through these measures, the system incorporated requirements for the ethical handling of sensitive information from the start, rather than retrofitting them later.

4.1 Information Scenarios

The five scenarios address information that falls under one or more of five broad types of information. Each scenario traces handling of the information through collection, archiving, analysis, and presentation (report generation). Three of the five scenarios are presented below; collectively they illustrate all the issues encountered in the scenario analyses.

- **Scenario 1: Telecommunications Information**
  - ***Scenario type*:** Managing individually identifiable information
  - ***Source of information*:** Individually identifiable call detail records (CDRs)
  - ***Biosurveillance information*:** Ongoing deidentified summaries of calls between selected regions of the world
  - ***Analytic objective*:** Identify significant deviations from baseline rates of calls between these regions of the world

Step 1. Acquire information

Project Argus eventually decided not to use telephone call data for any purpose. The doctrine team developed an approach to deidentifying telephone data before this decision was made. Each call on a telephone network generates a CDR that stores the number of the telephone used to make the call (the originating number), the dialed number, the telephone number that received the call (the terminating number), the time at which the call was placed, and the duration of the call. CDRs constitute International Telecommunications Company (ITC) proprietary business information because the information is generated in the course of ITC's business and is collected and stored by ITC pursuant to its arrangements with its customers. ITC regularly and legally monitors CDRs for a variety of routine business purposes, such as ensuring network reliability and detecting international fraud, and compiles CDR reports. ITC cannot provide the CDRs or CDR reports to third parties such as Project Argus without the permission of the customers involved. To make it possible to establish the baseline as well as the ongoing rate of telephone traffic between regions of the world of interest to Project Argus, it was established that ITC aggregate the CDRs in a manner that removes all individually identifiable data (see below) and retain the CDRs themselves. It was also established that no one from Project Argus ever participate in acquiring the data, see the CDRs or handle any individually identifiable information.

Step 2. Aggregate data

ITC could aggregate the CDR data by preparing graphs that illustrate the volume of calls between regions of the world specified by Project Argus analysts. This step deidentifies the data and, thereby, makes the aggregate results available for such purposes as biosurveillance. The preparation of these graphs does not constitute a routine business practice for ITC; however, ITC agreed to provide the graphs as part of its participation in Project Argus.

Because regions with small call volume would not yield meaningful results, Project Argus agreed to specify a minimum call volume required per geographic area for ITC to produce a data point. Although ITC retained the ability to identify the individuals represented in the graphs through the CDRs, through this mechanism Project Argus could establish a technical design requirement to prevent project analysts from recovering individually identifiable call data from the graphs. ITC and Project Argus agreed to produce a Memorandum of Understanding (MOU) to specify the terms and conditions for the production, transfer, and use of the graphs had the project been implemented.

Step 3. Produce graphical reports and deliver to Project Argus

The protocol specified that ITC produce monthly graphical reports of call volumes for regions specified by Project Argus and transmit the reports to analysts in the ISIS Center. Had call volumes between regions of interest equaled or exceeded an Argus-defined threshold, ITC would have shifted to daily reporting.

Step 4. Archive information

According to the protocol, Project Argus analysts at the Imaging Science and Information Systems (ISIS) Center would receive, index, and digitally store the above reports, producing a comprehensive archive of all information received from ITC. From an information security perspective, the ITC reports contain sensitive but unclassified information. The ISIS Center has established policies and procedures for protecting the confidentiality, integrity, and availability of sensitive information (see http://www.isis.georgetown.edu). During periods when call volumes fall below the alert threshold, requirements for data integrity and timeliness remain consistent with the everyday research and development (R&D) environment of the ISIS Center. When call volumes exceed the alert threshold and ITC reports arrive daily, Project Argus should consider escalating these requirements. The project reevaluates its information security requirements, considers the need for new requirements, and recommends special controls if needed. Archives will exist only for the life of the project.

Step 5. Analyze aggregated ITC information

The ISIS Center proposed to analyze the aggregated call volume information received from ITC and compare them with other datasets so as to identify anomalies that may represent indications and warnings of an emerging bioevent in a region of interest. Project Argus produces and posts several types of reports on a restricted website "For Official Use Only."

- **Scenario 2: Open-Source Media Information**
  - ***Scenario type***: Managing foreign copyrighted information
  - ***Source of information***: Foreign news media websites
  - ***Biosurveillance information***: Whole news stories and abstracted information about disease and social disruption in regions of interest outside the United States
  - ***Analytic objective***: Identify direct indicators of disease and indirect indicators of social disruption secondary to an emerging bioevent

Step 1. Acquire electronic data

Using geographic selection criteria based on Project Argus research requirements, analysts identify online newspapers and other websites of interest to the project. They also identify the information on these websites that is not relevant to the project. For example, on a community website, only sections that might publish articles about school closings would be of interest to the project; sections reporting on local sports scores and entertainment would not be as relevant. Argus engineers write a script specifying the interval of retrieval and the content to be excluded for isisMiTAP, an integrated suite of human-language technologies that processes semistructured textual data. On the established retrieval schedule, isisMiTAP retrieves and stores in its archive all content ("articles") from the identified websites that has not been specifically excluded. Project Argus treats all these reports as if they were copyrighted. To comply with the principles of fair use, the project excludes all website sections irrelevant to its research requirements and labels all archived articles as "for research purposes only." Thus archived articles are not distributed, republished, or disseminated for any commercial or noncommercial purpose under any conditions. Moreover, as with aggregated CDR data, Project Argus archives the articles only for the duration of the project.

Step 2. Translate and catalogue articles of interest

isisMiTAP processes the articles retrieved and presents the information they contain to users through various interfaces. isisMiTAP processing includes machine translation of foreign-language content, information extraction in the form of identifying named entities and keywords (e.g., diseases, locations, people), categorization (binning and posting to a news server), archiving (storing raw and derivative files on disk), and indexing (to support full-text searches). Human linguists translate selected articles to allow for more complete understanding. Analysts prepare summaries of selected articles upon demand.

Step 3. Archive information

Project Argus created and maintains a temporary electronic archive of selected articles at the ISIS Center, including all original-language texts and English translations, article summaries, and web links. All retrieved and archived articles are treated as "confidential – copyrighted" information subject to appropriate administrative, physical, and technical information security controls. The requirements for data integrity and timeliness will remain consistent with the everyday R&D environment of the ISIS Center until a bioevent is suspected, at which point the need to escalate the requirements will be considered. Project Argus reevaluates its information security requirements, considers the need for new requirements, and recommends special controls when needed. For example, a research and development prototype must not typically operate at all times and can tolerate some downtime. Were Project Argus to evolve into a mission-critical operational unit, it would require a business continuity plan that includes tactics to recover from interruptions in its IT system that currently does not exist.

Step 4. Analyze archived articles

Project Argus analysts identify, categorize, and evaluate the significance of salient events in the archived articles. Project Argus produces and posts several types of reports on a restricted website "For Official Use Only."

- **Scenario 3: Remote-Sensing Information**
  - ***Scenario type*:** Managing U.S. federal government information
  - ***Source of information*:** Website of the U.S. National Aeronautics and Space Administration (NASA)
  - ***Biosurveillance information*:** Aggregated remote-sensing information on weather conditions in selected countries of interest
  - ***Analytic objective:*** Identify the existence of weather conditions favorable to the development and spread of selected infectious diseases, such as Rift Valley Fever

Step 1. Acquire information

The U.S. National Oceanic and Atmospheric Administration (NOAA) collects weather data using remote satellite sensing and distributes the data to NASA, which analyzes it and posts it at http://www.nasa.gov. The U.S. government makes these reports available to the public for unlimited use, at no cost. When the doctrine team conducted this analysis, Argus analysts intended to download the aggregated weather reports on a monthly basis from the NASA website to the Argus electronic archive. In practice, these specific reports did not prove useful. Analysts do periodically consult the NASA MODIS Rapid

Response System website to retrieve remote-sensing images of recent fires, volcano eruption, and storms around the globe, data that shares the same properties of aggregated weather data for the purposes of fair use.

Step 2. Archive information

Project Argus downloads various public data from the NASA website to Argus's electronic archive. The ISIS Center's policies and procedures for protecting the confidentiality, integrity, and availability of the content of its archives that include sensitive information apply to such data. As with articles from foreign media, the requirements for data integrity and timeliness of remote-sensing information remain consistent with the everyday R&D environment of the ISIS Center until a bioevent is suspected, at which point Project Argus reevaluates its information security requirements for this information.

Step 3. Analyze remote-sensing information

Argus analysts examine the NASA information to establish the presence or absence of conditions favorable to the development and spread of infectious diseases such as Rift Valley Fever. Project Argus produces and posts several types of reports on a restricted website "For Official Use Only."

4.2 Laws, Regulations, and Good Practice in Managing Sensitive Information

Guided by the results of the scenario analyses, the doctrine team examined laws, regulations, and best practice pertaining to the management of sensitive information federal laws and regulations, including executive orders, international directives, agency policies and procedures, Congressional testimony, government and nongovernment reports, best practices and ethical principles. Certain guidance, such as the principles of fair use of copyrighted materials, bears directly on the types of information Project Argus acquires. Other guidance, such as the Security and Privacy Rules of HIPAA and the European Privacy Directive, may affect directly or indirectly how the project shares information with potential partners. For example, no Project Argus investigators qualify as "covered entities" under HIPAA. The Project Argus doctrine must nonetheless refer to HIPAA because at some point the project may collaborate with healthcare providers who should share patient information only with HIPAA-compliant partners. HIPAA and the European Privacy Directive also embody good privacy and security practice. By aligning its practice with principles expressed in these regulatory regimes, Project Argus demonstrates good faith in protecting sensitive information obtained from its partners or through its own initiatives.

4.3 Case Study: The Terrorism Information Awareness Program

Just prior to the launch of Project Argus, the U.S. Congress discontinued funding for the TIA program, a large counterterrorism effort organized by the Defense Advanced Research Projects Agency (DARPA). At first glance, the TIA program, with its focus on counterterrorism and law enforcement, has little in common with Project Argus. Yet both initiatives had to address common issues related to privacy and security, as well as functionality. Whether beginning with individually identifiable information, such as CDR data, or discovering identities in the course of analysis, as with public health or medical surveillance information, organizations conducting both terrorist investigations and biosurveillance must obey relevant privacy laws; establish pertinent policies and procedures; train their workforces; and implement risk-based administrative, physical, and technical privacy and security safeguards. In preparing the Project Argus Doctrine, the team examined TIA and other programs for lessons regarding these core controls [6–23].

In addition to implementing one of the key lessons learned from these case studies – the need to address such issues in policies and procedures from the outset of a project – the Project Argus doctrine team conducted a detailed analysis of such programs to identify other potential pitfalls and lessons learned. We recognize that the American public basically accepts as legitimate the aims of both scientific research and counterterrorism. However, individual programs must carefully assess and clearly explain the tradeoffs that exist between individual and societal welfare in specific instances, particularly in times of threat and conflict.

Thus biosurveillance investigators must not take for granted the good will of their subjects, their institutions, or their funding agencies. Rather, they must take personal responsibility for ensuring the implementation of appropriate privacy controls. We identified specific means to that end in our research, including:

- Incorporate privacy and security controls into technical design requirements for computerized biosurveillance information systems.
- Take full advantage of the privacy functions of the Institutional Review Board (IRB). As suggested by the report on TIA of the Department of Defense Inspector General, the IRB is fully equipped to advise and monitor researchers on privacy policies, procedures, and practices. In most academic medical research institutions, HIPAA has strengthened the IRB's awareness of and competence to deal with privacy issues.

- Devote great care to preparing the privacy and security portions of the IRB review forms, particularly the informed consent form. The IRB forms can function for an individual research project much like the privacy impact assessment prepared by federal agencies, helping to identify and propose mitigation plans for privacy risks associated with a project. The informed consent form provides an ideal vehicle for explaining to subjects a project's privacy protections.
- When affiliated with a medical center, cultivate an effective relationship with the center's HIPAA privacy and security officers. Like the privacy ombudsmen in federal agencies, these individuals facilitate communication on these matters among researchers, subjects, the institutions involved in the work and external agencies, such as the Office of Civil Rights and the Department of Health and Human Services.
- Consider using an external project advisory board when conducting research or using "data mining" methods that could raise privacy concerns. If properly composed and chartered, such a group can provide useful expertise in policy, privacy, and legal matters beyond a researcher's own institution and enhance the credibility of a project's good-faith efforts in the event of controversy.
- Formally develop and document in writing privacy and security policies and procedures for the research project or its parent unit. As HIPAA and the report of the Department of Defense Inspector General emphasize, these written policies and procedures should explain protections identified in the IRB forms, including administrative, physical, and technical controls for privacy and security.
- Work with relevant information security officers in the home institutions of project members to establish sound controls protecting the confidentiality, integrity, and availability of research data, including individually identifiable information.
- Train project team members in the ethical principles and institutional policies and procedures governing information privacy and security in the project.

5. RESULTS AND ANALYSIS

Based on the analyses described above, we developed three key elements of the Project Argus Biosurveillance Doctrine: policies and procedures, technical requirements, and doctrine management.

5.1 Policies and Procedures

The Argus policies and procedures express the information protection and use guiding principles in succinct form.

General Policy Statements. Project Argus will not acquire, archive, analyze, or distribute information in a form that is prohibited by applicable laws, regulations, or good ethical practice. The project will establish a risk-based IA program to protect the confidentiality, integrity, and availability of all project-related information, including public and unclassified but sensitive information; the project will not handle classified information of any type. Participating institutions will apply their own IA policies in acquiring, archiving, analyzing, and distributing any Project Argus information at their locations. Memoranda of Agreement (MOA) will establish the conditions for sharing information among collaborating organizations. Sharing is defined as providing collaborators access to an original owner's information by any means, including remote electronic access to an owner's archive or transfer of an owner's information to a collaborator's archive.

Specific Policies and Procedures. The scenarios discussed earlier established conditions governing the acquisition, analysis, archiving, and distribution of individually copyrighted, identifiable, proprietary, confidential, publicly available, and business information. For the doctrine, these conditions were translated into specific policies and procedures for each of the five types of information, such as the following ones for individually identifiable information.

- Original owners of individually identifiable information bear responsibility for complying with applicable laws, regulations, and good ethical practice in making such information in their possession available to Project Argus investigators.
- Original owners of individually identifiable information bear responsibility for obtaining permission from subjects as necessary or required for the use of such information in Project Argus.
- Only original owners of individually identifiable information may view, change, analyze, or otherwise use such information in Project Argus unless otherwise agreed upon and justified in writing.
- Original owners of individually identifiable information will limit access to the purposes of Project Argus, its data, data parameters, and data destinations to those members of their own organizations with a need to know such things.
- Project Argus will strive to acquire, archive, analyze, and distribute only aggregated or deidentified information from original owners of individually identifiable information.

- Written MOAs between original owners of individually identifiable information and Georgetown University on behalf of Project Argus will incorporate any relevant rules for the acquisition, archiving, analysis, and distribution of such information to be shared in the course of the project.
- All Project Argus participants, including original owners of individually identifiable information, will abide by the terms and conditions of relevant MOAs regarding the acquisition, archiving, analysis, and distribution of such information obtained from an original owner.
- If Project Argus investigators should inadvertently acquire individually identifiable information, they will discard it and seek no further information about the individual.
- Project Argus participants will receive training in the appropriate handling of any such inadvertently acquired information before incorporating information into the project archive.
- No Project Argus participant will share individually identifiable information with any person, within or external to the project, who is unauthorized to view, receive, or otherwise use it. Nor will they share it with any such organizations or entities.
- Project Argus will classify individually identifiable information as "confidential – individually identifiable" and treat it and any associated information according to the protections for confidential information on the ISIS Center network.

5.2 Technical Requirements

To enable and enforce compliance with the policies and procedures detailed above, the doctrine team developed technical requirements to guide engineers in designing the Project Argus biosurveillance information system to incorporate technical controls that enable and enforce compliance with the policies and procedures detailed above. Consistent with the IA philosophy of defense in depth – which requires multiple, overlapping privacy and security protections – the system's specific application controls will function within an administrative, physical, and technical infrastructure that complements and sustains them. Some of the technical requirements apply to all types of information; while others apply only to specific types, as follows.

- All types of information:
  - Requirements for data integrity and timeliness will remain consistent with the everyday R&D environment of the ISIS Center, unless special circumstances arise, such as a suspected bioevent.

  - Technical requirements will reflect good industry practice for protecting unclassified sensitive information, such as individually identifiable, proprietary, copyrighted, and research information.
  - Project Argus will regularly reevaluate its IA requirements, consider the need for new requirements, and recommend special controls if needed.
- "Confidential – individually identifiable" information:
  - In collaboration with Project Argus investigators, original owners of such information will develop, implement, and monitor the performance of methods for aggregating or deidentifying individually identifiable information when such aggregated information is required for research.
  - Project Argus participants will not be able to reidentify the subjects of the aggregated individually identifiable information to which they have access. Methods used for deidentifying information will minimize the chances of such reidentification.
  - Project Argus participants who are not authorized original owners of such information will not have access to the individually identifiable information from which aggregated information is prepared.
- "Confidential – proprietary" information: Project Argus will develop, implement, and monitor the performance of technical measures designed to enforce any special requirements listed in an MOA between Georgetown University and an original owner of such information.
- "Confidential – copyright" information: Project Argus will develop, implement, and monitor the performance of technical measures designed to enforce the time limit on storage of copyrighted information in the project archives.
- Publicly available information: Project Argus will develop, implement, and monitor the performance of technical measures designed to enforce any special requirements for protecting publicly available information that becomes research data, as deemed necessary by a relevant risk assessment.

5.3 Doctrine Management Process

The Project Argus doctrine team realized that promulgating policies, procedures and technical requirements alone would not necessarily assure compliance. We sought to institutionalize the information protection through a series of organizational measures including routine doctrine team meetings, special oversight procedures, specifications for groups working with but not members of Project Argus and, critically, training for Project Argus staff.

- **Doctrine Team Meetings.** The doctrine team stayed active throughout the initial development of Project Argus and developed a concept of operations for sharing Argus results with its customers. The team has reported to the oversight boards described below.
- **Oversight Boards.** Project Argus established internal and external oversight boards to review its reports and provide overall project guidance including compliance with the Biosurveillance Doctrine. The internal board consists of the doctrine team members and all Project Argus task team leaders. The external board consists of representatives of key government and academic stakeholders.
- **Doctrine Specifications for User Communities.** Sponsors, consumers, and unwitting participants all will have varying needs throughout the life of Project Argus. The doctrine team was instrumental in establishing the organizational conditions for safe sharing of biosurveillance data among a range of government agencies that has proven invaluable in numerous actual biothreat scenarios.
- **Doctrine Training and Awareness Program.** The doctrine team developed a code of ethics that embodies the key elements of the doctrine and served as the foundation for a comprehensive training and awareness program for all Project Argus participants that remains active.

6. CONCLUSION

The American public basically accepts as legitimate the aims of scientific research and counterterrorism, including biosurveillance for pandemics. Individu programs must carefully assess and clearly explain the tradeoffs that exist between individual and societal welfare involved in specific instances, particularly in times of heightened concern about possible attacks by an elusive and possibly indigenous foe. To this end, Project Argus has developed the Biosurveillance Doctrine described in this chapter, designed to ensure the appropriate acquisition, analysis, protection, and use of sensitive information, particularly individually identifiable, copyrighted, and proprietary information. The doctrine will be reviewed and revised as necessary throughout the life of the project. Through these efforts, Project Argus aims to keep faith with its sponsoring agencies, Congress, and the American people.

ACKNOWLEDGMENTS

The authors wish to thank Alexander W. Joel, Esq., Glen C. Graber, Ph.D. and Ted Cooper, M.D. for their helpful comments on earlier drafts of this chapter. Georgeann Higgins and Sandra Sinay made important contributions. The Intelligence Technology Innovation Center (ITIC) funded all work associated with this chapter. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the ITIC.

QUESTIONS FOR DISCUSSION

1. What are the advantages and disadvantages of planning information protection and use policies and procedures before designing an information technology project?
2. What are the similarities and differences in protecting the various types of sensitive information including proprietary, copyrighted and individually identifiable information?
3. Do you think that government or commercial organizations are capable of protecting or appropriately using sensitive or personal information?
4. Are codes of good information practice sufficient to protect sensitive information in a company or government organization? If not, what else do you think is necessary?
5. What difficulties can you imagine might arise in monitoring compliance with an organization's information protection and use policies and procedures after an information management system is deployed?
6. If you were in charge of developing an IT system to handle sensitive information, how would you incorporate information protection issues into your design process?

REFERENCES

1. Wilson, J., Polyak, M., Blake, J., and Collmann, J. A Heuristic Indication and Warning Staging Model for Detection and Assessment of Biological Events. *Journal of American Medical Informatics Association* 15(2) (2008 Mar/Apr), 158–171.
2. Fisher, K., Hogstedt, K., Rogers, A., and Smith, F. (2002). *Hancock 2.0.1 Manual*, AT&T Labs Shannon Laboratory http://www.research.
3. Odom, W.E. (2003). *Fixing Intelligence for a More Secure America.* New Haven, CT: Yale University Press.
4. Department of Health and Human Services; Office of the Secretary. (2002). Final Rule. 45 CFR Part 160, 162, and 164, standards for privacy of individually identifiable health information. *Federal Register* 67, no. 157 (14 August), 53181–53273.

5. Department of Health and Human Services; Office of the Secretary. (2003) Final rule. 45 CFR Part 160, 162, and 164, security standards. *Federal Register* 68, no. 34 (20 February), 8333–8381.
6. Crews, Jr., C. The Pentagon's total information awareness project: Americans under the microscope? *National Review Online*, (2002 25 November).
7. Cooper, T., and Collmann, J. (2005). Managing Information Security and Privacy in Health Care Data Mining. In, Advances in Medical Informatics: Knowledge Management and Data Mining in Biomedicine, New York, NY: Springer Science; pp. 95–137.
8. Defense Advanced Research Projects Agency; Information Awareness Office. (2003). Report to Congress regarding the terrorist information awareness program: in response to consolidated appropriations resolution, Pub. L. No. 108–7, Division M, § 111(b); 20 May.
9. Department of Defense; Office of the Inspector General, Information Technology Management. (2003). Terrorist information awareness program. Report D-2004–033; 12 December.
10. Department of Health and Human Services; Office for Human Research Protections. (2004). Guidance on research involving coded private information or biological specimens. 10 August. Available at http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf.
11. Department of Health and Human Services; National Institutes of Health Office for Protection from Research Risks. (2001). Protection of human subjects. 45 C.F.R. § 46.
12. Department of Health and Human Services. (2003). Protecting personal health information in research: understanding the HIPAA privacy rule. NIH publication no. 03-5388.
13. Federal Trade Commission. (1999). In brief: the financial privacy requirements of the Gramm-Leach-Bliley Act. Available at http://www.ftc.gov/bcp/conline/pubs/buspubs/glbshort.htm.
14. Fisher, K., Hogstedt, K., Rogers, A., and Smith, F. (2002). *Hancock 2.0.1 Manual*. Florham Park, NJ: AT&T Labs Shannon Laboratory.
15. Mack, G., Bebee, B., Shafi, I., Wenzel, G., Medairy, B., and Yuan, E. (2002) Total Information Awareness Program (TIA) System Description Document (SDD) v1.1. Defense Advanced Research Projects Agency Information Awareness Office. White Paper; 19 July.
16. National Institutes of Health. (2004). Clinical research and the HIPAA privacy rule. NIH publication no. 04–5495.
17. Safire, W. You are a suspect. *The New York Times*, (2002 14 November). Available from: http://www.commondreams.org/views02/1114-08.htm.
18. Simons, B., and Spafford, E. Letter to Honorable John Warner, Chairman, Senate Committee on Armed Forces; (2003 23 January).
19. Stanley, J., and Steinhardt, B. (2003) Bigger monster, weaker chains: the growth of an American surveillance society. American Civil Liberties Union, Technology and Liberty Program. White Paper.
20. Sweeney, L. ed. (2003). Navigating computer science research through waves of privacy concerns. Carnegie Mellon University, School of Computer Science, Pittsburgh, Technical report, CMU CS 03-165, CMU-ISRI-03-102.
21. Taipale, K. Data mining and domestic security: connecting the dots to make sense of data. *The Columbia Science and Technology Law Review* 5 (2003), 5–83.
22. Taylor, S. Big brother and another overblown privacy scare. *Atlantic Online*, (2002 10 December).
23. *The Washington Post*. Total information awareness. Editorial, (2002 16 November).

SUGGESTED READING

1. Cooper, T., and Collmann, J. (2005) Managing Information Security and Privacy in Health Care Data Mining. In, *Advances in Medical Informatics: Knowledge Management and Data Mining in Biomedicine*, New York, NY: Springer Science; pp. 95–137.
2. Collmann, J., and Cooper, T. Breaching the Security of the Kaiser Permanente Internet Patient Portal: The Organizational Foundations of Information Security. *Journal of the American Medical Informatics Association* 14 (2007 Mar), 239–43.
3. Cooper, T., Collmann, J., and Neidermeier, H. Organizational repertoires and rites in health information security. *Cambridge Quarterly of Healthcare Ethics*, Volume 17 <http://journals.cambridge.org/action/displayJournal?jid=CQH&volumeId=17&bVolume=y#loc17>, Issue 04 <http://journals.cambridge.org/action/displayIssue?jid=CQH&volumeId=17&seriesId=0&issueId=04>, October 2008, pp. 441-452.
4. Department of Health and Human Services; Office of the Secretary. (2003). Final rule. 45 CFR Part 160, 162, and 164, security standards. *Federal Register* 68, no. 34 (20 February), 8333–8381.
5. Department of Health and Human Services. (2003). Protecting personal health information in research: understanding the HIPAA privacy rule. NIH publication no. 03-5388.
6. Odom, W.E. (2003) *Fixing Intelligence for a More Secure America.* New Haven, CT: Yale University Press.
7. Sweeney, L. ed. (2003). Navigating computer science research through waves of privacy concerns. Carnegie Mellon University, School of Computer Science, Pittsburgh, Technical report, CMU CS 03–165, CMU-ISRI-03-102.
8. Taipale, K. Data mining and domestic security: connecting the dots to make sense of data. *The Columbia Science and Technology Law Review* 5 (2003) 5–83.

ONLINE RESOURCES

1. US Fair Trade Commission, Fair Information Practice Principles, http://www.ftc.gov/reports/privacy3/fairinfo.shtm
2. US Department of Commerce, Safe Harbor, http://www.export.gov/safeharbor/sh_overview.html
3. Health Information Management System Society, *HIMSS Privacy and Security Toolkit* http://www.himss.org/ASP/privacySecurityTree.asp?faid=78&tid=4#PSToolkit#PSToolkit
4. US Department of Health and Human Service, HIPAA Privacy Support, http://www.hhs.gov/ocr/hipaa/
5. US Department of Health and Human Service, HIPAA Privacy and Security Rules, http://aspe.hhs.gov/ADMNSIMP/
6. Caralli, R.A., Stevens, J.F., Wallen, C.M., White, D.W., Wilson, W.R., and Young, L.R. *Introducing the CERT Resiliency Engineering Framework: Improving the Security and Sustainability Processes*, Software Engineering Institute, http://www.sei.cmu.edu/publications/documents/07.reports/07tr009.html

Chapter 3

USING EMERGENCY DEPARTMENT DATA FOR BIOSURVEILLANCE: THE NORTH CAROLINA EXPERIENCE

ANNA E. WALLER*, MATTHEW SCHOLER, AMY I. ISING,
and DEBBIE A. TRAVERS

CHAPTER OVERVIEW

Biosurveillance is an emerging field that provides early detection of disease outbreaks by collecting and interpreting data on a variety of public health threats. The public health system and medical care community in the United States have wrestled with developing new and more accurate methods for earlier detection of threats to the health of the public. The benefits and challenges of using Emergency Department data for surveillance are described in this chapter through examples from one biosurveillance system, the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT). ED data are a proven tool for biosurveillance, and the ED data in NC DETECT have proved to be effective for a variety of public health uses, including surveillance, monitoring and investigation. A distinctive feature of ED data for surveillance is their timeliness. With electronic health information systems, these data are available in near real-time, making them particularly useful for surveillance and situational awareness in rapidly developing public health outbreaks or disasters. Challenges to using ED data for biosurveillance include the reliance on free text data (often in chief complaints). Problems with textual data are addressed in a variety of ways, including preprocessing data to clean the text entries and address negation.

* *Carolina Center for Health Informatics, Department of Emergency Medicine. University of North Carolina at Chapel Hill, 100 Market Street, Chapel Hill, NC 27516, USA, awaller@med.unc.edu*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_3,

The use of ED data for public health surveillance can significantly increase the speed of detecting, monitoring and investigating public health events. Biosurveillance systems that are incorporated into hospital and public health practitioner daily work flows are more effective and easily used during a public health emergency. The flexibility of a system such as NC DETECT helps it meet this level of functionality.

1. INTRODUCTION

Biosurveillance is an emerging field that provides early detection of disease outbreaks by collecting and interpreting data on a variety of public health threats, including emerging infectious diseases (e.g., avian influenza), vaccine preventable diseases (e.g., pertussis) and bioterrorism (e.g., anthrax). With the Centers for Disease Control and Prevention's (CDC) initial focus on bioterrorism preparedness at the state and local level in 1999 and the subsequent anthrax outbreak of 2001, the public health system and medical care community in the United States have wrestled with developing new and more accurate methods for earlier detection of threats to the health of the public. Earlier detection, both intuitively and as illustrated through predictive mathematical models, is believed to save lives, prevent morbidity and preserve resources (Kaufman et al., 1997). Biosurveillance systems use health-related data that generally precede diagnoses and that signal a sufficient probability of a case or an outbreak to warrant further public health response (Buehler et al., 2004).

Rapid detection of disease outbreaks rests on a foundation of accurate classification of patient symptoms early in the course of their illness. Electronic emergency department (ED) records are a major source of data for biosurveillance systems because these data are timely, population-based and widely available in electronic form (Lober et al., 2002; Teich et al., 2002). There are more than 115 million ED visits annually in the United States, and EDs represent the only universally accessible source of outpatient healthcare that is available 24 h a day, 7 days a week (Nawar et al., 2007). EDs see patients from all age groups and socioeconomic classes. Patients may present with early, nonspecific symptoms or with advanced disease. The accessibility of EDs provides a likely healthcare setting for many of the patients involved in a disease outbreak of public health significance. In recent years, EDs have steadily adopted electronic medical records technology (Hirshon, 2000), which has facilitated the replacement of drop-in manual surveillance using ED data with ongoing, real-time surveillance. ED data have been shown to detect outbreaks 1–2 weeks earlier than traditional public health reporting channels (Heffernan et al., 2004; Lober et al., 2002; Tsui et al., 2001; Wagner et al., 2004).

The ED data elements that are used for biosurveillance include the chief complaint (a brief description of the patient's primary symptom(s)), the triage nurse's note (an expansion of the chief complaint that includes the history of present illness), other clinical notes (e.g., physician and nurses' progress and summary notes), initial measured temperature, and diagnosis codes. The most widely used ED data element is the chief complaint because it is recorded electronically by most EDs and may precede entry of a diagnosis or transcription of physician notes by days or weeks (Travers et al., 2003, 2006). The triage note increases the amount of data available, which makes it more likely that biosurveillance algorithms will detect disease outbreaks. Triage notes are becoming more available in electronic form, and one study found that adding triage notes increased the sensitivity of outbreak detection (Ising et al., 2006).

Several challenges to using ED data for biosurveillance have been identified (Hirshon, 2000; Varney & Hirshon, 2006), including costs to EDs and public health, the lack of standardization of ED data, and security and confidentiality. Many EDs still document patient symptoms manually; even when the data are electronic, they are often entered in free text form instead of using standardized terms. Timeliness is also a concern; while some ED data elements are entered into electronic systems at the start of the ED visit, other elements are added hours, days or even weeks later. Even though there is no formal standard or best practices dictating how soon data should be available after an ED visit or other health system encounter for early detection, most surveillance systems aim for near real-time data, available within hours.

The benefits and challenges of using ED data for surveillance will be described in more detail through examples from one biosurveillance system, the North Carolina Disease Event Tracking and Epidemiologic Collection Tool (NC DETECT). NC DETECT evolved from a pilot project in 1999 to demonstrate the collection of timely, standardized ED data for public health surveillance and research. NC DETECT has since grown to incorporate ED visit data from 98% of 24/7 acute care hospital EDs in the state of North Carolina and has developed and implemented many innovative surveillance tools, including the Emergency Medicine Text Processor (EMT-P) for ED chief complaint data and research-based syndrome definitions. NC DETECT now provides twice-daily ED data feeds to CDC's BioSense and has over 200 registered users at the state, regional and local levels across North Carolina. This chapter will review the use of ED data for biosurveillance, including appropriate case studies from NC DETECT.

2. LITERATURE REVIEW/OVERVIEW OF THE FIELD

2.1 History of ED Data Use for Biosurveillance

ED data have been collected for decades for a variety of public health surveillance needs and have been incorporated into electronic systems designed to analyze data related to trauma, injury and substance abuse, among others. Public health officials have used event-based or drop-in biosurveillance systems that include ED data during major events, including the Olympic Games, political conventions, heat waves, after major hurricanes, and after the identification of known widespread outbreaks (Weiss et al., 1988; Davis et al., 1993; Lee et al., 1993; Rydman et al., 1999). Many of these systems have required users to do manual abstractions from medical charts or to enter data into stand-alone systems for specific symptoms of interest. For example, the EMERGEncy ID NET program, established in the late 1990s, created a network of select EDs to manually collect data to study syndromes related to emerging infections of interest to the CDC, using paper forms and standardized computer screens (Talan et al., 1998).

Secondary data, data that are generated as part of normal patient treatment and billing, are generally extracted from hospital information system(s) through automated programs either in real-time (at the time of record generation) or near real-time (hourly, every 12 h, daily). Surveillance systems that use secondary data are intended to be less burdensome to ED staff and less costly than systems requiring manual abstraction (Rodewald et al., 1992). This methodology of ED data collection has become standard practice for biosurveillance systems using ED data, including NC DETECT, RODS, ESSENCE and EARS, among others (Hutwagner et al., 2003; Ising et al., 2006; Lombardo, 2003; Wagner et al., 2004, Waller et al., 2007). While there are several different approaches to automated extraction programs, most rely either on delimited text batch files or HL7 messages.

2.2 Current Status of ED Data Use for Biosurveillance

According to the International Society for Disease Surveillance (ISDS) State Syndromic Surveillance Use Survey, 76% of responding states ($n = 33$) performing syndromic surveillance use ED data (http://isds.wikispaces.com/Registry_Project, accessed June 4, 2008). (Figure 3-1).

While most states and regions rely on ED chief complaint data, there is interest in increasing the number of ED data elements collected, as evidenced by the American Health Information Community's Biosurveillance Minimum Data Set (http://www.hhs.gov/healthit/ahic/materials/meeting10/

bio/BDSG_Minimum_DataSet.doc, accessed June 4, 2008). The recommendations from the American Health Information Community include additional emergency department data elements, such as triage notes, vital signs and ICD-9-CM-based diagnoses. The Biosurveillance Minimum Data Set is currently under formal evaluation for its utility at CDC-funded sites in Indiana, New York and Washington/Idaho.

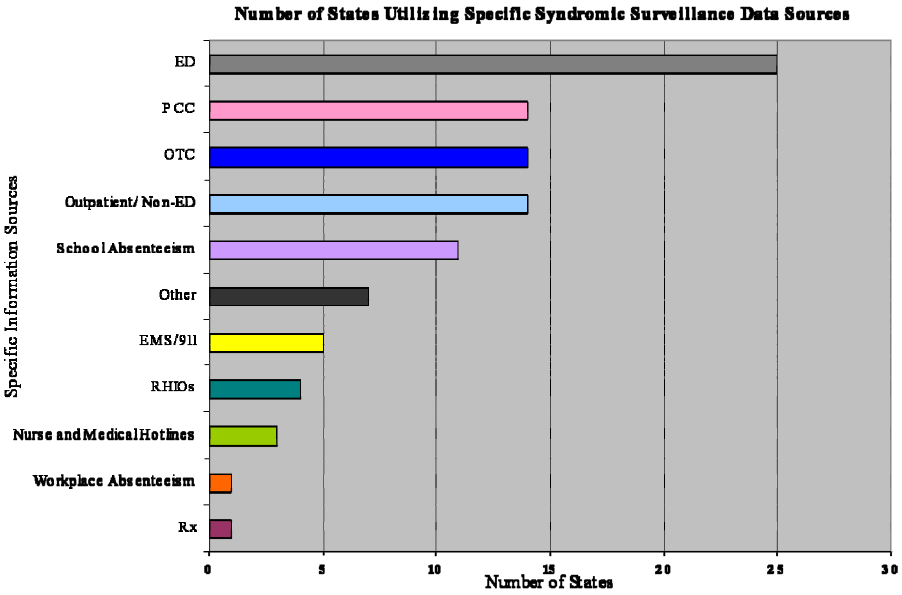


Figure 3-1. ISDS state syndromic surveillance use survey: 41 respondents; 33 use syndromic surveillance

2.3 Infectious Disease Syndrome-Based Surveillance Using ED Data

While infectious disease surveillance has traditionally relied on laboratory results and the reporting of mandated reportable conditions by medical practitioners, ED visit data and timely symptom-based analysis provide additional means for early identification of infectious disease outbreaks. Areas of particular interest include the CDC's list of potential bioterrorism agents, as well as post-disaster (e.g., hurricane, earthquake, chemical spill) surveillance (CDC, 2003). The ability to create effective syndromes to use with ED visit data is of paramount importance to their timely use for public health surveillance.

The structure of syndrome definitions used in biosurveillance is dependent on the design of the system and the nature of the data under surveillance. Individual systems use different methods to identify specific disease symptoms in the chief complaint and triage note data. This includes deterministic methods, such as keyword searching, and probabilistic methods, such as naïve Bayesian and probabilistic machine learning (Espino et al., 2007). Syndrome definitions then classify records into syndromic categories based on which symptoms are identified. To date, no best practices exist to guide syndrome definition development and evaluation (Sosin & DeThomasis, 2004). Which syndromes are monitored and which symptoms are associated with each syndrome varies according to the system under consideration. Furthermore, syndrome structure may vary depending upon which data elements, in addition to chief complaint, are available and their timeliness. Syndrome structure refers to how many symptoms are required, within which data fields they must be found, and which Boolean operators are employed to determine whether a certain record matches a particular syndrome.

2.4 ISDS Consultative Syndrome Group

In September 2007, the ISDS sponsored a consultative meeting on chief complaint classifiers and standardized syndromic definitions (Chapman & Dowling, 2007). At this meeting, representatives from eleven syndromic surveillance systems throughout the country, including NC DETECT, met to discuss which syndromes they monitor and which chief complaint-based clinical conditions they include in each syndrome. Clinical conditions are medical concepts which may be represented by multiple possible data inputs to the system. For example, the concept of "dyspnea" may be represented by signs and symptoms of dyspnea as recorded in an ED chief complaint or triage note field (e.g., "Shortness of Breath" (SOB)), a clinical observation such as an abnormal vital sign (e.g., low oxygen saturations or increased respiratory rate), clinical findings (e.g., abnormal breath sounds on pulmonary exam), abnormal lab findings (e.g., abnormal ABG or positive culture results), imaging studies (e.g., infiltrate on chest x-ray), or certain ICD-9-CM diagnosis codes (e.g., 486, pneumonia). The meeting participants reached consensus on best practices for which clinical conditions to associate with each of six different syndromes (sensitive and specific versions of respiratory syndrome and gastrointestinal syndrome, constitutional syndrome and influenza-like illness syndrome). Through online collaboration and periodic conference calls, the group continues the process of defining specific chief complaint search terms/keywords which best represent these clinical conditions.

3. TECHNICAL APPROACHES FOR GROUPING ED DATA INTO SYNDROMES FOR BIOSURVEILLANCE

While the process of identifying specific chief complaint search terms/ keywords to group into syndromes presents several technical challenges, the timeliness of chief complaints outweighs the benefits of any standardized data that are not available within hours of the ED visit. Textual data such as chief complaint and triage note present several problems, including misspellings and use of ED-specific and locally-developed acronyms, abbreviations and truncations (Travers & Haas, 2003). There are two main approaches to dealing with the variability in textual surveillance data: (1) incorporating keywords in the actual search query statements; or (2) preprocessing the data. In systems that build various keyword searches (e.g., lexical variants, synonyms, misspellings, acronyms, and abbreviations) into the actual surveillance tools, elaborate search statements are constructed, employing statistical software such as SAS (Cary, NC), or standard query language (SQL, Microsoft, Redmond, WA) (Forbach et al., 2007; Heffernan et al., 2004). In systems with preprocessors, the data are cleaned prior to application of a syndromic classification algorithm (Mikosz et al., 2004; Shapiro, 2004). The preprocessors clean text entries, replacing synonyms and local terms (e.g., *throat pain, throat discomfort, ear/nose/throat problem*), as well as misspellings, abbreviations, and truncated words (e.g., *sorethroat*, *sore throaf*, *soar throat*, *ST*, *S/T*, *sore thrt*, *sofe throat, ENT prob*), with standard terms (e.g., *sore throat*) or standard identifiers (e.g., UMLS® concept unique identifier C0242429) (NLM, 2007). Preprocessors often include normalization tools to eliminate minor differences in case, inflection and word order and to remove stop words (NLM, 2006).

While there is no consensus about which approach is best, many biosurveillance programs are implementing preprocessors to improve operations (Dara et al., 2007; Hripscak et al., 2007; Komatsu et al., 2007). Use of preprocessors can streamline maintenance of existing and development of new surveillance queries. Query processing time is also faster, resulting in better overall biosurveillance system performance. One such preprocessor is the Emergency Medical Text Processor (EMT-P), which was developed to process free text chief complaint data (e.g., *chst pn, ches pai, chert pain, CP, chest/abd pain*) in order to extract standard terms (e.g., *chest pain*) from emergency departments (Travers & Haas, 2003). EMT-P has been evaluated by biosurveillance researchers in Pennsylvania and found to improve syndromic classification (Dara et al., 2007). The developers continue to improve EMT-P and have made it publicly available (Travers, 2006).

3.1 Dealing with Negation

While clinical text such as triage notes can improve the accuracy of keyword-based syndrome queries, the data require processing to address negated terms (Ising et al., 2006; Hripcsak et al., 2007). One study evaluated NegEx, a negation tool developed at the University of Pittsburgh (Chapman et al., 2001). NegEx is a simple regular expression algorithm that filters out negated phrases from clinical text. The NegEx system was modified (to include the negation term (-)) and then combined with selected modules from EMT-P that replaced synonyms (e.g., *dec loc* with *consciousness decreased*) and misspellings (*nasaue* with *nausea*) for use in NC DETECT. The pilot results show that this combination of EMT-P and NegEx leads to more accurate negation processing (Ising et al., 2007).

3.2 Issues with Diagnosis Code Data

Another ED data element available for biosurveillance is the final diagnosis, which is widely available in electronic form and is standardized using the International Classifications of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) (USDHHS, 2006). All EDs generate electronic ICD-9-CM diagnoses as they are required for billing purposes (NCIPC, 1997; Travers et al., 2003). There is, however, some evidence that diagnosis data are not always available in a timely manner. In contrast to chief complaint data, which are generally entered into ED information systems by clinicians in real-time, ICD-9-CM diagnoses are often entered into the system by coders well after the ED visit. Sources of ICD-9-CM data may vary, which may influence the quality of the data. Traditionally, diagnoses have been assigned to ED visits by trained coders who are employed by the hospital and/or physician professional group. The primary purpose of the coding is billing, as opposed to secondary uses such as surveillance. Recently, emergency department information systems (EDIS) have come on the market that allow for diagnosis entry by clinicians. These systems typically include drop-down boxes with text that corresponds to ICD-9-CM codes; clinicians can then select a "clinical impression" at the end of the ED visit and the corresponding ICD-9-CM code becomes part of the EDIS data available for surveillance.

In a 2003 study of regional surveillance systems in North Carolina and Washington, biosurveillance developers found that over half of the EDs did not have electronic diagnosis data until 1 week or more after the ED visit (Travers et al., 2003). In a follow up study, researchers prospectively measured the time of availability of electronic ICD-9-CM codes in NC DETECT for all ED visits on 12/1/05 (Travers et al., 2006). The study confirmed that fewer than half of the EDs sent diagnoses within 1 week of the visit, and that it took 3 weeks to get at least one diagnosis for two-thirds of the visits. Seven (12%) of the hospitals had diagnoses for less than two-thirds of their ED visits at the 12 week mark. Diagnosis data are universally available from NC EDs, and studies have shown that ICD-9-CM data alone or in combination with chief complaint data are more valid than chief complaint data alone for syndromic surveillance (Beitel et al., 2004; Fleischauer et al., 2004; Reis & Mandl, 2004). This study corroborated the earlier study, however, that indicated the majority of North Carolina hospitals cannot send diagnosis data soon enough for timely, population-based biosurveillance.

4. BIOSURVEILLANCE IN NORTH CAROLINA

The North Carolina Emergency Department Database (NCEDD) project, spearheaded by the UNC Department of Emergency Medicine (UNC DEM) in 1999, laid the groundwork for electronic ED data collection in North Carolina by developing best practices for collecting and standardizing quality ED data. The focus on using ED data in North Carolina specifically for biosurveillance began in 2002 through a collaboration between UNC DEM and the North Carolina Division of Public Health (NC DPH). In 2004, a partnership between the North Carolina Hospital Association (NCHA) and NC DPH was instrumental in establishing ED data transmissions from the hospitals not yet participating in NC DETECT, including support for a new law mandating reporting as of January 1, 2005 (North Carolina General Statute 130A, http://www.ncleg.net/EnactedLegislation/Statutes/HTML/ ByChapter/ Chapter_130A.html, accessed January 17, 2008). As of October 1, 2010, there are 112/114 (98%) acute care, 24/7 hospital EDs submitting over 11,000 new visits to NC DETECT daily. These data are also transmitted twice daily to CDC's BioSense program.

In addition to ED data, NC DETECT receives data hourly from the statewide Carolinas Poison Center, and daily data feeds from the statewide Emergency Medical System (EMS) data collection center, a regional wildlife center, selected urgent care centers, and three laboratories of the NC State College of Veterinary Medicine (microbiology, immunology and vector-borne diseases laboratories) (Waller et al., 2008).

NC DETECT assists local, regional and state public health professionals and hospital users in identifying, monitoring, and responding to potential terrorism events, man-made and natural disasters, human and animal disease outbreaks and other events of public health significance. This system makes it possible for public health officials to conduct daily surveillance for clinical syndromes that may be caused by infectious, chemical or environmental agents. Suspicious syndromic patterns are detected using the CDC's EARS CUSUM algorithms, which are embedded in the NC DETECT Java-based Web application. The system also provides broader surveillance reports for ED visits related to hurricanes, injuries, asthma, vaccine-preventable diseases and environmental health (Waller et al., 2008). Role-based access provides hospital and public health access to NC DETECT data at local, regional and state levels; multi-tiered access provides tight controls on the data and allows all types of users to access the system, from those who need only an aggregated view of the data, to those who are able to decrypt sensitive protected health information when needed for investigation.

NC DETECT provides an excellent example of an early event detection and situational awareness system using ED visit data for disease surveillance. It is well established, statewide, and utilized daily by a variety of public health practitioners. A recently completed study found that NC ED information in NC DETECT compared favorably with national estimates of ED data made by the National Hospital Ambulatory Care Survey, despite differences in data collection methods (Hakenewerth et al., 2008). This finding is an indication of a well designed and robust system (Aylin et al., 2007).

4.1 History of Syndrome Definitions in NC

The syndromes monitored in NC DETECT are derived from the CDC's text-based clinical case definitions for bioterrorism syndromes (CDC, 2003). These syndromes were selected because they encompass both potential bioterrorism-related and community acquired disease processes. They include botulism-like illness (botulism), fever-rash (anthrax, bubonic plague, smallpox, tularemia, varicella), gastrointestinal (gastrointestinal anthrax, food/water-borne gastrointestinal illness, viral gastroenteritis), influenza-like-illness (epidemic influenza, pandemic influenza), meningoencephalitis (meningitis, encephalitis) and respiratory (respiratory anthrax, pneumonic plague, tularemia, influenza, SARS). Clinical case definitions are converted to syndrome definitions by expressing them in SQL, in most cases requiring both a syndrome specific and a constitutional keyword in either the chief complaint or triage note field. For example, a record containing the syndrome specific term "cough" and the constitutional term "fever" would match the respiratory syndrome. Documentation of a fever by vital sign measurement in the ED is also accepted in lieu of a constitutional keyword. The SQL code is written to identify common synonyms, acronyms, abbreviation, truncations, misspellings and negation in the free text data. The NC DETECT syndrome definitions have been modified over several years in an iterative fashion according to the knowledge and experience of the NC DETECT Syndrome Definition Workgroup. This workgroup meets monthly and includes public health epidemiologists who are regular users of NC DETECT for biosurveillance at the state and local levels, as well as clinicians and technical staff at NC DETECT. The continued improvement of the syndrome definitions for the purposes of syndromic surveillance requires more than this local effort, however. It requires collaboration with other system developers to determine the best practices nationally, as well as evidence-based research to support existing practices and/or develop new methodologies.

4.2 The Importance of Data Quality

The effectiveness of systems such as NC DETECT depends on the quality of the data provided by the data sources and on the system's capacity to collect, aggregate and report information. Perfect data, however, rarely exist and there are no perfect data systems. Thus, assessing and improving data quality must be ongoing tasks.

In NC DETECT, both automated and manual data quality checks are conducted daily and weekly. A Data Quality Workgroup meets monthly to review ongoing data quality concerns and strategize ways to address them. Major data quality issues range from failure to submit data at all to incorrect mapping of county of residence to extended delays in submitting diagnosis and procedure code data. Issues of particular concern include incomplete daily visit data, missing chief complaint data, failure to submit data in a timely fashion, and submission of invalid codes. Successfully addressing ED data quality issues requires constant monitoring of the data and ongoing communication with the hospitals submitting the data to NC DETECT.

4.3 NC DETECT Case Studies

NC DETECT has been used for a variety of public health surveillance needs including, but not limited to, early event detection, public health situational awareness, case finding, contact tracing, injury surveillance and environmental exposures (Waller et al., 2008). Those disease outbreaks that are first identified by traditional means are still aided by ED-based surveillance systems for identifying additional suspected cases and documenting the epidemiology of the affected individuals.

4.3.1 Public Health Surveillance During and After Hurricanes

Several major hurricanes have made landfall or passed through North Carolina in the past 10 years, including Floyd in 1999, Isabel in 2003, and Ophelia in 2005. In addition, hundreds of Katrina evacuees entered North Carolina in August and September 2005. While ED data were used in all instances to monitor the hurricanes' effects, the methodologies used show the evolution of ED data collection for public health surveillance in North Carolina.

In the fall of 1999, Hurricane Floyd, preceded by Hurricane Dennis and followed by Hurricane Irene, caused massive rainfalls that flooded eastern regions of North Carolina along the Neuse, Tar, Roanoke, Lumbar and Cape Fear Rivers. As NCEDD was still in early development in 1999, a disaster response team and ED staff in 20 hospitals worked together to define and apply standardized illness and injury classifications in order to conduct surveillance for the period of September 16 to October 27, 1999 and to compare results to similar periods in 1998. These classifications were applied manually based on diagnosis or chief symptoms for each patient visit abstracted from daily ED logs. Based on these analyses, Hurricane Floyd resulted in increases in insect stings, dermatitis, diarrhea and mental health issues as well as hypothermia, dog bites and asthma. The leading cause of death related to the storm was drowning (CDC, 2000). Surveillance for this time period required the dedicated efforts of EIS officers, medical students and other field staff, as well as ED staff and public health epidemiologists over an extended time period.

NC DPH conducted similar surveillance after Hurricane Isabel in 2003, manually surveying 35 hospitals to document hurricane-related morbidity and mortality (Davis et al., 2007). Officials updated the survey instrument to collect more information on injuries and asked hospitals to complete and fax the information to NC DPH. While less labor intensive overall than the surveillance that took place after Hurricane Floyd, the reliance on ED staff to provide information resulted in a relatively slow and extended collection of data from EDs.

Federal officials evacuated two large groups to North Carolina from Katrina-hit areas of the Gulf Coast in August and September 2005. For this event, NC DPH relied on NC DETECT and hospital-based public health epidemiologists in Wake and Mecklenburg counties for ED data collection. While the epidemiologists at two hospitals were able to identify more Katrina-related visits ($n = 105$) than the automated NC DETECT reports ($n = 90$), the NC DETECT reports required no manual tabulations and took only 2 h to develop and implement. In addition, the epidemiologist count included patients not included in the NC DETECT database, such as patients who were directly admitted to the hospital, without receiving treatment in the ED (Barnett et al., 2007). Furthermore, during the time the Katrina evacuee visits were being monitored, Ophelia approached the NC coast, where it stalled and resulted in the evacuation of coastal communities for several days. NC DETECT was used to monitor Ophelia-related ED visits simultaneously with the Katrina evacuee monitoring effort.

While manual tabulations may result in greater specificity, near real-time automated ED data collection for post-disaster surveillance provides a very low cost approach for monitoring the public's health if a system is already in place and operational. Queries can be continually refined to capture specific keywords in the chief complaint and triage note fields without added burden to hospital and/or public health staff. ED data collection provides an excellent complement to rapid needs assessments and other on-the-ground survey tools. Automated ED data collection assumes that EDs remain operational and that computerized health information systems continue to be used in times of mass disaster, an assumption that has not yet been put to the test in North Carolina.

4.3.2 Influenza

The NC DETECT influenza-like illness (ILI) definition, based on ED data, is used to monitor the influenza season in NC each year. The ED ILI definition follows the same trend as North Carolina's traditional, manually tabulated Sentinel Provider Network but is available in near real-time, as shown in Figure 3-2.

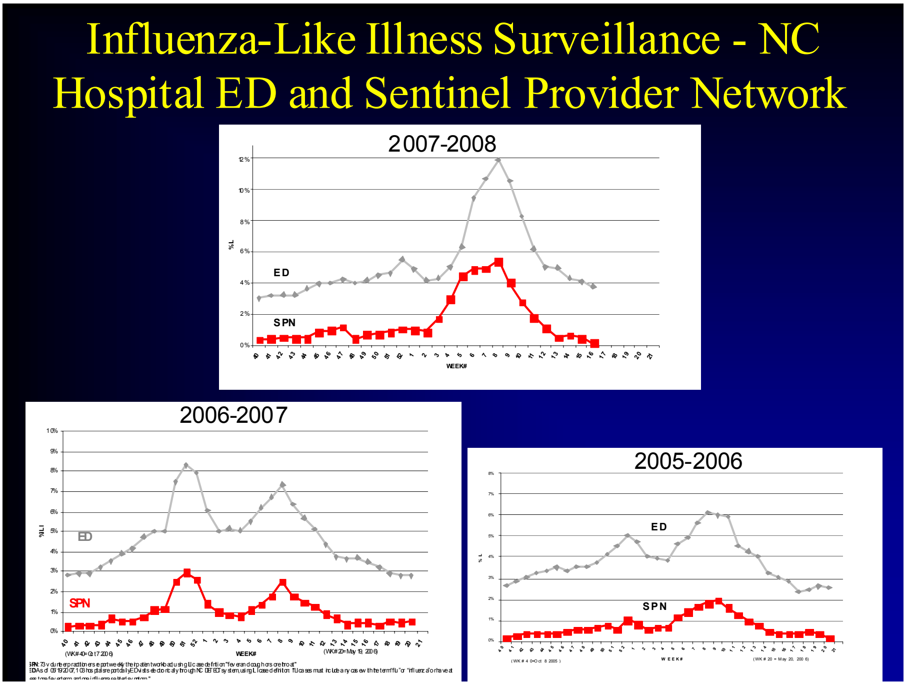

Influenza-Like Illness Surveillance - NC Hospital ED and Sentinel Provider Network

Figure 3-2. Comparing ED ILI (ED) from NC DETECT to the traditional sentinel provider network (SPN)

While North Carolina continues to maintain its Sentinel Provider Network, monitoring influenza with ED data provides several superior surveillance capabilities. In addition to timeliness, collecting ED data for influenza surveillance allows jurisdictions to assess impact on populations rather than samples, test case definition revisions on historical data, stratify ED visits by disposition type (admitted vs. discharged) and incorporate age stratification into analyses. The use of age groups in influenza surveillance has been shown to provide timely and representative information about the age-specific epidemiology of circulating influenza viruses (Olsen et al., 2007). Several states and large metropolitan areas, along with North Carolina, transmit aggregate ED-based ILI counts by age group to an ISDS-sponsored proof-of-concept project called the Distributed Surveillance Taskforce for Real-time Influenza Burden Tracking (DiSTRIBuTE). Although the ILI case definitions are locally defined, the visualizations that DiSTRIBuTE provides show striking similarities in ILI trends across the country (http://www.syndromic.org/projects/DiSTRIBuTE.htm).

4.3.3 Early Event Detection

While syndromic surveillance systems have clearly shown benefit for public health situational awareness and influenza surveillance, early event detection has been more of a challenge. Symptom-based detection systems are often overly sensitive, resulting in many false positives that can drain limited resources (Baer et al., 2007; Heffernan et al., 2004). Hospital and public health users who incorporate syndromic surveillance into their daily work flows, however, are able to accommodate these false positives more efficiently and still derive benefit from monitoring ED data for potential events. Investigating aberrations based on ED data that do not result in detecting an outbreak can still be important to confirm that nothing out of the ordinary is occurring. A recent investigation of gastrointestinal signals in Pitt County, North Carolina, for example, resulted in more active surveillance by the health department (checking non-ED data sources for similar trends) and the hospital (increased stool testing), as well as a health department press release promoting advice for preventing foodborne illnesses. Although a true outbreak or signal causative agent was not detected, this work results in improved coordination and communication among the hospital, healthcare providers and health department, which will make collaboration more efficient in any future large scale response efforts.

4.3.4 Bioterrorism Agent Report

To complement the more sensitive symptom-based syndromes, system developers may also include reports looking for specific mention of Category

A bioterrorism agents, such as anthrax, botulism, etc. In NC DETECT, for example, the Bioterrorism (BT) Agent Case Report searches for keywords and ICD-9-CM diagnoses related to 21 different bioterrorism agent groups. A statewide search on all 21 agents on average returns only ten cases (averaging one case a day over 10 days). In comparison to the specificity of this report, a statewide search on botulism-like illness for 10 days in NC DETECT produces approximately 200 cases while a search on a broad definition of gastrointestinal illness produces approximately 16,000 cases statewide over a 10-day period.

While the BT agent case report does include false positive cases, it provides an effective, unobtrusive monitoring mechanism that complements the astute clinician. It is also an important backup when notifiable diseases go unreported to the public health department, which actually occurred in March 2008 with a single case of tularemia.

4.3.5 Case Finding & Infectious Disease Outbreak Monitoring

Similar to the periods during and after natural disasters, monitoring ED data during a known infectious disease outbreak can assist with case finding and contact tracing. During known outbreaks, NC DETECT is used to identify potential cases that may require follow up. To assist in this effort, the NC DETECT Custom Event Report allows users to request new reports in just 2 h, with specific keyword and/or ICD-9-CM diagnostic criteria (Ising et al., 2007). This report has assisted North Carolina's public health monitoring in several events, including, but not limited to, nationwide recalls of peanut butter (February 2007), select canned foods (July 2007), nutritional supplements (January 2008), as well as localized Hepatitis A (January 2008) and Listeriosis (December 2007) outbreaks. Allowing users to access reports with very specific keywords (e.g., "peanut," "canned chili," "selenium") provides them with an efficient, targeted mechanism for timely surveillance of emerging events, all with the intention of reducing morbidity and mortality.

4.3.6 Infectious Disease Retrospective Analyses

When syndromic surveillance systems collect ICD-9-CM diagnosis codes in addition to chief complaints, users can conduct retrospective analysis effectively. For example, users can search on the ICD-9-CM code V04.5 (need for prophylactic vaccination and inoculation against certain viral diseases: rabies) to review how many ED patients received rabies prophylaxis in a given time period. Using the ED chief complaint, users can go a step further and view how many ED patients with chief complaints related to animal bites/animal exposures were NOT documented as having received a V04.5 code. Investigation of the results may reveal hospital coding errors or hospital practices that are not in line with public health requirements that can then be corrected.

4.3.7 Injury Surveillance

The Injury and Violence Prevention Branch of NC DPH has added ED data from NC DETECT to its data sources for injury surveillance efforts. In addition to ED visit data, they also use hospital discharge, death certificate, and medical examiner data. Injury surveillance efforts involving ED data have included falls, traumatic brain injury, fire-related injury, self-inflicted injury, heat-related visits, and unintentional drug overdoses. Furthermore, they have used ED data when working with trauma Regional Advisory Committees to evaluate injury patterns and are exploring the possibility of incorporating ED data into NC's violent death reporting system. While ED data have long been used for injury surveillance, the availability of near real-time data provides opportunities for more timely documentation of intervention outcomes.

4.4 Conclusions and Discussion

ED data are a proven tool for biosurveillance, and the ED data in NC DETECT have proved to be effective for a variety of public health uses, including surveillance, monitoring and investigation. Biosurveillance systems that are incorporated into hospital and public health practitioner daily work flows are more effective and easily used during a public health emergency. The flexibility of a system such as NC DETECT helps it meet this level of functionality.

4.5 Evaluation of NC DETECT

Any surveillance system should undergo rigorous evaluation to make sure it is meeting user needs effectively and efficiently. The ED data stream of NC DETECT has undergone two such evaluations. In 2007, it was evaluated by the North Carolina Center for Public Health Preparedness at the charge of the NC DPH. The evaluation was designed to determine the usefulness of the ED data and the ease with which it is used for both real-time and non-real-time public health surveillance activities. Interviews were conducted with stakeholders to learn about the specifics of the ED data, data flow, and the aberration detection algorithms. In addition, local, regional and state public health authorities, as well as hospital-based public health epidemiologists (PHEs), were asked to complete a Web-based survey about their experience using the ED data via NC DETECT. Key findings included:

- ED data permit public health authorities to identify human health events as a result of bioterrorism, natural or accidental disaster, or infectious disease outbreak, but the rapidity of detection is contingent on the extent of the event and affected individuals, the ability of chief complaint data to be successfully transmitted to NC DETECT in a timely manner, and the frequency and timing of aberration detection and investigation by public health authorities;
- The NC statute mandating provision of ED visit data for public health surveillance and the availability of UNC DEM staff to provide technical and analytical expertise have been instrumental in assuring that timely, quality data are available for public health surveillance;
- ED data are useful to public health authorities;
- The system showed a low positive predictive value (PPV), indicating that users must examine a large number of false positives in order to identify a potentially true threat to public health.

Based on these findings, this evaluation recommended additional efforts to encourage public health authorities to routinely use the ED data, increased communication among hospitals, business organizations and public health authorities, examination and evaluation of different aberration detection algorithms, and a cost-benefit study of using ED data for public health surveillance.

A second evaluation of the emergency department data stream of NC DETECT was conducted in 2007 by the Research Triangle Institute to assess the impact of this biosurveillance system on public health preparedness, early detection, situational awareness, and response to public health threats. This study used key informant interviews and found the following:

- Biosurveillance has been used in North Carolina for situational awareness and early detection of disease outbreaks;
- Public health epidemiologists in hospitals and regional state-based response teams have integrated use of NC DETECT with traditional surveillance activities;
- Biosurveillance has added timeliness and flexibility to traditional surveillance, increased reportable disease reporting and case finding, and increased public health communication.

This evaluation recommended the addition of urgent care center data to complement the ED visit data for biosurveillance and exploring the use of diagnosis data, when available in a timely manner, to minimize false positive alerts.

5. CONCLUSION

A distinctive feature of ED data for surveillance is their timeliness. With electronic health information systems, these data are available in near real-time, making them particularly useful for surveillance and situational awareness in rapidly developing public health outbreaks or disasters. The use of ED data for public health surveillance can significantly increase the speed of detecting, monitoring and investigating public health events. Combined with other timely data sources such as data from poison centers, EMS, ambulatory care data, and animal health data, ED data analyses are an important source of information for mitigating the effects of infectious disease.

ACKNOWLEDGEMENTS

Financial support for NC DETECT is currently provided by the North Carolina Division of Public Health Contract 01322–11 and the Department of Emergency Medicine of the University of North Carolina at Chapel Hill. The authors acknowledge the collaboration of Lana Deyneka, MD, MPH, of the North Carolina Division of Public Health, and Scott Wetterhall, MD, MPH, and Anne-Lyne McCalla, MPH, of RTI, International in the syndromic surveillance research and development efforts of NC DETECT through BioSense grant R01 PH000038–0.

QUESTIONS FOR DISCUSSION

1. Are timely ED data systems for public health surveillance cost effective? How would you measure this?
2. How can biosurveillance systems and electronic lab reporting for reportable conditions best complement each other?
3. What other data sources could and should be used with ED data for an exemplar biosurveillance system?
4. Can an automated biosurveillance system ever really replace the astute clinician at detecting and responding to an infectious disease outbreak of public health significance?
5. What statistical approaches are available for aberration detection and what are the pros and cons of each? How does a biosurveillance system determine which aberration detection method(s) to use?
6. What are the major data quality issues related to conducting public health surveillance with ED data? How can these be identified and addressed?

7. Discuss the challenges and benefits of using secondary ED visit data for public health surveillance.
8. What are some of the security and privacy issues surrounding the use of ED visit data for public health surveillance?

REFERENCES

Aylin P, Bottle A, Majeed A. Use of Administrative data or clinical databases as predictors of risk of death in hospital: comparison of models. BMJ 2007; 334:1044. Originally published online 23 Apr 2007.

Baer A, Jackson M, Duchin JS. What is the value of a positive syndromic surveillance signal? Advances in Disease Surveillance 2007; 2:192.

Barnett C, Deyneka L, Waller AE. Post-Katrina situational awareness in North Carolina. Advances in Disease Surveillance 2007; 2:142.

Beitel AJ, Olson KL, Reis BY, Mandl KD. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatric Emergency Care 2004; 20(6):355–360.

Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks. MMWR. Recommendations and Reports 2004; 53(RR05):1–11.

Centers for Disease Control and Prevention. Morbidity and mortality associated with Hurricane Floyd – North Carolina, September-October 1999.MMWR. Morbidity and Mortality Weekly Report 2000; 49(17):369–372.

Centers for Disease Control and Prevention (October 23, 2003). Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Available at http://www.bt.cdc.gov/surveillance/syndromedef/index.asp. Accessed June 24, 2005.

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics 2001; 34(5):301–310.

Chapman WW and Dowling JN. Consultative meeting on chief complaint classifiers and standardized syndromic definitions. Advances in Disease Surveillance 2007; 4:47.

Dara J, Dowling JN, Travers D, Cooper GF, Chapman WW. Chief complaint preprocessing evaluated on statistical and non-statistical classifiers. Advances in Disease Surveillance 2007; 2:4.

Davis MV, MacDonald PDM, Cline JS, Baker EL. Evaluation of public health response to hurricanes finds North Carolina better prepared for public health emergencies. Public Health Reports 2007; 122:17–26.

Davis SF, Strebel PM, Atkinson WL, Markowitz LE, Sutter RW, Scanlon KS, Friedman S, Hadler SC. Reporting efficiency during a measles outbreak in New York City, 1991. American Journal of Public Health 1993; 83(7):1011–1015.

Espino JU, Dowling JN, Levander J, Sutovsky P, Wagner MM, Cooper GF. SyCo: A probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. Advances in Disease Surveillance 2007; 2:5.

Fleischauer AT, Silk BJ, Schumacher M, Komatsu K, Santana S, Vaz V, et al. The validity of chief complaint and discharge diagnosis in emergency-based syndromic surveillance. Academic Emergency Medicine 2004; 11(12):1262–1267.

Forbach C, Scholer MJ, Falls D, Ising A, Waller A. Improving system ability to identify symptom complexes in free-text data. Advances in Disease Surveillance 2007; 2:7.

Hakenewerth A, Waller A, Ising A, Tinitinalli J. NC DETECT and NHAMCS: Comparison of emergency department data. Academic Emergency Medicine 2009; 16(3):261–269.

Heffernan R, Mostashari F, Das D, Besculides M, Rodriguez C, Greenko J, et al. New York City syndromic surveillance systems. MMWR. Morbidity and Mortality Weekly Report Supplement 2004; 24(53):23–27.

Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis [serial on the Internet]. 2004 May http://www.cdc.gov/ncidod/EID/vol10no5/03–0646.htm. Accessed July 20, 2007.

Hirshon JM. For the SAEM Public Health Task Force Preventive Care project. The rational for developing public health surveillance systems based on emergency department data. Academic Emergency Medicine 2000; 7:1428–1432.

Hripscak G, Bamberger A, Friedman C. Fever detection in clinic visit notes using a general purpose processor. Advances in Disease Surveillance 2007; 2:14.

Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System. Journal of Urban Health. 2003; 80 (2 Suppl 1): i89–i96.

Ising A, Li M, Deyneka L, Barnett C, Scholer M, Waller A. Situational awareness using web-based annotation and custom reporting. Advances in Disease Surveillance 2007; 4:167.

Ising A, Travers D, Crouch J, Waller A. Improving negation processing in triage notes. Advances in Disease Surveillance 2007; 4:50.

Ising AI, Travers DA, MacFarquhar J, Kipp A, Waller A. Triage note in emergency department-based syndromic surveillance. Advances in Disease Surveillance 2006; 1:34.

Kaufman AF, Meltzer MI, Schmid GF. Economic impact of a bioterrorist attack: are prevention and postattack intervention programs justifiable? Emerging Infectious Diseases 1997; 3:83–94.

Komatsu K, Trujillo L, Lu HM, Zeng D, Chen H. Ontology-based automatic chief complaints classification for syndromic surveillance. Advances in Disease Surveillance 2007; 2:17.

Lee LE, Fonseca V, Brett KM, Sanchez J, Mullen RC, Quenemoen LE, Groseclose SL, Hopkins RS. Active morbidity surveillance after Hurricane Andrew--Florida, 1992. JAMA 1993; 270(5):591–594. Erratum in: JAMA 1993; 270(19):2302.

Lober WB, Karras BT, Wagner MM, et al. Roundtable on bioterrorism detection: information system-based surveillance. Journal of the American Medical Informatics Association 2002; 9(2):105–115.

Lombardo J. A systems overview of the Electronic Surveillance System for Early Notification of Community-based Epidemics (ESSENCE II). Journal of Urban Health 2003; 80(2 Suppl 1):i32–i42.

Mikosz CA, Silva J, Black S, Gibbs G, Cardenas I. Comparison of two major emergency department-based free-text chief-complaint coding systems. MMWR. Morbidity and Mortality Weekly Report Supplement 2004; 53S:101–105.

National Center for Injury Prevention and Control (US). Data elements for emergency department systems, release 1.0. Atlanta, GA: Centers for Disease Control; 1997.

National Library of Medicine (2006). Specialist Lexicon and Lexical Tools Documentation, 2006AD. Retrieved May 11, 2007 from http://www.nlm.nih.gov/research/umls/meta4.html.

National Library of Medicine (2007). Unified Medical Language System Documentation, 2007AA. Bethesda, MD: National Library of Medicine. Retrieved May 11, 2007 from http://www.nlm.nih.gov/research/umls/meta2.html.

Nawar EW, Niska RW, Xu J. (2007). National Hospital Ambulatory Medical Care Survey: 2005 Emergency Department Summary. Advance data from vital and health statistics: no. 386. Hyattsville, MD: National Center for Health Statistics.

North Carolina General Statute 130A. http://www.ncleg.net/EnactedLegislation/ Statutes/ HTML/ByChapter/Chapter_130A.html Accessed January 17, 2008.

Olson DR, Heffernan RT, Paladini M, Konty K, Weiss D, Mostashari F. Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City. PLoS Medicine. 2007; 4(8):e247.

Reis BY, Mandl KD. Syndromic surveillance: The effects of syndrome grouping on model accuracy and outbreak detection. Annals of Emergency Medicine 2004; 44:235–241.

Rodewald LE, Wrenn KD, Slovis CM. A method for developing and maintaining a powerful but inexpensive computer data base of clinical information about emergency department patients. Annals of Emergency Medicine 1992; 21:41–46.

Rydman RJ, Rumoro DP, Silva JC, Hogan TM, Kampe LM. The rate and risk of heat-related illness in hospital emergency departments during the 1995 Chicago heat disaster. Journal of Medical Systems 1999; 23(1):41–56.

Shapiro A. Taming variability in free text: Application to health surveillance. MMWR. Morbidity and Mortality Weekly Report 2004; 24(53S):95–100.

Sosin DM, DeThomasis J. Evaluation challenges for syndromic surveillance – Making incremental progress. MMWR. Morbidity and Mortality Weekly Report 2004; 53S:125–129.

Talan DA, Moran GJ, Mower WR, Newdow M, Ong S, Slutsker L, Jarvis WR, Conn LA, Pinner RW. EMERGEncy ID NET: an emergency department-based emerging infections sentinel network. The EMERGEncy ID NET Study Group. Annals of Emergency Medicine 1998; 32:703–711.

Teich JM, Wagner MM, Mackenzie CF, Schafer KO. The informatics response in disaster, terrorism and war. Journal of the American Medical Informatics Association 2002; 9(2):97–104.

Travers DA. (2006). Emergency Medical Text Processor. Accessed on June 10, 2008 from http://nursing.unc.edu/research/current/emtp/.

Travers DA, Barnett C, Ising A, Waller A. (2006). Timeliness of emergency department diagnoses for syndromic surveillance. Proceedings of the AMIA Symposium 2006, 769–773.

Travers DA, Haas SW. Using nurses natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. Journal of Biomedical Informatics 2003; 36:260–270.

Travers DA, Waller A, Haas S, Lober WB, Beard C. (2003). Emergency department data for bioterrorism surveillance: Electronic availability, timeliness, sources and standards. Proceedings of the AMIA Symposium 2003, 664–668.

Tsui FC, Wagner MM, Dato V, Chang C. (2001) Value of ICD-9-coded chief complaints for detection of epidemics. Proceedings of the Fall Symposium of the American Medical Informatics Association 2001, 711–715.

U.S. Department of Health and Human Services (USDHHS). International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). 6th ed. Washington: Author; 2006.

Varney SM, Hirshon JM. Update on public health surveillance in emergency departments. Emergency Medicine Clinics of North America 2006; 24:1035–1052.

Waller AE, Ising AI, Deneka L. North Carolina Biosurveillance System. In: Voeller JG, editor. Wiley Handbook of Science and Technology for Homeland Security. New York: Wiley; April 2010, vol.3.5.

Waller AE, Ising AI, Deyneka L. North Carolina Emergency Department visit data available for public health surveillance. North Carolina Medical Journal 2007; 68(4):289–291.

Wagner MM, Espino J, Tsui FC, et al. Syndrome and outbreak detection using chief-complaint data – experience of the Real-Time Outbreak and Disease Surveillance project. MMWR. Morbidity and Mortality Weekly Report 2004; 24(53S):28–31.

Weiss BP, Mascola L, Fannin SL. Public health at the 1984 Summer Olympics: the Los Angeles County experience. American Journal of Public Health 1988; s78(6):686–688.

SUGGESTED READING

Hirshon JM. For the SAEM Public Health Task Force Preventive Care project. The rational for developing public health surveillance systems based on emergency department data. Academic Emergency Medicine 2000; 7:1428–1432.

Varney SM, Hirshon JM. Update on public health surveillance in emergency departments. Emergency Medicine Clinics of North America 2006; 24:1035–1052.

Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks. MMWR. Recommendations and Reports 2004; 53(RR05):1–11.

Fleischauer AT, Silk BJ, Schumacher M, Komatsu K, Santana S, Vaz V, et al. The validity of chief complaint and discharge diagnosis in emergency-based syndromic surveillance. Academic Emergency Medicine 2004; 11(12):1262–1267.

Centers for Disease Control and Prevention (October 23, 2003). Syndrome definitions for diseases associated with critical bioterrorism-associated agents. Available at http://www.bt.cdc.gov/surveillance/syndromedef/index.asp. Accessed June 24, 2005.

Nawar EW, Niska RW, Xu J. (2007). National Hospital Ambulatory Medical Care Survey: 2005 Emergency Department Summary. Advance data from vital and health statistics: no. 386. Hyattsville, MD: National Center for Health Statistics.

ONLINE RESOURCES

Website for the International Society for Disease Surveillance (http://www.syndromic.org) This website includes the online journal Advances in Disease Surveillance, as well as a variety of wikis addressing research and public health topics.

NC DETECT website (http://www.ncdetect.org) The NC DETECT website contains links to numerous abstracts and presentations related to ED data use for biosurveillance, as well as details on the technical components of NC DETECT.

Chapter 4

CLINICAL LABORATORY DATA FOR BIOSURVEILLANCE

EILEEN KOSKI

CHAPTER OVERVIEW

This chapter provides general background on the types of data used for biosurveillance, syndromic surveillance and traditional public health surveillance, with particular emphasis on the use of clinical laboratory data. Overviews of types of surveillance, the history of clinical laboratory testing and different types of laboratories are provided as background. A more detailed discussion of the roles and characteristics of clinical laboratory data with respect to biosurveillance includes comments on data standards and analytic requirements, with particular emphasis on understanding possible sources of variability or artifacts during data analysis and interpretation.

Keywords: Biosurveillance; Electronic laboratory reporting; Laboratory data; Public health surveillance; Syndromic surveillance

1. INTRODUCTION

Biosurveillance may be viewed as an extension of public health surveillance. The term biosurveillance is typically defined as the process of monitoring selected aspects of the health of a population in such a way as to facilitate early detection and monitoring of an event of public health concern, as well as supporting situational awareness and response to the event. In terms of data, biosurveillance utilizes a broader range of data types and sources than traditional public health surveillance.

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_4,

From a historical perspective, there have been dramatic changes in the volume and types of data available for public health surveillance since the sixteenth century when the introduction of the London Bills of Mortality [1] represented an early effort to systematically collect data on causes of death. Since that time, the scope, types, quantities and level of detail of the data collected have continued to expand. The availability of data and the ability to process it efficiently have progressed, of necessity, along parallel paths, particularly since the introduction of computers into healthcare. The earliest applications of computing to healthcare began in the 1950s, escalating with the development of MUMPS (Massachusetts General Hospital Utility Multi-Programming System) in 1966 and continuing to expand with the emergence of specialized medical computing professional societies [2].

The pace of these developments has escalated dramatically as computing power in general, and advances in health information technology (HIT) in particular, have continued to accelerate. This has, in turn, enabled both access to new data streams as well as supporting new, and increasingly complex, analytic technologies. As a result, a variety of new approaches to surveillance have been developed, each of which may take advantage of – and be best served by – different types of data and detection algorithms.

Clinical laboratory data represents a critical information component of data used for biosurveillance in multiple contexts. Specific elements of laboratory data have proven particularly well suited to different surveillance strategies, but virtually all forms of public health surveillance rely on laboratory data in one form or another.

2. TYPES OF SURVEILLANCE

There are many forms of public health surveillance focused on various aspects of population health including infectious disease, chronic disease, birth defects, mental health and even social issues. For purposes of this discussion, the principle distinction of interest relates to the following three broad categories:

Traditional public health surveillance for infectious disease
Syndromic surveillance
Biosurveillance.

In most current forms of what may be referred to as “traditional” public health surveillance, the goal is to detect evidence of a known, direct threat to population health. Two of the primary sources of data currently used are case reports from routine clinical practice, including case reports and surveillance-oriented laboratory testing from sentinel providers or networks, and direct reporting by clinical laboratories of positive laboratory results for known conditions of interest [3]. The requirements for what is reported from providers and clinical laboratories are generally specified in regulations and laws governing reporting to public health agencies. Some conditions are nationally notifiable [4], while others may only be reportable in specific jurisdictions where the condition is of particular concern. Some sentinel providers and networks, such as NREVSS, the National Respiratory and Enteric Viral Surveillance System, are based on data collected from collaborating university and community hospital laboratories, selected state and county public health laboratories, and commercial laboratories through a specific cooperative agreements, as opposed to regulatory requirements.

This form of surveillance is described as passive surveillance because there is no direct attempt to seek out possible cases, relying instead on identification of cases that present to the healthcare system in the normal course of medical care [5]. In the case of sentinel providers, there may be an agreement to test patients presenting with symptoms of a particular condition, such as influenza, that would not normally prompt testing in routine clinical practice. This can be considered passive surveillance for purposes of this discussion, however, since it still relies on the patient presenting to the healthcare system. In active surveillance, by comparison, potentially exposed individuals are specifically sought out for testing by the health department.

In recent years, some of the shortcomings of this approach with respect to emerging and re-emerging infectious disease, as well as bioterrorist attacks, have led researchers to introduce a number of new surveillance approaches and techniques. These approaches have included both direct detection methods, such as water, air or other environmental sensing, as well as more mathematically probabilistic techniques, such as syndromic surveillance [6, 7]. These new approaches rely on many more types and sources of data than traditional public health surveillance, for example:

Chief complaint for hospital and/or emergency department admission
Hospital discharge data
Prescription drug sales
Over-the-counter drug sales
Calls to poison control centers or other medically oriented call-in centers
Visits to health-related Internet sites
School and work absenteeism
News articles with health-related content
Laboratory order data.

Each of these types of data has specific advantages and disadvantages, although most require some form of specialized processing of the data to make it usable for large scale analytic purposes. This may include techniques such as text parsing of chief complaint data or operationalization of variables such as drug names and laboratory tests into meaningful syndromic groupings.

Some of these data sources represent indirect and less-specific evidence of disease, such as school absenteeism rates [8], but may prove effective in detecting the presence of a condition that may be widespread and debilitating, although not sufficiently severe to warrant a hospital or physician visit. Others are more directly health-related but may be used in a new way. Sales of over-the-counter medications [9] can be used to detect less severe conditions for which patients are likely to self-medicate rather than visit a physician, such as diarrheal or flu-like illnesses. Patterns of laboratory orders, as opposed to just positive results [10, 11], can be used to detect the presence of more severe conditions that warrant interaction with the healthcare system, but that are sufficiently novel that no positive results for known conditions are reported.

In addition to identification of confirmed or suspected cases of a specific disease, some forms of surveillance may also look for significant change in the behavior or other characteristics of a known disease vector. Examples of this are the use of sentinel chickens and mosquito pools [12], in which laboratory testing is done on potentially exposed animals or on the vectors themselves.

The concept of biosurveillance broadens the scope of study even further to conditions that may not pose an imminent or direct threat to human health, but may pose a direct threat to our food supply or economy in the form of veterinary or plant contagions, as well as direct or indirect threats to health via other environmental contagions or toxins. Again, laboratory data is crucial to detection and identification of any suspected or actual contagion.

One data source that is common to all three different types of surveillance is the clinical laboratory.

2.1 Laboratory Data for Biosurveillance

Serology and culture results are generally considered the “gold standard” for confirmation of many infectious diseases, and clinical laboratory tests have been well established to have excellent sensitivity and positive predictive value. As a result, laboratory confirmation of a suspected case of an infectious disease has long been, and continues to be, a key component of public health case definitions [13, 14], as well as being an essential element in surveillance systems in general.

The role of laboratory data in detection and identification of known, current threats is generally well understood and accepted. Under certain circumstances, however, laboratory data may take on a new role. In some cases of diseases that have declined in prevalence, such as malaria, the institution of laboratory confirmation has been instrumental in assuring that suspected or presumed cases were no longer reported as confirmed cases [15, 16]. In the case of a new, emerging, or re-emerging condition, the availability of confirmatory laboratory testing may lag behind recognition of the condition. For a novel infectious agent, in particular, availability of confirmatory testing is clearly dependent upon identification of the etiologic agent. Even in such cases however, laboratory testing is typically crucial in ruling out alternate causative agents, as well as confirming the presence of physiologic outcomes that may be part of the public health case definition for a specific condition.

Although all laboratory-confirmed cases of notifiable conditions should, in theory, be reported by the clinicians caring for the affected patients, compliance varies considerably. From an organizational perspective, there are far fewer laboratories than individual clinicians, each typically serving a much larger patient base. As a result, laboratory-based surveillance has proven to be an effective avenue to improving the completeness of reporting of notifiable conditions to public health departments [17].

The field of syndromic surveillance emerged in recent years in an attempt to expedite recognition of an outbreak caused by a new or emerging organism or agent by looking for evidence of spikes in presentation of specific symptoms, in the absence of a known causation. In the case of laboratory data, this has required operationalization of laboratory test orders based on the typical symptoms for which those tests might be ordered, regardless of the actual test results [10, 11]. For example, a spike in orders for stool or CSF (cerebrospinal fluid) cultures might indicate the presence of a novel gastrointestinal or neurologic pathogen in a community.

2.2 The Clinical Laboratory

2.2.1 Development of the Clinical Laboratory

The identification of the causative agents for such devastating infections as diphtheria and cholera in the 1880s and 1890s, as well as the subsequent development of tests for these diseases, paved the way for the establishment of the earliest dedicated clinical laboratories at the end of the nineteenth century [18]. Prior to that time, laboratory testing was very limited and typically performed by individual physicians and scientists.

By 1926, recognition of the role of the clinical laboratory had become so firmly established that the American College of Surgeons' accreditation standards required all hospitals to have a clinical laboratory under the direction of a physician. This ensured that clinical laboratories would develop mainly in hospitals for some time [19].

Shortages of trained personnel during World War I led to the development of a variety of training programs as well as certification requirements for "medical technologists." As the number of medical technologists grew, they eventually sought greater autonomy and independence from the historical structure established by the American Society of Clinical Pathologists (ASCP). Independent clinical testing laboratories began to appear and by the late 1970s, clinical laboratory science was becoming an independent profession [20].

The earliest laboratory testing for infectious diseases was focused on bacterial organisms that could be observed using technologies such as the light microscope. As knowledge of both infectious organisms and related issues of public health importance, such as antibiotic resistance, continued to grow, the breadth and complexity of laboratory testing expanded as well. The identification of viruses as major causes of infectious disease also necessitated the introduction of new technologies. The current repertoire of infectious disease testing includes molecular diagnostic techniques, including viral genomic testing, that can enhance the speed, specificity and sensitivity of diagnosis and identification of etiologic agents [21]. Although the majority of clinical laboratory testing is still performed in hospital laboratories, the introduction of increasingly complex laboratory testing techniques contributed to the emergence of specialized reference laboratories.

2.2.2 Laboratory Types

There are a number of different ways to classify laboratories, such as those based on:

1. Types of testing – e.g., clinical, environmental, routine, esoteric
2. Bio-safety level
3. Market segment served – e.g., hospital-based versus out-patient.

In the context of laboratory data for biosurveillance, there are four broad laboratory segments that account for most testing performed today: public health, hospital, commercial, and reference, which includes esoteric and specialized laboratories. Some laboratory testing is also performed in physician office laboratories. From the perspective of biosurveillance, most laboratory data comes from hospital, commercial or reference laboratories. Public health laboratories perform crucial functions with respect to confirmation and characterization of rare or new pathogens, as well as providing testing services during active case investigation. The type of passive surveillance that forms the core of biosurveillance, however, relies on detecting cases identified during routine medical practice, in which case the testing is more likely to have been performed in a hospital, commercial or reference laboratory.

There is a great deal of variety as well as overlap in the menu of laboratory testing offered by hospital, commercial and reference laboratories. The breadth and depth of the test menu offered by any of these types of laboratories will also generally depend on the size of the institution. For example, A major academic medical center will typically offer a broader laboratory test menu than a small community hospital.

While there are few absolute distinctions between what may be offered in each segment of the laboratory industry, there are some characteristics that typically differentiate the segments.

Public health laboratories normally focus on issues affecting the health of the population, including newborn screening for potentially life-threatening metabolic and genetic disorders, monitoring communities for pathogens in food or water, and testing for newly emerging infectious diseases including rapid identification of suspect agents.

Hospital laboratories generally place a greater emphasis on testing that requires almost immediate attention, such as arterial blood gas analysis; testing requiring rapid turnaround at any hour of the day or night, such as the types of testing typically performed in emergency departments and intensive care units; and other testing with special clinical characteristics, such as surgical pathology.

Commercial laboratories typically focus more on the type of testing performed in an out-patient setting, as well as testing from residential facilities such as nursing homes.

Reference laboratories usually focus on specialized testing, such as endocrinology, infectious disease, molecular diagnostics, or genetic testing. In addition, it is not unusual for new laboratory testing technologies to be deployed initially in a reference laboratory, particularly if the testing is still considered experimental, prior to more widespread availability of a test.

These are broad distinctions intended only to highlight the typical distinctions between laboratory test menus that may be available from different segments. In reality, the situation can be very dynamic. Reference laboratories may exist within academic medical centers or as part of large commercial laboratories, and small hospitals may contract out parts of their routine testing to commercial laboratories, so very few absolute distinctions exist in practice. Since the severity of disease seen in hospitalized and emergency department (ED) patients are often very different from patients seen on an out-patient basis, it can be very useful to understand the source of laboratory data used for surveillance purposes in order to put any findings into the appropriate context.

2.2.3 Sources of Laboratory Data for Biosurveillance

Laboratory data is currently made available for biosurveillance, either directly from a clinical laboratory or indirectly through another provider, via at least four distinct mechanisms:

1. As the basis for a specific case report to a public health agency by an individual clinician or a hospital
2. As reported directly to public health agencies in compliance with regulations pertaining to notifiable conditions
3. As data provided to specific programs within one or more public health agencies in conjunction with specific surveillance programs conducted by those agencies, including sentinel networks, such as NREVSS
4. From data previously provided to a Health Information Exchange (HIE) or Regional Health Information Organization (RHIO).

In the future, it is anticipated that data may also be made available directly from electronic health records (EHR) in providers' offices. Only a small percentage of providers currently use EHRs, so this is not a major source of laboratory data for biosurveillance today, nor is it likely to be in the next few years [22]. As adoption increases, however, it is likely to be a more robust option in the future, and there is already some preliminary data indicating that there may be potential value in using such data for surveillance [23].

Historically, laboratory result data has been delivered to public health departments primarily via paper, as well as by fax and telephone. Although most medical records are still paper-based, laboratory data has largely been managed in electronic form in laboratory information systems for many years. As a result, routine public heath data from laboratories has increasingly been reported electronically, increasing the speed and reliability with which data is available for biosurveillance purposes. In addition, numerous public health agencies have reported that the switch to electronic laboratory reporting (ELR) has generally resulted in more complete reporting as well [24, 25].

As with most new technological approaches, the introduction of electronic laboratory reporting was not without difficulty, particularly for early implementations. Although more data has typically been received, and in some cases the data has been available very rapidly, unanticipated problems were encountered. One type of problem experienced in multiple sites was the reporting of false positives due to the use of automated text parsers for detection of organism names. In most cases, the text parsers had incorrectly reported negative reports as positive if both reports included a text representation of a specific organism name [25–27]. In a similar vein, incorrect automapping of disease names to electronic codes has been reported to have resulted in an erroneous report of a rare condition [28]. Another problem encountered was the difficulty in addressing repeated reporting of the same result, such as preliminary and final culture results [26, 27]. Although preliminary results were considered of value in expediting detection, the need to combine or "de-duplicate" the multiple reports necessitated additional processing of the data. Although electronic reporting is expected to expedite reporting, automated systems have also been observed to be much slower than traditional reporting or to have had unpredictable performance lapses [26]. The increased burden on health department resources that has sometimes been necessitated by dramatic increases in reporting has also been reported to pose a problem in some cases [29].

Such problems should be ameliorated over time as various sources of error are identified and systems become more sophisticated. In the interim, these are issues that need to be addressed during data analysis by carefully attempting to identify duplicates or unusual patterns of positives that could indicate the presence of large numbers of false positives.

2.2.4 Components of Laboratory Data for Biosurveillance

In traditional public health surveillance, the data records transmitted to public health departments from laboratories are those that meet the criteria established by each jurisdiction for notifiable conditions. The precise data elements required for each condition are normally described in the applicable regulations and may vary from jurisdiction to jurisdiction and program to program, but typically include the following:

Reporting laboratory identifiers
Patient identifiers and demographics
Ordering clinician identifiers
Identification of ordered test
Test results, including reference ranges and interpretation
Additional information as required.

In some cases, all test results for a specific condition may be provided to public health agencies, but in most cases only positive results are considered reportable.

Clinical laboratories must track the requirements of each jurisdiction in which they perform testing to be certain that they have complied with all the applicable reporting requirements. In the case of reference laboratories, the reporting requirements may remain with the originating laboratory, i.e., the laboratory that originally received the specimen from the ordering clinician. In cases where testing is referred from one laboratory to another, which is often the case for reference laboratory work, the requisite patient and provider identifiers and demographics may not be available to the reference laboratory. The reference laboratory may only know the state of origin of the ordering laboratory, not of the patient or even the provider, leaving the primary reporting obligation with the originating laboratory.

In the case of newer surveillance methodologies and programs, such as syndromic surveillance and biosurveillance programs, the data transmitted may be significantly different from that provided for regulatory purposes. This is true primarily because such programs, by definition, are attempting to address issues not fully supported by the current reporting system. In addition, while local pubic health reporting is covered by regulation and generally requires full patient identification, some agencies conducting syndromic surveillance and biosurveillance programs may not automatically have the right to access patient demographic data.

It should be noted that while provision of data in compliance with regulations governing notifiable diseases is covered under the HIPAA (Health Insurance Portability and Accountability Act) privacy provisions, as are uses of data to support the response to a bio-event, the precise status of other secondary uses of clinical data for biosurveillance purposes is more ambiguous [30]. Clinical laboratories are also required to comply with CLIA (Clinical Laboratory Improvement Amendments) regulations that include provisions governing provision of laboratory data.

As a result, one primary difference between notifiable case reporting and data provided for programs not covered by regulation is that data may be provided in de-identified, aggregated form or else patient identifiers are anonymized or pseudonymized. Pseudonymization consists of providing a special linker identifier that can be linked back to the original record by the originating facility, but that provides no patient identifiable data directly to the recipient. Anonymization theoretically precludes any such subsequent linkage to the original record. Pseudonymization is used to protect patient confidentiality while retaining the ability to identify specific patients in the event of compelling public heath need and appropriate authorization [31].

The second primary difference is that data sets provided for syndromic surveillance and biosurveillance programs typically encompass a much broader array of test orders and results than are required by most jurisdictions for routine public health reporting. At the same time, some public health jurisdictions may require reporting on selected chronic diseases, such as diabetes. From a biosurveillance perspective, the focus is more commonly on infectious agents or toxins, so data on common chronic diseases might not always be included in biosurveillance systems.

2.2.5 Data Standards for Biosurveillance

As healthcare data becomes increasingly available by electronic means, the sheer quantity of data available for use has increased dramatically. The speed with which it is necessary to analyze data for biosurveillance requires data to be interoperable. Interoperability in this context essentially means that data from all sources must be structured, coded or represented in the same way so that data can be quickly aggregated and analyzed across all originating data sources.

While the use of standard coding systems such as ICD-9 and CPT have been routine in medical billing for many years, the introduction of data standards into transmission of laboratory data for biosurveillance is still a work in progress. Two of the most widely used vocabularies to achieve interoperability of laboratory data are LOINC [32, 33] to identify test result components and SNOMED [34] to identify subsets of laboratory results that are typically reported in text, such as the names of organisms identified on a culture result. HL7 (Health Level 7) is a message transport standard that is also widely used in healthcare, particularly for electronic laboratory reporting [35], including in the delivery of data for biosurveillance more broadly.

With the creation of the Office of the National Coordinator for Health Information Technology (ONC) under the Department of Health and Human Services (HHS) a number of projects have been initiated to develop use-cases and interoperability standards applicable to a broad range of healthcare transactions, including biosurveillance. Biosurveillance was, in fact, one of the first three use-cases addressed for which interoperability specifications (IS) were developed by the ANSI Healthcare Information Technology Standards Panel (HITSP). In addition, the ELINCS specification developed by the California Health Care Foundation and now maintained by HL7, seeks to address similar needs. While LOINC, SNOMED and HL7 are prominently featured in these IS, many other standards are also designated to address other elements of the data set and processes involved in reporting data for biosurveillance.

2.2.6 Data Analysis

The purpose of this section is to describe the types of issues that need to be addressed when analyzing laboratory data for biosurveillance. The precise statistical and analytic tools required will vary based on the specific applications and will not be described in this chapter.

If confirmatory laboratory data is analyzed in the context of complete public health case reports, there are two principle issues to be considered:

1. Whether the testing methodologies used by different laboratories were sufficiently reliable that all cases can be considered to have been comparably confirmed [36]
2. Whether the data provided is sufficiently granular to determine if cases reported at different points in time meet current or analysis-specific case criteria.

There are additional issues to consider when analyzing large data sets comprised primarily of laboratory-based reporting.

If data from more than one laboratory are to be aggregated together, the comparability of the nomenclature, testing methodologies, reporting units and reference ranges used become critically important. At the most basic level, it is often necessary to map local coding systems to a national standard, such as LOINC, before any analysis can begin. Further data cleaning and transformation may be required to assure that all data can be referenced to the same units. For example, if a single result analyte is typically reported in “cells/mL,” but sometimes appears in “cells/mL × 1,000,” a mathematical transformation will be required to render the data comparable.

While not yet fully realized, a great deal of progress has been made on both electronic availability of laboratory data for surveillance and on progress towards standardization. Nonetheless, careful consideration of source data characteristics that could introduce an artifact into any analysis performed is required to obtain the most value from the available data, particularly when gathered and aggregated from many different sources.

Preliminary descriptive data analysis is generally recommended to assess the quality and consistency of the data available. Validation steps can include looking for unexpected gaps or suspicious increases, as well as comparing observed trends to expected trends based on historical or other information about anticipated changes, such as major known changes in therapy or recent news reports about a condition of interest.

2.2.7 Underlying Data Characteristics

In general, analysis of laboratory data for biosurveillance is considered a secondary use of data since the laboratory tests were originally ordered and results generated for purposes of clinical care, not for surveillance. In the case of laboratory data from sentinel providers, some or all of the data may represent a primary use if the tests were ordered exclusively for public health purposes. As a secondary data use, it is critical to understand the nature of the data and the circumstances under which it originated. As mentioned earlier, it is particularly important to understand the nature of the data reported, specifically whether it has been reported as “raw” data, or if any elements of the data have been derived, such as by the use of text parsers, or transformed in any way, such as by mapping codes.

Data collected for a clinical trial is generally gathered under a protocol that specifies exactly what data, how and at what time intervals it will be collected as well as explicit inclusion and exclusion criteria for which patients to collect it on. Most data currently available for biosurveillance is based on routine clinical practice. In many cases there may be consensus group or other guidelines that outline the appropriate testing to perform in specific clinical settings. Unlike researchers following a protocol, however, independent practitioners may choose to order whatever tests they consider clinically appropriate. Patients, in turn, may choose not to have all the tests ordered performed, whether due to financial or other considerations. That means that it is important to identify and address as many variables influencing the data collected as possible.

Patient and provider characteristics, as well as external drivers, can all influence the nature and composition of the data available. Given the wide variability in patient behavior, clinical orientation of providers, and potential external drivers, accounting for this can be a daunting task. With thought and care however, it is often possible to either structure the analysis to address this issue, or to limit conclusions to the level supported by what is known about the data.

Patient Characteristics

Patient behavior is influenced by many variables. When analyzing laboratory data for biosurveillance, it is necessary to try to derive or address as many of these variables as possible.

First and foremost, of course, behavior is influenced by symptomatology. Patients typically visit a clinician when they perceive the presence of a physical ailment or abnormality that does not appear likely to resolve on its own. A patient visit may relate to the onset of a new condition, or to ongoing monitoring of an existing or chronic condition. Not all patients will present at the same time in the course of their illness, and not all patients will comply with providers' recommendations for laboratory testing. Patient behavior is contingent on a host of variables, including:

Severity of symptoms

Presence or absence of other relevant co-morbidity

Level of medical awareness – for example, whether or not a patient is likely to recognize the importance of seeking care for a specific set of symptoms

Access to care, including logistical issues and financial issues – for example, availability of health insurance or flexible clinic hours

Patient demographics, including cultural issues that may influence whether patients in certain communities are more likely to consult a traditional or alternative practitioner before presenting at a mainstream medical facility

Personal attitudes and characteristics.

Little or none of this information is available in data gathered directly from laboratories, which typically receive limited demographic or clinical information from ordering providers. In the case of reference laboratories, such information is particularly minimal, often limited to those data elements required in order to correctly interpret the test results. This means that any conclusions drawn must be carefully interpreted in light of the characteristics of the data.

In some cases, however, it may be possible to derive desired data elements from those that are available. For example, if testing typical of patients with diabetes – e. g., hemoglobin $A1_c$ results – appears on the same requisition as a culture that was positive for a particular micro-organism of interest, it may be reasonable to conclude that the result was on a patient with diabetes. If there is a perceived need to perform subset analysis based on co-morbidity, this may be a reasonable way to create a usable data set. Severity can similarly sometimes be inferred from the precise nature of the tests ordered, such as specific tests for uncommon respiratory disorders versus basic respiratory cultures.

Such classifications can not be considered totally reliable. From a data mining perspective, however, they may be helpful in examining patient subgroups under circumstances where complete clinical information is not available. Such derived data classifications may prove particularly valuable as part of hypothesis-generating activities [37], or in attempting to assess the face validity of a finding, as opposed to drawing definitive conclusions.

Provider Characteristics

Provider behavior is also influenced by a number of variables. For example, it is not unusual to observe different patterns of both orders and results from clinicians in different specialties or different parts of the country. Culture results for an infectious disease specialist, for example, may reveal a higher concentration of unusual pathogens than those from a general internist as a result of selection bias in the type of patient seen.

If a particular data set includes a balanced representation of specialists and generalists, the net effect may be that the analysis can be interpreted to represent the population as a whole. If possible, it can be very useful for some analyses to attempt to differentiate between tests ordered by different types of providers. Whether or not this is possible will often depend on the granularity of the provider data available. If the representation of provider type is skewed or simply unknown, however, the results could be confusing and even open to misinterpretation.

External Drivers

Providers and patients alike are influenced by trends in health education and awareness, mass media reporting of specific health events, changes in healthcare coverage and changing geographic population distributions. If laboratory results are trended over time, it is therefore critical to normalize the changes in results of concern as a function of the overall testing performed. In cases where only positive results are available, for example, it can be very difficult to understand if an observed increase in case reporting represents a true increase in the incidence or prevalence rate, as opposed to a better case detection rate as a result of increased testing. If the testing rate increases dramatically enough, such as might occur during a transition from symptom-oriented testing to screening, an absolute increase in the number of positive cases reported could even be associated with a decline in the true incidence or prevalence rate.

Transient rises in specific testing rates have occurred in the context of a perceived outbreak [11], and more sustained rises have occurred in response to major media stories. Although not reported in the scientific literature, rates of testing for hepatitis C at a large national laboratory were observed by the author to have increased dramatically after extensive broadcast and print news coverage of former surgeon general C. Everett Koop's announcements about the dangers and consequences of hepatitis C infection.

The approval of a new treatment for a disease can also lead to increased testing activity as both patients and providers may see a new impetus to confirm, rule-out, or monitor patients with a specific diagnosis. Again, though not reported in the literature, dramatic increases in HIV viral load testing volume at a large national laboratory were observed by the author following the introduction of protease inhibitors for treatment of HIV infection.

In some cases, testing volume and positivity for a particular test can appear to increase – or decrease – as an artifact of an overall change in testing rates in a particular area. This can be the result of a procedural change – such as a transition from paper-based to electronic laboratory reporting [24, 25], or the result of a local or global economic change that affects the ability of patients to seek healthcare at the same rate as was previously true. In some cases, depending on the data source, a particular data set may include a greater or lesser percentage of the total testing data in a particular area.

Periodicity and other re-occurring temporal variations, e.g., day of week effects, seasonal variations and major holidays, can also have an impact on apparent fluctuations in data. It is necessary to understand the normal seasonal variations over the course of the year in order to assess the importance of a detected change in any given year, particularly with respect to understanding the difference between a possible outbreak and the start of a predictable seasonal increase. Some temporal variations can be adjusted for statistically, by aggregating data at different units of time and by comparing data from earlier years. Weekly data counts, for example, can be used to address day-of-the-week variability, permitting comparisons over time with less distracting noise. In the case of seasonal variability, it is advisable to look at multiple years of data to understand how consistent the variability has been over time and whether to assess the relevance of any observed increase.

Performing data validation and required data transformations requires both a good grounding in current trends related to a specific disease to understand the context of any observed changes, as well as a strong statistical foundation in pattern recognition and anomaly detection to better identify suspicious anomalies that may be the source of an unrecognized artifact or bias in the data.

2.3 Relevant Experience and Case Studies

There is a growing body of experience available on both the benefits and challenges of using laboratory data for biosurveillance, as well as projections for the anticipated impact of electronic laboratory reporting on surveillance for notifiable conditions.

2.3.1 The Emergence of West Nile Virus in New York City

Jernigan et al. retrospectively evaluated data from a large national laboratory and identified a 30-fold spike in orders for St. Louis Encephalitis (SLE) over a 2-week period in New York City during the emergence of West Nile Virus (WNV) there in 1999 [11]. Although the spike occurred after the New York City Department of Health and Mental Hygiene (NYC DOHMH) had announced what appeared to be cases of SLE, the dramatic increase in testing in such a short time period produced a spike that would have easily been detectable using anomaly detection algorithms. The fact that all of the tests produced negative results meant that none of these tests would have been reported to the NYC DOHMH or any other public health agency, even if some of the patients had WNV. This investigation supported the premise that laboratory orders, as well as results, could be useful in detecting outbreaks, particularly in the case of emerging infectious disease where no confirmed laboratory test results would be available.

2.3.2 Lyme Disease in New Jersey

Lyme disease is most prevalent in the northeast and New Jersey has one of the highest rates in the United States. In 2002, the state had begun to include electronic laboratory reporting (ELR) for notifiable diseases. An evaluation of the state's Lyme disease surveillance system in 2006 [29] revealed a nearly fivefold increase in reported cases from 2001 to 2004, although confirmed reports only increased 18%. ELR represented 51% of the cases reported from 2001 to 2006, but only 29% were confirmed upon investigation. Differences were noted in the data obtained from ELR and non-ELR reports, with proportionally more non-ELR cases confirmed in high prevalence areas and during peak LD transmission season. The increased burden of investigations created greater demand on limited public health department resources. Ultimately a decision was made to classify ELR-reported cases as "suspected cases" unless there is also a concurrent report from a healthcare provider.

2.3.3 Hepatitis in New York City

The experiences of the NYC DOHMH helped to identify areas in which ELR could be particularly valuable, such as high-volume or time-sensitive diseases [28]. In the case of hepatitis A, for which prompt administration of postexposure prophylaxis (PEP) to contacts is important, ELR improved reporting time considerably with a median decrease of 17 days for laboratories certified to report electronically, as well as a 35% increase in reports received leading to a fourfold increase in PEP administration [38]. It was noted that ELR did not improve surveillance for all diseases equally and that many of the previously mentioned challenges warranted a higher level of technical skill and experienced surveillance staff to obtain the best results [28].

2.3.4 Projections in Florida

Based on the timeliness of reporting for four notifiable diseases, and based on the expectation that ELR could reduce reporting time from completion of laboratory testing to 1 day, a study in Florida estimated that the total time from onset of symptoms to reporting to the county health department could be cut from 12 days to 7 days for salmonellosis and from 10 days to 6 days for shigellosis, but would produce no change for meningococcal disease [39]. In the case of meningococcal disease, clinicians were noted to already place great emphasis on timely reporting, while historical records showed lower rates of timely reporting for the other conditions examined. The findings of the analysis echo the experience of the NYC DOHMH that the benefits of ELR are likely to vary by disease.

3. CONCLUSIONS AND DISCUSSION

Clinical laboratory data is a key component of traditional public health case reporting as well as syndromic surveillance and biosurveillance. Laboratory result data can be precise and highly specific when used in case confirmation. Laboratory order data may be less specific, but it can be more sensitive in detecting emerging infectious diseases and bioterrorism.

The implementation of electronic laboratory reporting systems has improved timeliness and completeness of laboratory-based reporting, although the benefits vary by disease. Continued improvements in the technology used for such systems should continue to reduce problems historically experienced with these systems, such as false positives and duplicate records.

There are variations in the type of data available from different laboratory segments that may influence the specific utility of the data in various situations. Furthermore, data gathered directly from laboratories with limited clinical context must be interpreted with care when attempting to extrapolate to conclusions on a population basis.

Clinical laboratory data alone will never provide all of the information required by public health agencies to detect and characterize events of public health concern, but it is – and will continue to be – a core component of such efforts.

QUESTIONS FOR DISCUSSION

1. Identify some of the potential advantages and disadvantages of each different type of data mentioned as having been used for biosurveillance. Can you think of other types of data that could be useful today? Are there other types of data that might become more valuable in the future?
2. What are some of the challenges associated with analyses based on secondary use of clinical data, as opposed to the type of data that is normally collected for research based on a protocol?
3. Describe at least one public health situation where data from each different laboratory market segment would be particularly valuable. Explain the reasoning behind each choice and consider what contribution the other segments could make in the same situation.

4. Discuss the differences in data requirements when attempting to detect an outbreak or a trend.
5. Explain the role of standards in using laboratory data for biosurveillance.
6. Suggest possible reasons that laboratory or other components of public health case definitions might change over time.

REFERENCES

1. Greenberg SJ. The "Dreadful Visitation": public health and public awareness in seventeenth-century London. *Bull Med Libr Assoc* 1997 Oct;85(4):391–401.
2. Collen MF. *A History of Medical Informatics in the United States.* Indianapolis: American Medical Informatics Association, 1995.
3. Stroup NE, Zack MM, Wharton M. Sources of routinely collected data for surveillance. In: Teutsch SM, Churchill RE, eds., *Principles and Practice of Public Health Surveillance.* New York: Oxford University Press, 1994:31–82.
4. McNabb SJN, Jajosky RA, et al. Summary of notifiable diseases – United States, 2006. *MMWR Morb Mortal Wkly Rep* 2008 Mar;55(53):1–94.
5. Teutsch SM. Considerations in planning a surveillance system. In: Teutsch SM, Churchill RE, eds., *Principles and Practice of Public Health Surveillance.* New York: Oxford University Press, 1994:18–28.
6. Sosin DM. Draft framework for evaluating syndromic surveillance systems. *J Urban Health* 2003 Jun;80(2 Suppl 1):i8–13.
7. Henning KJ. What is syndromic surveillance? *MMWR Morb Mortal Wkly Rep* 2004 Sep:53(Suppl):7–11.
8. Heffernan R, Mostashari F, et al. New York City syndromic surveillance systems. *MMWR Morb Mortal Wkly Rep* 2004 Sep;53(Suppl):25–7.
9. Das D, Metzger K, Heffernan R, Balter S, Weiss D, Mostashari F. Monitoring over-the-counter medication sales for early detection of disease outbreaks. *MMWR Morb Mortal Wkly Rep* 2005 Aug;54(Suppl):41–6.
10. Ma H, Rolka H, Mandl K, Buckeridge A, Fleischauer A, Pavlin J. Implementation of laboratory order data in BioSense early event detection and situation awareness system. *MMWR Morb Mortal Wkly Rep* 2005 Aug;54(Suppl):27–30.
11. Jernigan DB, Koski E, Shoemake HA, Mallon RP, Pinner RW. Evaluation of Laboratory Test Orders and Results in a National Laboratory Data Repository: Implications for Infectious Diseases Surveillance. *International Conference on Emerging Infectious Diseases 2000 Program and Abstracts Book*; 2000;58/88:139.
12. Patnaik JL, Juliusson L, Vogt RL. Environmental predictors of human West Nile virus infections, Colorado. *Emerg Infect Dis* 2007 Nov;13(11):1788–90.
13. Wharton M, Chorba TL, Vogt RL, Morse DL, Buehler JW. Case Definitions for Public Health Surveillance *MMWR* 1990 Oct;39(RR-13):1–43.
14. Case Definitions for Infectious Conditions Under Public Health Surveillance *MMWR* 1997 May;46(RR-10):1–55.
15. Andrews JM, Quinby GE, Langmuir AD. Malaria eradication in the United States. *Am J Public Health* 1950;40:1405–11.
16. Langmuir AD. The surveillance of communicable diseases of national importance. *N Engl J Med* 1963;268:182–92.

17. Backer HD, Bissell SR, Vugia DJ. Disease reporting from an automated laboratory-based reporting system to a state health department via local county health departments. *Public Health Rep* 2001 May–Jun;116:257–65.
18. Kotlarz VR. Tracing our roots: origins of clinical laboratory science. *Clin Lab Sci* 1998 Jan/Feb;11(1):5–7.
19. Kotlarz VR. Tracing our roots: the beginnings of a profession. *Clin Lab Sci* 1998 May/Jun;11(3):161–6.
20. Delwiche FA. Mapping the literature of clinical laboratory science. *J Med Libr Assoc* 2003 July;91(13):303–10.
21. Dick J, Parrish NM. Microbiology tools for the epidemiologist. In: Nelson KE, Williams CM, Graham NMH, eds., *Infectious Disease Epidemiology: Theory and Practice*. Sudbury, MA: Jones and Bartlett Publishers, 2005:171–204.
22. Ford EW, Menachemi N, Phillips MT. Predicting the adoption of electronic health records by physicians: when will health care be paperless? *JAMIA* 2006 Jan–Feb; 13(1):106–12. Epub 2005 Oct 12.
23. Centers for Disease Control and Prevention (CDC). Automated detection and reporting of notifiable diseases using electronic medical records versus passive surveillance-massachusetts, June 2006-July2007. *MMWR Morb Mortal Wkly Rep* 2008 Apr;57(14):373–6.
24. Overhage M, Grannis S, McDonald CJ. A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. *Am J Public Health* 2008 Feb;98(2):344–50. Epub 2008 Jan 2.
25. Effler P, Ching-Lee M, Bogard A, Ieong M, Nekomoto T, Jernigan D. Statewide system of electronic notifiable disease reporting from clinical laboratories: comparing automated reporting with conventional methods. *JAMA* 1999;282:1845–50.
26. M'ikanatha NM, Southwell B, Lautenbach E. Automated laboratory reporting of infectious diseases in a climate of bioterrorism. *Emerg Infect Dis* 2003 Sep;9(9):1053–7.
27. Panackal AA, M'ikanatha NM, Tsui FC, et al. Automatic electronic laboratory-based reporting of notifiable infectious diseases at a large health system. *Emerg Infect Dis* 2002 Jul;8(7):685–91.
28. Nguyen TQ, Thorpe L, Makki HA, Mostashari F. Benefits and barriers to electronic laboratory results reporting for notifiable diseases: The New York City Department of Health and Mental Hygiene experience. *Am J Public Health* 2007 April;97(Suppl 1):S142–45.
29. Centers for Disease Control and Prevention (CDC). Effect of electronic laboratory reporting on the burden of Lyme disease surveillance – New Jersey, 2001–2006. *MMWR Morb Mortal Wkly Rep* 2008 Jan;57(2):42–5.
30. Hodge JG Jr, Brown EF, O'Connell JP. The HIPAA privacy rule and bioterrorism planning, prevention, and response. *Biosecur Bioterror* 2004;2(2):73–80.
31. Winter A, Funkat G, Haeber A, Mauz-Koerholz C, Pommerening K, Smers S, Stausberg J. Integrated information systems for translational medicine. *Methods Inf Med* 2007;46(5):601–7.
32. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, Fiers T, Charles L, Griffin B, Stalling F, Tullis A, Hutchins K, Baenziger J. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clin Chem* 1996 Jan;42(1):81–90.
33. Baorto DM, Cimino JJ, Parvin CA, Kahn MG. Combining laboratory data sets from multiple institutions using the logical observation identifier names and codes (LOINC). *Int J Med Inform* 1998 Jul;51(1):29–37.
34. Rothwell D, Cote R, et al. Developing a standard data structure for medical language – the SNOMED proposal. *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, Washington, DC: McGraw-Hill, 1995:695–9.

35. Zeng D, Chen H, Lynch C, Eidson M, Gotham I. Infectious disease informatics and outbreak detection. In: Chen H, Fuller SS, Friedman C, Hersh W, eds., *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. New York: Springer Science + Business Media, Inc., 2005:369–70.
36. Ravel R. Various factors affecting laboratory test interpretation. In: Ravel R, *Clinical Laboratory Medicine*, *6th Edition*. St. Louis: Mosby-Year Book, Inc., 1995:1–8.
37. Bean NH, Martin SM. Implementing a network for electronic surveillance reporting from public health reference laboratories: an international perspective. *Emerg Infect Dis* 2001 Oct;7(5):773–9.
38. Moore KM, Reddy V, Kapell D, Balter S. Impact of electronic laboratory reporting on hepatitis A surveillance in New York City. *J Public Health Manag Pract* 2008 Sept–Oct;14(5):437–41.
39. Centers for Disease Control and Prevention (CDC). Potential effects of electronic laboratory reporting on improving timeliness of infectious disease notification – Florida, 2002–2006. *MMWR Morb Mortal Wkly Rep* 2008 Dec;57(49):1325–8.

SUGGESTED READING

1. Collen MF. *A History of Medical Informatics in the United States*. Indianapolis: American Medical Informatics Association, 1995.
2. Nelson KE, Williams CM, Graham NMH, eds., *Infectious Disease Epidemiology: Theory and Practice*. Sudbury, MA: Jones and Bartlett Publishers, 2005: Particularly Chapters 1, 7, 8.
3. O'Carroll PW, Yasnoff WA, Ward ME, Ripp LA, Martin EL, eds., *Public Health Informatics and Information Systems*. New York: Springer-Verlag, 2003.
4. Teutsch SM, Churchill RE, eds., *Principles and Practice of Public Health Surveillance*. New York: Oxford University Press, 1994.

ONLINE RESOURCES

APHL (Association of Public Health Laboratories) http://www.aphl.org/Pages/default.aspx
ASCP (American Society for Clinical Pathology) http://www.ascp.org/
Centers for Disease Control and Prevention http://www.cdc.gov/
HIPAA http://www.hipaa.org/
HL7 http://www.hl7.org/
International Society for Disease Surveillance – syndromic surveillance http://syndromic.org
Lab Tests Online http://www.labtestsonline.org/
LOINC http://www.regenstrief.org/
MedLine Plus – Laboratory Tests http://www.nlm.nih.gov/medlineplus/laboratorytests.html
Morbidity and Mortality Weekly Report (MMWR) http://www.cdc.gov/mmwr/
National Library of Medicine http://www.nlm.nih.gov/
NREVSS – National Respiratory and Enteric Viral Surveillance System http://www.cdc.gov/surveillance/nrevss/
Public Health Case Definitions http://www.cdc.gov/ncphi/disss/nndss/casedef/index.htm
SNOMED http://www.ihtsdo.org/

Chapter 5

BIOSURVEILLANCE BASED ON TEST ORDERS FROM VETERINARY DIAGNOSTIC LABS

LOREN SHAFFER*

CHAPTER OVERVIEW

The threat of disease epidemics resulting from zoonotic pathogens is a valid concern for public health authorities and others involved with healthcare of people and animals. A significant number of epidemics over the past two decades have resulted from infection by a zoonotic pathogen. Detection of these outbreaks has typically relied upon the identification of human cases in spite of the fact that humans are a primary reservoir for a small percentage of zoonotic pathogens. Detection of an outbreak of zoonotic disease in animals before it has reached outbreak status in humans could provide the opportunity for earlier interventions that greatly reduce human morbidity and mortality. This chapter discusses test orders made to veterinary diagnostic laboratories as a potential source of data for biosurveillance efforts and provides sample methods for evaluating these data. Examination of these data has provided promise of their utility for such surveillance systems and demonstrated the ability to detect outbreaks of certain diseases earlier than traditional reporting methods.

Keywords: Animal diseases; Bioterrorism; Communicable diseases; Emerging; Disease outbreaks; Epidemiology; Population surveillance; Sentinel surveillance; Zoonoses

* *The Ohio State University, College of Veterinary Medicine, 1920 Coffey Road, Columbus, OH 43210, USA, loreneshaffer@gmail.com*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_5,

1. INTRODUCTION

It should not be surprising that a 2003 report published by the Institute of Medicine cites animal contact as one of the major contributing factors to emerging infectious disease. Over 75% of the outbreaks attributed to emerging infectious disease that occurred in the last two decades of the twentieth century were the result of zoonotic pathogens (Taylor et al., 2001). Indeed, pathogens with more than one host species are two to four times more likely to result in an emerging disease than their single host counterparts (Cleaveland et al., 2001; Woolhouse, 2002). Contacts with animals in petting zoos, farms, and other animal exhibits have often been associated with outbreaks of zoonotic disease in humans (Bender and Shulman, 2004). However, the entire extent that animals might provide us a means for early detection of outbreaks of zoonotic disease remains to be determined.

1.1 Wildlife as Sentinels of Disease

Certain animals have already demonstrated some of their potential value as sentinels of infectious disease in humans. Crows are a good example. In New York State, a web-based dead bird surveillance project identified an increase in the density of dead crow sightings and West Nile Virus (WNV) positive dead birds 2 weeks prior to the first reported human case of WNV (Eidson et al., 2001a). Additional investigations support that avian morbidity and mortality surveillance provides information that is helpful in predicting WNV onset in humans (Guptill et al., 2003), sometimes as much as 3 months in advance (Eidson et al., 2001b). Other examples of animals serving as sentinels for infectious disease in humans include pheasants for Eastern equine and St. Louis encephalitis (Morris et al., 1994), and white-tailed deer for Lyme disease (Gill et al., 1994).

1.2 Pets as Sentinel Indicators of Disease

Domestic species, including pets, should not be overlooked as both potential sources of and sentinels for emerging infectious diseases. A study conducted in the U.S. from 1993 to 1995 found 13.1% of owned cats were infected with *C. parvum, Giardia, Toxocara cati, Salmonella enterica,* or *Campylobacter jejuni* (Hill et al., 2000). Tauni and Österlund (2000) also associated cats as the intermediary host of *Salmonella thyphimurium* from wild birds to humans. Recently, researchers have speculated that cats might serve a similar role as an intermediary host in the transmission of avian influenza viruses from wildlife to humans (Tansey, 2006). Domestic poultry (Kaye and Pringle, 2005) along with dogs (Crawford et al., 2005) are candidates for such roles as well. Although the path of transmission remains debatable, all of these species are potential hosts of at least certain influenza viruses, making the development of a multi-species surveillance system both a benefit and a challenge when preparing for a potential pandemic of influenza.

Pets have already served as sentinel indicators for certain diseases associated with exposure to pesticides and asbestos (Backer et al., 2001). They often share much of the same environment they live in with their owners. When infected at the same time, pets will develop signs of some disease before humans (Babin et al., 2003). Hence, pets might fill a role as sentinels for the earlier detection of outbreaks of infectious disease, including those resulting from a purposeful release of a disease-causing agent. Pathogens could also first infect pets, making them a source of infection for their owners and others they contact. Identifying the presence of disease early in pets for either scenario could provide an alert that summons the attention of health officials and provides for earlier response to an outbreak, whether occurring naturally or intentionally.

1.3 "One Medicine"

Inspired by his observations of Sudanese Dinka pastoralist healers treating both animals and humans for what were often common ailments, Calvin Schwabe (1984) coined the term "One Medicine" to describe his vision of veterinary and human medicine working together to address disease and improve upon public health. Effective surveillance of zoonotic pathogens and control of emerging diseases that they may cause is felt by some to be outside the scope of traditional medicine, requiring integration across human and animal populations (Woolhouse, 2002; National Research Council, 2005).

However, such a holistic approach is lacking in contemporary veterinary and medical communities. A 2000 report from the Chemical and Biological Arms Control Institute refers to this absence of strong links between public health and the veterinary community as "an important disconnect." Another report, published by the National Research Council (2005), recognizes the need for better integration of animal and public health surveillance in order to improve upon the ability to rapidly detect outbreaks caused by zoonotic pathogens. The United Nations Food and Agriculture Organization, the World Organization for Animal Health (OIE), and the World Health Organization consider the weaknesses in disease detection to be a contributor to the spread of diseases of animal origin (Center for Infectious Disease Research and Policy, 2006).

2. SURVEILLANCE FOR OUTBREAKS OF ZOONOTIC DISEASE

The earlier that an outbreak involving zoonotic disease is detected can mean earlier action to reduce the threat or impact in terms of morbidity, mortality, and economic loss, to people and animals. Enhanced disease surveillance in animals, coordinated in a "One Medicine" approach as envisioned by Calvin Schwabe, may lead to quicker response by both veterinary and public health authorities that greatly minimizes the impact of an outbreak of zoonotic disease. Integrated efforts could also serve to identify cases in human and animal populations that might otherwise go unnoticed. Such efforts would aid in the concurrent treatment of cases in all affected species and lead to a more thorough eradication of the pathogen implicated in the outbreak and reduce the chances of reinfection resulting from an unchecked reservoir.

Outbreaks of zoonotic disease typically rely on the identification of human cases (Childs et al., 1998) although humans are the primary reservoir for a mere 3% of zoonotic pathogens (Taylor et al., 2001). The potential economic implications of an outbreak rather than the zoonotic potential of the pathogen have been the driving forces behind surveillance for disease in agricultural animals (Conner, 2005). Most outbreaks from zoonotic pathogens that occur in animals do not have the same economic implications or direct impact on humans as other diseases and therefore are sometimes not considered to have the same degree of importance (Wurtz and Popovich, 2002). Hence, most disease surveillance in animals targets agricultural species and tends to be very disease specific. The disease specificity of such surveillance greatly reduces the ability to detect outbreaks of other than target diseases. The following list of surveillance systems that utilize animal-based data often provide for examples of this trend.

2.1 National Animal Health Reporting System

The National Surveillance Unit (NSU) of the United States Department of Agriculture (USDA) has identified surveillance for emerging diseases as a priority for the National Animal Health Reporting System (NAHRS) (http://www.aphis.usda.gov/vs/ceah/ncahs/nahrs). The NAHRS receives data from state veterinarians in participating states on the presence of confirmed clinical disease. Diseases are limited to those that are reportable to the OIE in specific commercial species in the United States including cattle, sheep, goats, equine, swine, poultry, and food fish. Only six of the diseases on the OIE list also appear on the list of United States nationally notifiable infectious diseases for humans, of which 38 out of 58 (65.5%) are zoonotic.

Wurtz and Popovich (2002) criticize the NAHRS as a passive, voluntary system without quality control, verification, or feedback. Passive collection of data, such as that utilized by the NAHRS, is limited by the inconsistency in collection for different diseases and among different states. Additionally, the NAHRS provides little benefit as far as timeliness since it compiles a national summary report from data on a monthly basis.

2.2 National Animal Health Laboratory Network

The National Animal Health Laboratory Network (NAHLN) (http://www.nahln.vs/) is another USDA project intended to provide for earlier detection and tracing of outbreaks. The USDA National Veterinary Services Laboratories promised the NAHLN, as part of a strategy to coordinate the Federal, State, and university laboratory resources, to be "a cornerstone of animal health surveillance that will electronically connect surveillance data systems to laboratory diagnostics." Initiated in 2002, the NAHLN included only 12 laboratories as of February 2006 (Mark, 2006). Focus is again disease specific and limited to African swine fever, avian influenza, classical swine fever, contagious bovine pleuropneumonia, exotic Newcastle disease, foot and mouth disease, lumpy skin disease, and rinderpest in agricultural animals. Some have considered the NAHLN to lack the capacity to deal with massive or multiple outbreaks in the United States (Kearney, 2005).

2.3 Veterinary Services Electronic Surveillance Project

The Center for Emerging Issues (CEI) within the United States Department of Agriculture (USDA) maintains the Veterinary Services Electronic Surveillance project. At the heart of this system is Pathfinder, a data mining tool that is used to search the electronic open records of the internet. The CEI users determine words and word combinations of interest that Pathfinder uses to complete its search. Since searched sources include all those of the WorldWide Web, poor translation of foreign sources tends to be a limitation (Johnson, 2004). Another limitation is the sometimes inaccurate reporting that occurs in the media where terminology may appear out of proper context.

CEI uses data from this system to contribute to another syndromic surveillance system located within the USDA, the Offshore Pest Information System (OPIS). CEI designed the OPIS system to enhance information sharing between the USDA's Veterinary Services and International Services. OPIS combines the CEI data from Pathfinder with field reports from International Services of the USDA to generate weekly reports for USDA users.

2.4 Rapid Syndrome Validation Project for Animals

The Rapid Syndrome Validation Project for Animals (RSVP-A) was an initiative of the Kansas State University system to have attending veterinarians determine specific syndromes from patient signs and upload them via internet connection or Palm® device. Syndrome categories were non-neonatal diarrhea, neurologic dysfunction or inability to rise, abortion or birth defect, unexpected death, erosive or ulcerative lesions of the skin, mucosa, or coronet, and feed refusal or weight loss without clear explanation. The system included only cattle. Reporting limitations occurred because of the provider-dependent design and the frequent unavailability of service to support the wireless devices (DeGroot, 2005).

2.5 Other Manual Entry Systems

The Petsavers Companion Animal Disease Surveillance system was an initiative of the British Small Animal Veterinary Association. Investigators collected data from 15 small animal veterinary practices in the form of regular surveys requiring written responses to questions about patients treated within a reporting period of up to 4 days. Some questions in the survey had non-response rates of 30%. The conclusion of the investigators was that a more robust technique for collection and preparation of data that is less time consuming and more accurate is required (Robotham and Green, 2004).

Michigan State University conducted a similar project that involved dairy farmers providing daily records of animal events and veterinarians recording diagnoses and treatments. While burdensome because of the manual reporting, weekly and monthly reports from the data provided back to the participants were determined to be useful in managing the health of herds. The usefulness of the reports served as a sufficient incentive for continued participation in the surveillance system (Bartlett et al., 1986).

3. IMPROVING OUTBREAK DETECTION

The disease surveillance programs that are in place for agricultural animals do not exist for pets. No regulatory agency is exclusively charged with the collection of disease reports in pets, so it can become confusing for veterinarians treating these animals when deciding what needs to be reported and to whom. In the absence of disease reports for pets, it can become challenging for public health officials to find any data about co-morbidity in these animals. Such data could provide information that would aid them in identifying and controlling outbreaks of zoonoses (Table 5-1).

Table 5-1. Select animal-based biosurveillance programs

System	Agency	Species	System Design	Pros & Cons
VS Electronic Surveillance (Pathfinder)	Center for Emerging Issues, USDA	Multiple	searches internet open records for specified word and word combinations	Pro - extensive number of sources included Con - subject to translation error and inaccurate reporting
Offshore Pest Information System	Veterinary and International Services, USDA	Multiple	field reports and results from Pathfinder to create weekly reports	Pro - enables increased communications between areas Con - lack of timeliness and Pathfinder limitations
Rapid Syndrome Validation Project for Animals	Kansas State University	Cattle	veterinarians determine syndrome and report via internet or wireless service at or near time of service	Pro - presumed increase in specificity from provider-based reports Con - creates additional burden on providers, unavailability of utility service, limited to include a single species
Petsavers Companion Animal Disease Surveillance	British Small Animal Veterinary Association	Pets	veterinarians submit written survey of cases treated in practice during multi-day period	Pro - presumed increase in specificity from provider-based reports Con - creates additional burden on providers and poor reporter compliance
Computerized Dairy Herd Health Database	Michigan State University	Dairy cattle	farmers report daily animal events and veterinarians report diagnoses and treatments	Pro - presumed increase in specificity from provider participants, provides feedback to reporters Con - creates additional burden on providers and farmers, limited to single species

3.1 Syndromic Surveillance

Syndromic surveillance is the systematic and ongoing collection, analysis, and interpretation of data that precede diagnosis and can signal a sufficient probability of an outbreak to warrant public health investigation (Sosin, 2003). Syndromic surveillance systems developed for detection of outbreaks in humans have used emergency department chief complaints (Begier et al., 2003; Tsui et al., 2003), electronic medical records (Lazarus, 2001), sales of over-the-counter medications (Das et al., 2005; Wagner et al., 2003), contents of grocery baskets (Feinberg and Shmueli, 2005), medical laboratory orders (Ma, 2005), and diagnostic codes (Bradley et al., 2005). By utilizing predominantly pre-diagnostic data, these systems have the potential for greatly improving the timeliness of outbreak detection. The application of syndromic surveillance methods have been recommended to detect novel and emerging zoonoses including those associated with potential bioterrorist action (Kruse et al., 2004). However, sources of veterinary-based data for such methods are not as immediately obvious compared to those used for human-based surveillance.

3.1.1 Preferred Data

The selection of data sources for syndromic surveillance are initially influenced by evidence or the belief of the system developers that the source can provide an early signal for the disease/s or condition/s of interest (Zeng and Wagner, 2002). It may be difficult to measure the true value of non-traditional data sources, especially those not related to healthcare, as outbreaks of the diseases of interest are potentially rare (Johnson et al., 2005). In these cases, system developers must rely on retrospective studies of other diseases with similar presentation or surveys to measure behaviors that could influence the data.

The most valuable data sources for syndromic surveillance are those that are stored electronically, permit robust syndromic grouping, and are available in a timely fashion (Mandl et al., 2004). Systems that require additional data entry and increase workloads are undesirable, especially for large-scale sustained surveillance. Using existing data collected for other purposes, stored electronically, and available for automatic transfer are preferable, as they do not depend on changes in workflow.

3.1.2 Data Criteria

Certain criteria for evaluating data to be used for surveillance systems are recommended.

1. The Centers for Disease Control and Prevention have identified the ability to provide baseline information on incidence trends and geographic distribution as a prerequisite to detecting new or re-emerging infectious disease threats. The baseline becomes especially important to determining when counts are abnormally elevated. Making accurate interpretations from the results of detection analyses is difficult without first establishing what is normal. Baselines help to determine the noise in the data and provide for establishing expected values required in the analyses. Such indices are important to validate the

predictive models used by detection systems to determine abnormal patterns of distribution or counts.

2. Representativeness describes how well records in the system describe the population and indicates the potential of accurately determining the distribution of cases by time and place. The presence of a species may be a more important measure of representativeness for early outbreak surveillance. If the goal of the system is to detect emerging diseases in pets then it follows that the data need to include information for companion animals. Changes in the relative representation of these species reflected in the records might provide indication of an outbreak that is limited to only certain species.
3. The availability of data reflects the potential gain in terms of timeliness of detection, the time from the disease event to the time the event is discovered (Sonesson and Bock, 2003). Timeliness has become a major objective of surveillance systems used to detect outbreaks of infectious disease. This potential gain establishes the value of data for earlier detection of disease outbreaks compared to traditional disease reporting and detection systems.

3.2 Veterinary Diagnostic Laboratories

Veterinary diagnostic laboratories (VDLs) are a potential source of data for outbreak detection and considered important tools for surveillance in animals (Conner, 2005). For many outbreaks, detection may require central aggregation and analysis of data from many sources (Dato et al., 2004). Commercial laboratories offer the advantage of providing data for patients seen by many different providers over a sizeable geographic area from a single source. This is no less true for commercial veterinary laboratories where the majority of approximately 22,500 private veterinary practices in the United States submit patient samples for analysis (Glickman et al., 2006).

3.2.1 Case Evidence to Support Using Data

The author completed a retrospective study of microbiology test orders submitted to a commercial veterinary diagnostic laboratory for specimens originating from veterinary clinics in Central Ohio during 2003 (Shaffer et al., 2008). As the VDL received samples, accession personnel created an electronic record that included data useful for outbreak surveillance (e.g., ZIP code of clinic, species of animal, date of specimen collection). Serfling regression methods used historic data to develop models of expected counts (i.e., baselines) of genera-specific pathogens for each week (see Figure 5-1). Comparison of these pathogen trends in animals with the incidence of disease reports in humans from the same area sometimes revealed significant increases in the same pathogen infecting humans and animal species at the same time such as *Escherichia coli* (see Figure 5-2). In this instance, only ten of the total *E. coli* (O157:H7) cases in humans with onset of illness reported between weeks 40 and 42 were attributed to contaminated food. Eight of the total cases in humans not associated with food resulted from pathogen subgroups that were indistinguishable by pulsed-field gel electrophoresis (PFGE).

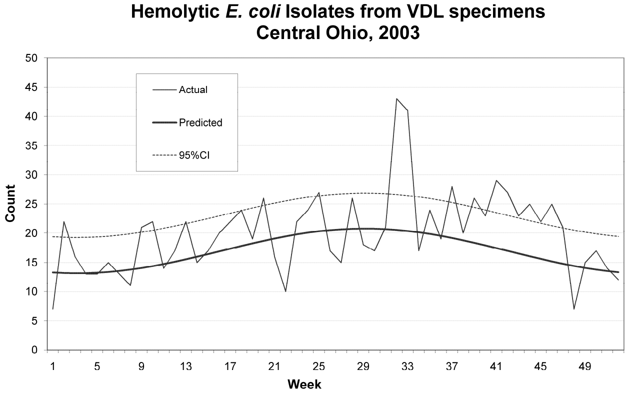


Figure 5-1. Counts of hemolytic *E. coli* isolates by date of submission to VDL from specimens originating from veterinary clinics in Central Ohio during 2003 with Serfling model prediction.

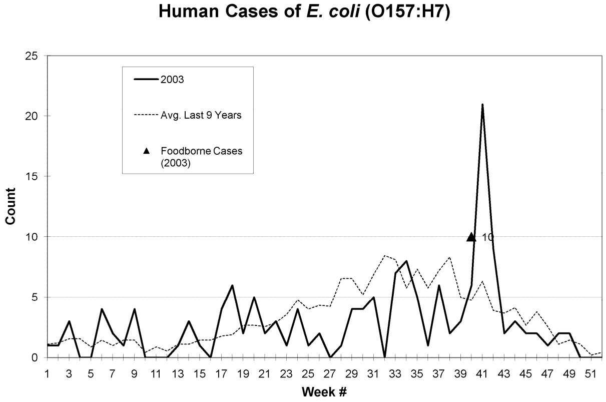


Figure 5-2. Reported cases of *E. coli* (O157:H7) in humans living in Central Ohio by onset of illness with average counts over previous 9 years. *Triangle marker* indicates the number of cases associated with contaminated food during weeks 40–42.

3.2.2 Determining Animal Representation

How well companion animals might serve as a proxy for humans is a consideration for surveillance system developers. The extent that human and animals share disease-causing organisms is only one aspect of this consideration. Knowing how well animals represent humans in the same area is another aspect that provides us with information helpful in estimating prevalence, exposure, and system sensitivity. The author explored this by estimating the number of households included in surveillance by a system that would use data from a veterinary diagnostics laboratory.

The author examined a dataset provided from a commercial veterinary diagnostics laboratory that contained test orders for specimens originating from patients treated at clinics located in Ohio received between April 1, 2005 and March 31, 2006 (Shaffer, 2007). Information in the dataset included an accession number that identified specimen submissions from a single animal. From this dataset, the author determined the number of specimens originating from dogs, cats, horses, and pet birds that veterinarians submitted to the laboratory for testing during the 1-year study period by counting the unique accession numbers grouped by genus (see Figure 5-3).

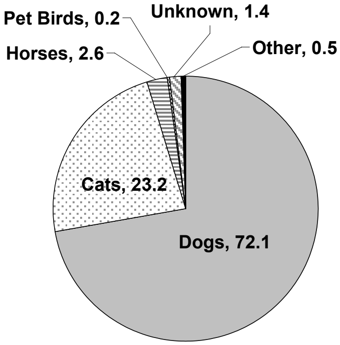


Figure 5-3. Percentage of laboratory specimen accessions by genera received by VDL from clinics located in Ohio, April 1, 2005–March 31, 2006.

3.2.3 Estimating Human Representation

During the first week of January 2002, the American Veterinary Medical Association mailed the Household Pet Survey to 80,000 randomly selected households in the United States. This survey asked a number of questions regarding animals owned anytime during 2001 including veterinary care, number of animals owned, and household demographics. The AVMA used the responses to estimate the proportion of veterinary visits that resulted in laboratory submissions, the average number of animals by species per household, and the rate of annual veterinary visits per animal by genera. Using these figures from the AVMA survey, the author was able to calculate the number of households represented in the VDL records (see Table 5-2).

The number of animal-owning households was compared to the estimated total number of Ohio households from the 2002 United States Census (http://www.factfinder.census.gov) to estimate the percentage of households represented by these data. These calculations were repeated for each group by adding and subtracting 10% to each estimate generated by the AVMA survey (number of animals per household, number of annual veterinary visits, and percentage of visits that involved a laboratory submission) to consider how much that degree of simultaneous over or under estimation might change the final estimate (see Figure 5-4). Visual examination of maps depicting the number of accessions sent to the laboratory during this time reveals how they approximate the relative population of people living in each area (see Figure 5-5).

Table 5-2. Estimating the number of Ohio households represented in IDEXX dataset (April 1, 2005–March 31, 2006) from the number of laboratory specimen accessions.

		A	B	C	D	E	F	G
		No. of Lab Accessions	% of Visits Resulting in Lab Testing[a]	Est. No. of Vet Visits (A/B)	No. of Annual Vet Visits/ Animal[a]	Est. No. of Animals/ House-hold	No. of Animals /House-hold[a]	Est. No. of House-holds (E/F)
Dogs	+10% Error	150,309	20.5	733,215	2.1	349,150	1.7	205,383
	−10% Error		18.6	808,113	1.9	425,323	1.5	283,549
			16.7	900,054	1.7	529,444	1.4	378,174
Cats	+10% Error	48,276	14.9	324,000	1.1	294,546	2.4	122,728
	−10% Error		13.5	357,600	1.0	357,600	2.2	162,546
			12.2	395,705	0.9	439,673	2.0	219,836
Horses	+10% Error	5,402	10.2	52,961	1.0	52,961	2.3	23,027
	−10% Error		9.3	58,086	0.9	64,540	2.1[b]	30,734
			8.4	64,310	0.8	80,387	1.9	42,309
Pet Birds	+10% Error	350	24.0	1,459	0.2	7,292	3.4	2,145
	−10% Error		21.8	1,606	0.2	8,030	3.1[b]	2,591
			19.6	1,786	0.2	8,929	2.8	3,189

[a] U.S. Pet Ownership and Demographics, American Veterinary Medical Association, 2002

[b] Rate based on northeast central region of U.S that includes Ohio

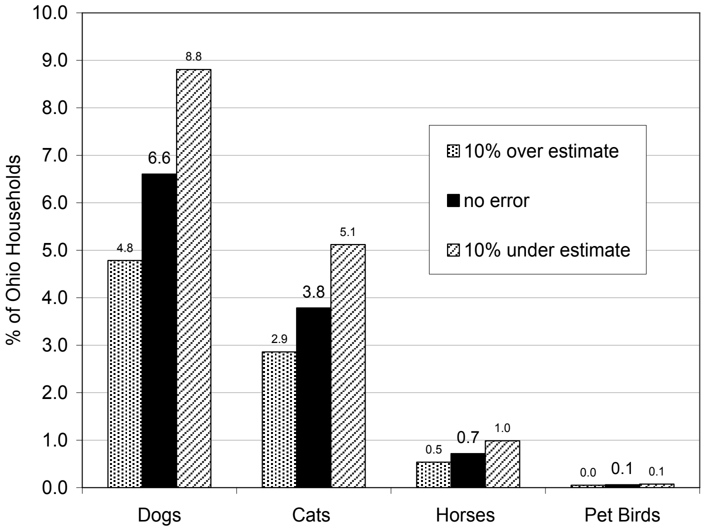


Figure 5-4. Estimates of Ohio household representation by animals with consideration of possible over- and underestimation of AVMA Household Survey values. Total Ohio households estimated at 4,293,649 from the 2002 U.S. census.

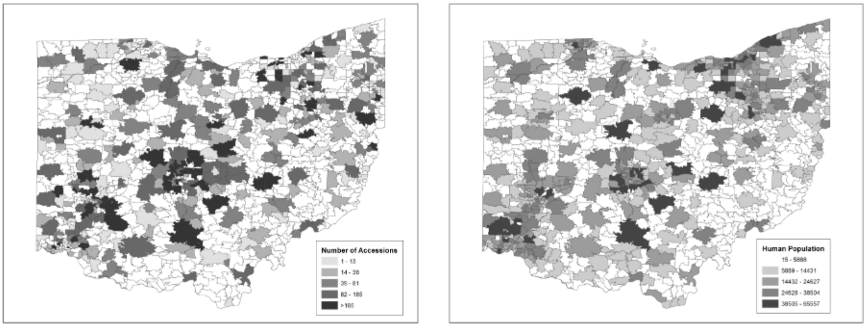


Figure 5-5. Distribution by ZIP code of specimens submitted for veterinary laboratory analysis from clinics in Ohio between April 1, 2005 and March 31, 2006 (**a**) and human population (**b**).

3.2.4 Availability and Timeliness

Timely detection of disease outbreaks is a critical concern of biosurveillance system users and thus should be a primary goal for biosurveillance system developers. Timeliness of detection can be measured in terms of latency (the difference in time between the event of interest and when it is discovered) and is reflected by the availability of data.

In a prospective study, the author assessed availability by determining the lag that existed between the time VDL personnel created a record and when that complete record was received for analysis (Shaffer, 2007). The VDL sent data in batches via file transfer protocol (ftp) once each day at 11 AM. Each message contained all the available records from the previous 24-h period, including those created the same morning. The VDL received samples mostly via private delivery service (e.g., FedEx) each morning. The delivery time to the laboratory was not representative of the time of sample collection. In other words, samples would not arrive at the VDL in the same order or with the same delay from when they were collected. While the VDL could provide records in real-time as they were created, the normal workflow that included the transport of specimens from the veterinary provider to the VDL did not support this added expense.

This assessment determined that each batch record message contained, on average, approximately one-third of the total number of records created that same day. With the receipt of the subsequent day's message, 95% of records were captured (see Figure 5-6). Other results from this same prospective study indicated that discovery of disease using these veterinary laboratory test order records could precede reports of an associated outbreak in humans by as much as 21 days.

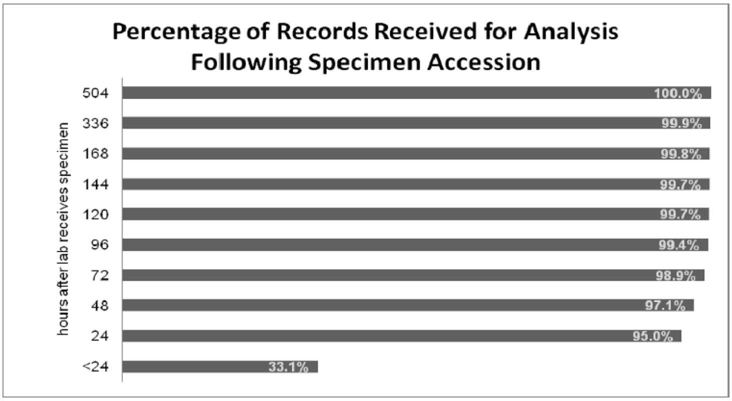


Figure 5-6. Availability of VDL order records for surveillance activities.

4. CONCLUSION

Following the introduction of *Bacillus anthracis* into the United States postal system in October 2001, there has been a heightened interest in developing disease surveillance that is capable of detecting outbreaks earlier than is generally possible through traditional reporting methods. Results from the retrospective investigation discussed here indicated that the true extent of outbreaks involving multiple species can go unrecognized by surveillance more commonly used by public health. Treating only humans in a multiple-species outbreak could leave a disease reservoir unchecked and provide for re-infection. Including pet animals in disease surveillance programs may also provide us with earlier recognition of an outbreak that has the potential of affecting humans. The prospective study completed by the author indicated that such recognition could come significantly earlier than the first reports of human cases.

Surveillance is critical to effective disease management and the early detection of zoonotic disease outbreaks. While no single data source is capable of capturing all of the information required for early outbreak detection, surveillance that includes data from both human and animal populations is critical to protect the health of each these groups. Additional investigations are necessary to thoroughly evaluate the sources of veterinary-based data and identify those that will be most beneficial for biosurveillance activities.

Healthcare providers, public health authorities, and others have come to recognize that an outbreak of disease does not always confine itself within political boundaries. We must also remain cognizant that, in many instances, disease does not confine infection to a single animal species. We have become more adept at working across our political jurisdictions to control disease and must now bridge our species-defined disciplines to improve our disease surveillance capabilities – for the health of all concerned.

QUESTIONS FOR DISCUSSION

1. Some feel that more clinically ill animals have an increased potential for receiving veterinary care that includes laboratory testing. Discuss how this might affect the sensitivity and specificity of a biosurveillance system. What other potential biases should be considered?
2. Although there is no legislation that is comparable to what exists for/that regarding human medical information (HIPAA), governing the release of veterinary data (and) privacy is still a concern. Discuss how privacy may be an issue for animal owners, veterinarians, and VDLs and how biosurveillance system developers might address this.
3. While agricultural species represented a small percentage of VDL accessions in the case example, how might outbreak surveillance for these species occur using these or similar data? In your discussion, you may want to consider how healthcare for pet animals is akin to human healthcare with individuals as the focus while most agricultural veterinary care is focused at the herd level.

4. The business of diagnostic testing necessitates that VDLs are regularly adding tests and marketing various tests as a single product in their catalog of available services. Discuss the process that biosurveillance system developers must include so that surveillance categories remain accurate. How might this affect the system architecture selected by the system developers?

REFERENCES

American Veterinary Medical Association, 2002, *U.S. Pet Ownership and Demographics Sourcebook*, AVMA, Schaumburg, IL.

Babin, S. M., Casper, J., Witt, C., Happel-Lewis, S. L., Wojcik, R. A., Magruder, S. F., et al., 2003, Early detection of possible bioterrorist events using sentinel animals. Paper presented at the 131st Annual Meeting of the American Public Health Association, November 15–19, 2003.

Backer, L., Grindem, C. B., Corbett, W. T., Cullins, L., and Hunter, J. L., 2001, Pet dogs as sentinels for environmental contamination, *Sci Total Environ*, **274**:161–169.

Bartlett, P. C., Kaneene, J. B., Kirk, J. H., Wilke, M. A., and Martenuik, J. V., 1986, Development of a computerized dairy herd health data base for epidemiologic research, *Prev Vet Med*, **4**(1):3–14.

Begier, E. M., Sockwell, D., Branch, L. M., Davies-Cole, J. O., Jones, L. H., Edwards, L., et al., 2003, The national capitol region's emergency department syndromic surveillance system: Do chief complaint and discharge diagnosis yield different results? *Emerg Infect Dis*, **9**(3):393–396.

Bender, J. B., and Shulman, S. A., 2004, Reports of zoonotic disease outbreaks associated with animal exhibits and availability of recommendations for preventing zoonotic disease transmission from animals to people in such settings, *J Am Vet Med Assoc*, **224**(7):1105–1109.

Bradley, C. A., Rolka, H., Walker, D., and Loonsk, J., 2005, BioSense: Implementation of a national early event detection and situational awareness system, *MMWR Morb Mortal Wkly Rep*, **54**(Suppl):11–19.

Center for Infectious Disease Research and Policy, 2006, Agencies launch warning system for animal diseases, (July 27, 2006); http://www.cidrap.umn.edu/cidrap/content/influenza/avianflu/news/jul2506warning.html.

Centers for Disease Control and Prevention, 2001, Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group, *MMWR Recomm Rep*, **50**(RR-13):1–35.

Childs, J., Shope, R. E., Fish, D., Meslin, F. X., Peters, C. J., Johnson, K., et al., 1998, Emerging zoonoses, *Emerg Infect Dis*, **4**(3):453–454.

Cleaveland, S., Laurenson, M. K., and Taylor, L. H., 2001, Diseases of humans and their domestic mammals: Pathogen characteristics, host range and the risk of emergence, *Philos Trans R Soc Lond B Biol Sci*, **356**:991–999.

Conner, C. F., 2005, Review of efforts to protect the agricultural sector and food supply from a deliberate attack with a biological agent, a toxin or a disease directed at crops and livestock, *Bio-Security and Agro-Terrorism*, United States Department of Agriculture, Washington, DC.

Crawford, P. C., Dubovi, E. J., Castleman, W. L., Stephenson, I., Gibbs, E. P. J., Chen, L., et al., 2005, Transmission of equine influenza virus to dogs, (September 29, 2005); http://www.sciencexpress.org/10.1126/science.1117950.

Das, D., Metzger, K., Heffernan, R., Balter, S., Weiss, D., and Mostashari, F., 2005, Monitoring over-the-counter medication sales for early detection of disease outbreaks – New York City, *MMWR Morb Mortal Wkly Rep*, **54**(Suppl):41–46.

Dato, V., Wagner, M. M., and Fapohunda, A., 2004, How outbreaks of infectious disease are detected: A review of surveillance systems and outbreaks, *Public Health Rep*, **119**:464–471.

DeGroot, B., 2005, *The Rapid Syndrome Validation Project for Animals – Augmenting Contact with the Network of Accredited Veterinarians*, NAHSS Outlook, (April), United States Department of Agriculture, Ft. Collins, CO.

Eidson, M., Kramer, L., Stone, W., Hagiwara, Y., and Schmit, K., 2001a, Dead bird surveillance as an early warning system for West Nile Virus, *Emerg Infect Dis*, **7**(4):631–635.

Eidson, M., Miller, J., Kramer, L., Cherry, B., and Hagiwara, Y., 2001b, Dead crow densities and human cases of West Nile Virus, New York State, 2000, *Emerg Infect Dis*, **7**(4):662–664.

Fienberg, S. E., and Shmueli, G., 2005, Statistical issues and challenges associated with rapid detection of bio-terrorist attacks, *Stat Med*, **24**:513–529.

Gill, J. S., McLean, R. G., Shriner, R. B., and Johnson, R. C., 1994, Serologic surveillance for the Lyme disease spirochete, *Borrelia burgdorferi*, in Minnesota by using white-tailed deer as sentinel animals, *J Clin Microbiol*, **32**(2):444–451.

Glickman, L. T., Moore, G. E., Glickman, N. W., Caldanaro, R. J., Aucoin, D., and Lewis, H. B., 2006, Purdue University-Banfield national companion animal surveillance program for emerging and zoonotic diseases, *Vector Borne Zoonotic Dis*, **6**(1):14–23.

Guptill, S., Julian, K. G., Campbell, G. L., and Marfin, A. A., 2003, Early-season avian deaths from West Nile Virus as warnings of human infection, *Emerg Infect Dis*, **9**(4):483–484.

Hill, S. L., Cheney, J. M., Taton-Allen, G. F., Reif, J. S., Bruns, C., and Lappin, M. R., 2000, Prevalence of enteric zoonotic organisms in cats, *J Am Vet Med Assoc*, **216**(5):687–692.

Johnson, C. L., 2004, Electronic surveillance for emerging diseases, Paper presented at the National Veterinary Electronic Surveillance Meeting, July 8, 2004, Raleigh, NC.

Johnson, H. A., Wagner, M. M., and Saladino, R. A., 2005, *A New Method for Investigating Non-traditional Biosurveillance Data: Studying Behaviors Prior to Emergency Department Visits*, RODS Laboratory, Pittsburgh, PA.

Kaye, D., and Pringle, C. R., 2005, Avian influenza viruses and their implication for human health, *Clin Infect Dis*, **40**:108–112.

Kearney B. Strengthening Safeguards Against Disease Outbreaks. *In Focus*. Washington, D.C.: The National Academy of Sciences, 2005.

Kruse, H., Kirkemo, A.-M., and Handeland, K., 2004, Wildlife as source of zoonotic infections, *Emerg Infect Dis*, **10**(12):2067–2072.

Lazarus R., Kleinman K.P., Dashevsky I., et al., Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *Public Health* 2001; 1.

Ma H., Rolka H., Mandl K., et al., Implementation of laboratory order data in BioSense Early Event Detection and Situation Awareness System. *MMWR Morb Mortal Wkly rep* 2005; **54**:27–30.

Mandl, K. D., Overhage, M. J., Wagner, M. M., Lober, W. B., Sebastiani, P., Mostashari, F., et al., 2004, Implementing syndromic surveillance: A practical guide informed by the early experience, *J Am Med Inform Assoc*, **11**(2):141–150.

Mark L. Strengthening Veterinary Diagnostic Lab Network is Critical. *USAHA News*. Richmond, VA: United States Animal Health Association, 2006.

Morris, C. D., Baker, W. G., Stark, L., Burgess, J., and Lewis, A. L., 1994, Comparison of chickens and pheasants as sentinels for eastern equine encephalitis and St. Louis encephalitis viruses in Florida, *J Am Mosq Control Assoc*, **10**(4):545–548.

National Research Council, 2005, *Animal Health at the Crossroads: Preventing, Detecting, and Diagnosing Animal Diseases*, National Academy of Sciences, Washington, DC.

Robotham, J., and Green, L. E., 2004, Pilot study to investigate the feasibility of surveillance of small animals in the UK, *J Small Anim Pract*, **45**(4):213–218.

Schwabe, C. W., 1984, *Veterinary Medicine and Human Health*, 3rd ed., Williams and Wilkins, Baltimore, MD.

Shaffer, L., Funk, J., Rajala-Schultz, P., Wallstrom, G., Wittum, T., Wagner, M., et al., 2007, Early outbreak detection using an automated data feed of test orders from a veterinary diagnostic laboratory, *Springer Lect Notes Comput Sci*, **4506**:1–10.

Shaffer, L. E., 2007, *Estimate of Human Population Representation from Veterinary Diagnostic Laboratory Orders*, Health Monitoring Systems, Inc., Pittsburgh, PA.

Shaffer, L. E., Funk, J. A., Rajala-Schultz, P., Wagner, M. M., Wittum, T. E., and Saville, W. J. A., 2008, Evaluation of microbiology orders from veterinary diagnostic laboratories as a potential data source for early outbreak detection, *Adv Dis Surveill*, **6**(2):1–7.

Sonesson, C., and Bock, D., 2003, A review and discussion of prospective statistical surveillance in public health, *J R Stat Soc A*, **166**(Part 1):5–21.

Sosin, D. M., 2003, Draft framework for evaluating syndromic surveillance systems, *J Urban Health*, **80**(2 Suppl 1):i8–i13.

Tansey, B., 2006, Debate over pets' role in spread of avian flu: No cases passed yet by dogs or cats, but experts want study, June 12, 2006, *San Francisco Chronicle*.

Tauni, M. A., and Österlund, A., 2000, Outbreak of *Salmonella typhimurium* in cats and humans associated with infection in wild birds, *J Small Anim Pract*, **41**:339–341.

Taylor, L. H., Latham, S. M., and Woolhouse, M. E. J., 2001, Risk factors for human disease emergence, *Philos Trans R Soc Lond B Biol Sci*, **356**:983–989.

Tsui, F.-C., Espino, J. U., Dato, V. M., Gesteland, P. H., Hutman, J., and Wagner, M. M., 2003, Technical description of RODS: A real-time public health surveillance system, *J Am Med Inform Assoc*, **10**(5):399–408.

Wagner, M. M., Robinson, J. M., Tsui, F. C., Espino, J. U., and Hogan, W. R., 2003, Design of a national retail data monitor for public health surveillance, *J Am Med Inform Assoc*, **10**(5):409–418.

Woolhouse, M. E. J., 2002, Population biology of emerging and re-emerging pathogens, *Trends Microbiol*, **10**(Suppl):S3–S7.

Wurtz, R. M., and Popovich, M. L., 2002, *Animal Disease Surveillance: A Framework for Supporting Disease Detection in Public Health*, Scientific Technologies Corporation Tuscon, AZ.

Zeng, X., and Wagner, M., 2002, Modeling the effects of epidemics on routinely collected data, *J Am Med Inform Assoc*, **9**(Suppl 6):S17–S22.

SUGGESTED READING

Salman, M.D., ed., 2003, *Animal Disease Surveillance and Survey Systems: Methods and Applications*, Iowa State Press, Ames, Iowa.

Burroughs, T., Knobler, S., and Lederberg, J., eds., 2002, *The Emergence of Zoonotic Diseases: Understanding the Impact on Animal and Human Health – Workshop Summary*, National Academy Press, Washington, DC.

Karlen, A., 1995, *Man and Microbes: Disease and Plagues in History and Modern Times*, Simon and Schuster, New York, NY.

Smolinski, M.S., Hamburg, M.A., and Lederberg, J., eds., 2003, *Microbial Threats to Health: Emergence, Detection, and Response*, National Academy Press, Washington, DC.

Drexler, M., 2001, *Secret Agents: The Menace of Emerging Infections*, Joseph Henry Press, Washington, DC.

Schwabe, C.W., 1984, *Veterinary Medicine and Human Health*, Williams and Wilkins, Baltimore, MD.

ONLINE RESOURCES

The American Association of Veterinary Laboratory Diagnosticians (AAVLD). http://www.aavld.org/mc/page.do.

American Veterinary Medical Association (AVMA). http://www.avma.org/.

Association for Veterinary Informatics. http://www.avinformatics.org/index.htm.

Canary Database. *Animals as Sentinels of Human Environmental Health Hazards*. http://www.canarydatabase.org/.

Centers for Disease Control and Prevention. *Nationally Notifiable Infectious Diseases United States, 2006*. http://www.cdc.gov/epo/dphsi/phs/infdis2006.htm.

National Center for Infectious Diseases (NCID). *Healthy Pets Healthy People*. http://www.cdc.gov/healthypets/.

Office International des Epizooties (OIE). *Diseases Notifiable to the World Organisation for Animal Health*. http://www.oie.int/eng/maladies/en_classification.htm.

UNIT II: SURVEILLANCE ANALYTICS

Chapter 6

MARKOV SWITCHING MODELS FOR OUTBREAK DETECTION

HSIN-MIN LU*, DANIEL ZENG, and HSINCHUN CHEN

CHAPTER OVERVIEW

Infectious disease outbreak detection is one of the main objectives of syndromic surveillance systems. Accurate and timely detection can provide valuable information for public health officials to react to major public health threats. However, disease outbreaks are often not directly observable. Moreover, additional noise caused by routine behavioral patterns and special events further complicates the task of identifying abnormal patterns caused by infectious disease outbreaks. We consider the problem of identifying outbreak patterns in a syndrome count time series using the Markov switching models. The outbreak states are treated as hidden (unobservable) state variables. Gibbs sampler then is used to estimate both the parameters and hidden state variables. We cover both the theoretical foundation of the estimation methods and the technical details of estimating the Markov switching models. A case study is presented in the last section.

Keywords: Markov switching models; Infectious disease informatics; Markov chain Monte Carlo; Gibbs sampler; Bayesian inference

* *Department of Information Management, National Taiwan University, Taipei, Taiwan. luim@ntu.edu.tw*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_6,

1. INTRODUCTION

Recent efforts in building syndromic surveillance systems try to increase the timeliness of the data collection process by incorporating novel data sources such as emergency department (ED) chief complaints (CCs) and over-the-counter (OTC) health product sales. Studies show that these data sources do contain valuable information reflecting current public health status (Espino and Wagner, 2001; Ivanov et al., 2002; Chapman et al., 2005a, b). However, they usually carry substantial noise that may interfere with the detection of infectious disease outbreaks.

To overcome the problem, researchers have been working on developing statistical methods that can extract disease outbreak signals from the real-time data provided by syndromic surveillance systems. Typically, the data are classified and aggregated to generate univariate or multivariate time series at daily frequency. A univariate time series may be the daily counts of patients with a particular syndrome (for example, the gastrointestinal syndrome) from an ED. A multivariate time series may be the daily number of patients with a particular syndrome from multiple EDs. If geographic information such as the ZIP code is available, the multivariate time series may be the daily counts of patients with a particular syndrome from the ZIP code areas near an ED.

A popular time series outbreak detection method in current literature is a two-step procedure (Reis and Mandl, 2003; Reis et al., 2003). At the first step, a baseline model describing the "normal pattern" is estimated using the training data (usually a historical time series without outbreaks). The baseline model then is used to predict future time series values. At the second step, statistical surveillance methods such as the Stewart control chart (Shewhart, 1939; Montgomery, 2005) or the Cumulated SUM (CUSUM) (Page, 1954) method then take the prediction error (observed value minus predicted value) as the input and output outbreak scores. Higher outbreak scores are usually associated with higher risk of having an outbreak. When the outbreak scores exceed a predefined threshold, the alarm is triggered.

The two-step procedure operates based on the assumption that there are no outbreaks in the training data. This assumption, nevertheless, can only be verified when a simulated time series is used. When a real-world dataset is used, the assumption is very hard to verify because researchers have no control over the health status of the community involved in generating the dataset. Moreover, a full investigation of disease outbreaks during the data collection period is usually too expensive to conduct.

The validity of the detection results can be seriously impaired if we cannot verify that the training data are outbreak-free. The estimated parameters of the baseline model may be biased by outbreak-related observations.

Subsequent prediction and outbreak detection, as a result, may be negatively affected. When no outbreaks are detected, we cannot rule out the possibility that the assumption is violated. The problem can seriously reduce the practical value of the outbreak detection method.

To deal with the problem of having outbreak-related observations in training data, we need to have a model that is flexible and smart so that it can adjust itself automatically when outbreak-related observations exist. This is usually referred to as modeling endogenous structure changes (Clements and Hendry, 2006) in the econometrics literature. A structure change is the change of the underlying system parameters such that the system dynamics become different. The outbreak of infectious disease causes a structure change of the observed time series and thus fits naturally into the framework of endogenous structure change modeling.

A natural way of modeling structure changes in a time series is introducing additional hidden state variables which determine the underlying time series dynamics. One popular model of this kind is the Markov switching models proposed by Hamilton (Hamilton, 1989). This family of models has a hidden state variable that may have a different value in each period. The hidden state variable takes the value of either 0 or 1 and controls the conditional mean, variance, and autocorrelation of the time series. It evolves following a Markov process. That is, its current hidden state depends only on the historical values in the last few periods.

The estimation process for the Markov switching model, nevertheless, is much more complicated than that of the standard time series models such as the ARIMA models. After writing down the likelihood as a function of the parameters and the hidden state variable, we face the curse of dimensionality since the number of unknown hidden state values is at least as much as the number of periods. Typical numerical optimization routines can only deal with a function of no more than dozens of variables unless significant efforts are invested to fine-tune the routine. Fortunately, the expectation-maximization (EM) algorithm (Dempster et al., 1977), Gibbs sampler, and Markov Chain Monte Carlo (MCMC) (Albert and Chib, 1993; Carter and Kohn, 1994; Chib and Greenberg, 1995) can be used to estimate the parameters efficiently. Depending on the actual setting used, these methods often require a certain level of customization. We will have a more detailed discussion in the next section.

Equipped with the Markov switching models, we contribute to the outbreak detection research in at least two directions. First, the Markov switching models can be combined with existing two-step procedures to help improve the validity. For outbreak detection research, Markov switching can be used to identify potential historical outbreak periods if real-world datasets are used. The identified outbreaks can help researchers refine their gold standards used to compute their algorithm performance. For practitioners using existing outbreak detection methods on a daily basis, the Markov switching methods can be used to choose a training period that is outbreak-free. A wisely chosen training dataset can potentially improve the validity of detection results.

Second, existing Markov switching models can be modified to be applied directly to identify potential outbreaks in a prospective setting. Most applications of the Markov switching models are in a retrospective setting, i.e., identifying potential outbreaks in a historical dataset. However, in syndromic surveillance we need to identify potential outbreaks in real-time. To achieve this, we need to update the model every time a new observation is available. Due to the nature of the estimation methods, the estimated parameters may not have consistent semantics each time the model is re-estimated. That is, state 0 may represent either the outbreak or the non-outbreak state. This problem can seriously hinder the ability to automate the detection process. We are actively researching for novel approaches to solve the problem. This chapter will focus mainly on applying Markov switching models in a retrospective setting.

The rest of the chapter is structured as follows. Section 2 briefly introduces the Markov switching models and points out the technical difficulties involved in model estimation. Section 3 provides an illustrative example of Bayesian inference and then discusses the theoretical foundation. A sketch of proof for MCMC is given. Section 4 derives the conditional posterior distributions that are essential for Bayesian inference and discusses major steps for model estimation. Finally, Section 5 presents the BioPortal Outbreak Detection System (BiODS), which uses the Markov switching model for outbreak detections. The model estimation results using a real-world dataset are summarized with a brief discussion.

2. MARKOV SWITCHING MODELS

The Markov switching models use hidden state variables to control the dynamics of a time series. They belong to a broader family of models – state-space models. One well-known example of the state-space model is the linear dynamic model (Kalman, 1960; Harvey, 1989), which has one or more continuous state variables for each period. There are two kinds of equations in this model: measurement equations and transition equations (Kim and Nelson, 1999). The measurement equations define how unobservable states affect the observable random variables. The transition equations, on the other hand, define how the state variables evolve over time.

When the state variable is discrete, the state-space model is usually called the hidden Markov model (Baum and Petrie, 1966) or the Markov switching model (Hamilton, 1989) depending on the setting of the measurement equation. The measurement equation in the hidden Markov model is usually formulated so that the observed random variables at period t only depend on other observable random variables and unobservable state variables at the same period.

The observable random variables in the Markov switching model, on the other hand, usually depend on other observable random variables before period t. This setting makes the Markov switching models more suitable to deal with time series-related problems. Most applications of the Markov switching models fall in the field of economics and finance. Notable examples include the identification of macroeconomics business cycles (Hamilton, 1989) and the modeling of changing regimes of interest rates (Dahlquist and Gray, 2000).

To keep the discussion simple, we focus on univariate time series only. The case of multivariate time series can be readily generalized from the univariate case. Consider a time series $Y = (y_1, y_2, ..., y_T)$. There may be zero, one or more outbreaks during the observed periods. Let s_t denote the hidden outbreak state variable that takes value of either zero or one depending on whether there is an outbreak at period t. We assume that the observed time series is stationary and the unconditional (long-term) mean and variance exist. When an outbreak occurs, the time series dynamics change and the observed random variable moves toward the new long-term mean and variance associated with the outbreak. A simple Markov switching model that can capture the occurrences of outbreaks can be written as follows:

$$y_t = a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1} + e_t \tag{6-1}$$

$$s_t \in \{0,1\} \tag{6-2}$$

$$P(s_t = j \mid s_{t-1} = i) = p_{ij} \tag{6-3}$$

$$e_t \sim N(0, \sigma^2) \tag{6-4}$$

Equation 6-1 models the observed time series y_t using an auto-regressive model with one lag (AR(1)). The constant term is a_{00} or $a_{00} + a_{01}$ depending on the value of s_t; the coefficient associated with the first lag term y_{t-1} is a_{10} or $a_{10} + a_{11}$ depending on the value of s_t. The hidden state variable s_t is either 0 or 1. Having $s_t = 1$ ($s_t = 0$) indicates that there is an outbreak (no outbreaks) at period t. Equation 6-3 specifies that the evolution of s_t follows a Markov process with transition probability p_{ij}. That is, the state of the next period only depends on the state of the current period.

Consider the case when s_t is known and is constant across time, then under the stationary assumption, $E[y_t] = E[y_{t-1}] \equiv E[y]$. As a result, the unconditional mean of the time series, $E[y]$, can be calculated by taking expectation on both side of Equation 6-1:

$$E[y] = \frac{a_{0,0} + a_{0,1}s_t}{1 - (a_{1,0} + a_{1,1}s_t)} \tag{6-5}$$

Note that s_t determines the value of $E[y]$ in the above equation.

Equations 6-1 to 6-4 define a basic Markov switching model that can capture the changing dynamics caused by disease outbreaks. In real-world applications, one may want to enrich the model by incorporating day-of-week effect, seasonal effect, environmental variables, independent jumps or spikes, and other factors that may affect the observed time series Y. Though including additional effects may complicate the estimation process, the major steps remain unchanged. As such, our discuss focuses on the simple Markov switching model presented above.

2.1 Time Series Generated by Markov Switching Models: An Illustrative Example

To further illustrate the characteristics of the Markov switching models, we present a time series generated by the Markov switching models. By setting model parameters to reasonable values, we can generate a time series that looks like those observed from the syndromic surveillance systems. This can not only give us intuitions about the model but also provide indirect evidence about the outbreak detection ability of the Markov switching models.

We use Equations 6-1 to 6-4 to simulate a time series with outbreaks. Table 6-1 summarizes the parameters used in the simulation. During non-outbreak periods, the time series follows an AR(1) process with an unconditional (long-term) mean of 142.8. When an outbreak occurs, the autocorrelation increases from 0.3 to 0.7 and the unconditional mean increases to 233.3. The hidden outbreak states follow a Markov process with transition probability $p_{00} = 0.995$ and $p_{11} = 0.985$. The transition probability is consistent with the estimated values using real-world datasets.

Figure 6-1 plots the simulated time series and underlying true outbreak states. The association between the true outbreak states and the observed time series can be visually verified. Note that visual inspection may indicate that there are other outbreaks in this dataset. For example, there are two small peaks from days 400 to 550. The peaks, however, are caused by the underlying random process only and are not associated with any outbreaks.

Table 6-1. Simulation parameters for the Markov switching model.

Coefficient	Value
a_{00}	100
a_{01}	–30
a_{10}	0.30
a_{11}	0.40
σ^2	30^2
p_{00}	0.995
p_{11}	0.985
Non-outbreak average	142.8
Outbreak average	233.3

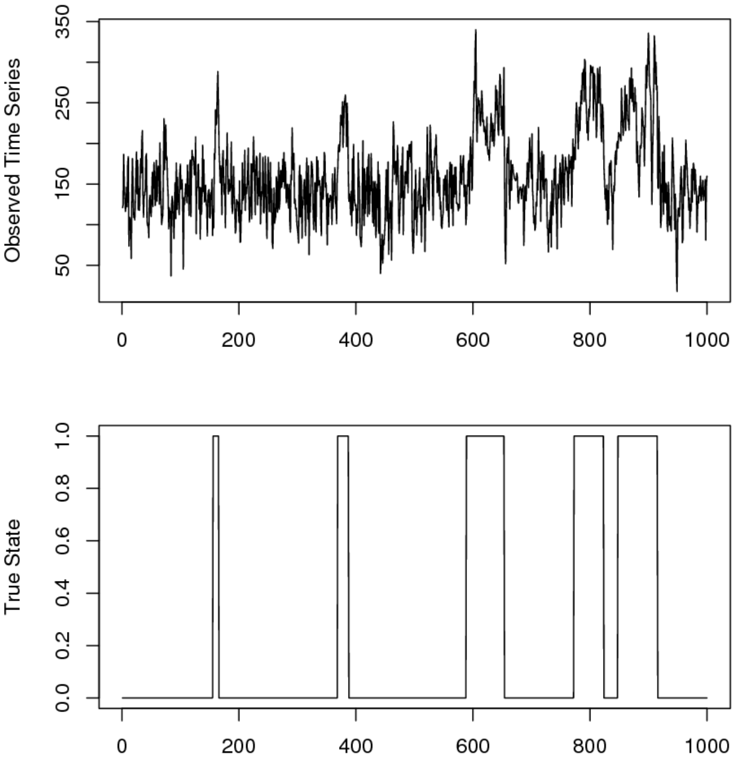


Figure 6-1. Simulated time series using a Markov switching model.

2.2 Estimation Methods for Markov Switching Models

The Markov switching models provide a powerful framework to capture the changing dynamics caused by infectious diseases. The hidden state variable can tell us, at each period, the probability of having an outbreak. However, the hidden state variable is unobservable and thus requires special attention during the estimation process.

Let us first pretend that the hidden state s_t is observable. The conditional distribution of y_t given y_{t-1} , s_t and parameters $\mathrm{A} \equiv \{a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1}, \sigma^2, p_{00}, p_{11}\}$ can be written as:

$$f(y_t \mid y_{t-1}, s_t, \mathrm{A}) = \frac{1}{\sqrt{2\pi\sigma}} \exp\{\frac{-[y_t - (a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1})]^2}{2\sigma^2}\}$$

It is more convenient to deal with the logarithm of the conditional distribution:

$$\ln f(y_t \mid y_{t-1}, s_t, \mathrm{A}) =$$

$$\frac{-1}{2}\{\ln 2\pi + \ln \sigma^2 + \frac{[y_t - (a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1})]^2}{\sigma^2}\}$$

Since y_t is independent given y_{t-1} and s_t, the log-likelihood of Y given initial value y_0, parameter A, and the state vector $S = \{s_1, s_2, ..., s_T\}$ is simply the summation of the individual logarithms of the conditional distributions:

$$\ln f(Y \mid y_0, S, \mathrm{A}) = \sum_{t=1}^{T} \ln f(y_t \mid y_{t-1}, s_t, \mathrm{A}) \qquad (6\text{-}6)$$

Equation 6-6 is usually referred to as the complete log-likelihood function of Y. To keep the notation clean and easy to understand, the term y_0 will be suppressed in the subsequent discussion.

The trajectory of hidden state S across period 1 to T, nevertheless, is not observable. To make things even more complicated, each state depends on the previous state since s_t evolves following a Markov process. To be able to estimate the parameter A, we need to eliminate S from the complete log-likelihood. One way to achieve this is to make use of the following identity:

$$f(Y|\mathrm{A}) = \int f(Y \mid S, \mathrm{A}) f(S \mid \mathrm{A})$$

The function $f(Y \mid \mathrm{A})$ is usually referred to as the incomplete log-likelihood. The intuition behind the identity is that we eliminate S by computing the average of $f(Y \mid S, \mathrm{A})$ weighted by the probability of S. In our case, the random variable s_t takes only two values and we can replace the integral with a summation:

$$f(Y \mid \mathsf{A}) = \sum_{S \in \mathsf{S}} f(Y \mid S, \mathsf{A}) f(S \mid \mathsf{A})$$

The set S is a collection of all possible trajectory $S = (s_1, s_2, ..., s_T)$. The size of S grows exponentially as T increases. As a result, the computation quickly grows beyond the capacity of modern computers. In the classical state-space literature, a filtering procedure similar to the Kalman Filter (Kalman, 1960) can be applied (Hamilton, 1989). However, certain model settings can only be estimated using approximation, which may lead to unsatisfactory estimation results.

Two techniques are commonly used to deal with the estimation problem. The first technique is the EM algorithm proposed by Dempster (Dempster et al., 1977). The second technique is Bayesian inference using Markov Chain Monte Carlo (MCMC) and Gibbs sampler (Albert and Chib, 1993). The EM algorithm iterates between two steps: the expectation (E) step and the maximization (M) step. In the E-step, the expected value of s_t given current parameter estimation is computed. In the subsequent M-step, the hidden state variables in the incomplete log-likelihood are replaced with their expectations computed in the E-step. The value of the incomplete log-likelihood then is maximized with respect to the parameters. The process is repeated until convergence. The incomplete log-likelihood will always increase from iteration to iteration. Also, the EM algorithm will always converge to a local maximum, often slowly. If the confidence intervals of parameters are needed, one needs to compute the covariance matrix separately from the first or second derivatives of the incomplete log-likelihood. Moreover, in more complicated models, there may be no analytical solution for the expectation of the state variables. In these cases, Monte Carlo simulation is employed for computing the expectation of state variables and additional convergence-related issues need to be taken care of before the EM algorithm can be applied.

Bayesian inference takes a different route to solve the problem. Instead of trying to find the parameters that maximize the incomplete log-likelihood, the Bayesian inference constructs the joint posterior distribution of parameters and hidden state variables given observed time series Y. Let $\Theta = (\mathsf{A}, S)$. Applying the Bayes theorem, we have:

$$f(\Theta \mid Y) \propto f(Y \mid \Theta) f(\Theta)$$

That is, the joint posterior of parameters and hidden state variables is proportional to the complete likelihood times the prior distribution of parameters and hidden state variables.

The point estimates of parameters are computed directly from the posterior distribution. Unlike the EM algorithm where the parameters are treated as fixed values to compute the expected value of hidden states, the parameter estimation risk is taken into consideration by Bayesian inference. The joint posterior distribution of parameters and hidden state variables are usually constructed from a set of conditional posterior distributions. The procedure allows a complicated model to be divided into small pieces, solved separately, and combined to form the joint posterior of all parameters and hidden state variables. We provide an overview of Bayesian inference in the following section.

3. BAYESIAN INFERENCE: AN OVERVIEW

We introduce Bayesian inference in this section. A simple example is presented in Sect. 3.1. The following subsections then discuss the theoretical foundations of Bayesian inference. Readers not interested in the theory can read Sect. 3.1 and skip the rest of this section.

3.1 Maximum Likelihood Estimation and Bayesian Inference: An Illustrative Example

The goal of Bayesian inference is to summarize the posterior distribution of parameters (and hidden state variables, if any) given observations. As an illustrative example, we present the process of estimating regression parameters via both likelihood maximization and Bayesian inference. Consider a linear regression model:

$$Y = X\beta + e, e \sim N(0, \sigma^2 I)$$

where $\beta = (b_0, b_1, ..., b_{k-1})$. The matrix X is the collection of row vectors $x_t = (1, x_{t,1}, x_{t,2}, ..., x_{t,k-1})$. The white noise vector e is defined as $e = (e_1, e_2, ..., e_T)'$, which follows a multivariate normal distribution. The matrix I is an identity matrix. The log-likelihood is:

$$\ln f(Y \mid \mathrm{E}, X) = \frac{-T}{2}\ln 2\pi - \frac{T}{2}\ln \sigma^2 - \frac{1}{2}(Y - X\beta)'(Y - X\beta)\frac{1}{\sigma^2} \qquad (6\text{-}7)$$

where $\mathrm{E} \equiv \{\beta, \sigma^2\}$.

3.1.1 Likelihood Maximization

The maximum likelihood estimator (MLE) of E can be found by computing the first derivative of $\ln f(Y|\mathrm{E}, X)$ with respect to E and set it to zero:

$$\frac{\partial \ln f(Y|\mathrm{E},X)}{\partial \mathrm{E}}=0$$

Solving for β and σ^2, we have:

$$\hat{\beta}=(X'X)^{-1}X'Y \tag{6-8}$$

$$\hat{\sigma}^2=\hat{e}'\hat{e}/T \tag{6-9}$$

where $\hat{e}=Y-X\hat{\beta}$.

$$VAR(\beta)=(X'X)^{-1}\sigma^2 \tag{6-10}$$

Equations 6-8 and 6-9 are the MLE estimator of the regression model. Equation 6-10 is the variance of parameters, which can be used to compute the confidence intervals of regression coefficients. It is straightforward to check that the likelihood function is globally concave and thus has a unique global maximum.

3.1.2 Bayesian Inference

The goal of Bayesian inference is to characterize the posterior distribution $f(\beta,\sigma^2|Y,X)$. Similar to likelihood maximization, we want to know the point estimators and confidence intervals of the parameters. One popular idea to achieve this is to draw a sample from the posterior distribution $f(\beta,\sigma^2|Y,X)$. The point estimators and confidence intervals then can be computed from the sample directly.

By the Bayes theorem, we know that:

$$f(\beta,\sigma^2|Y,X)\propto f(Y|\beta,\sigma^2,X)f(\beta,\sigma^2|X) \tag{6-11}$$

The likelihood function $f(Y|\beta,\sigma^2,X)$ is the same as that presented in the previous subsection and can be found in Equation 6-7. The prior distribution $f(\beta,\sigma^2|X)$ needs to be determined to set up the posterior distribution. If we assign a constant to the prior distribution, then the posterior distribution is proportional to the likelihood function. Bayesian inference, in this case, is dealing with the same density function as likelihood maximization. By assigning a constant to the prior distribution, we impose almost no additional information on the inference process. This kind of prior distribution, nevertheless, is not a valid probability distribution and may not always be a good choice for prior distributions.

The other popular option is to use so-called conjugate priors. Combined with the likelihood function, the posterior distribution belongs to the same distribution family as the prior distribution. Prior distributions are usually assigned without considering the observed data. That is, $f(\beta,\sigma^2 \mid X) = f(\beta,\sigma^2)$. In the case of regression, the conjugate prior is normal-inverse-gamma. In other words, we assume that σ^2 follows an inverse gamma distribution and given σ^2, β follows a multivariate normal distribution. The above discussion can be summarized as:

$$f(\beta,\sigma^2 \mid X) = f(\beta,\sigma^2) = f(\beta \mid \sigma^2) f(\sigma^2) = MV(\beta; M, V) IG(\sigma^2; a, b) \quad (6\text{-}12)$$

Where

$$MV(\beta; M, V) = (2\pi)^{-\frac{K}{2}} \mid \Sigma_0 \mid^{-1/2} \exp\{-\frac{1}{2}(\beta - M) V^{-1}(\beta - M)\}$$

$$IG(\sigma^2; v, \lambda) = \frac{\lambda^v}{\Gamma(v)} (1/\sigma^2)^{v+1} \exp\left(-\lambda/\sigma^2\right)$$

The function $MV(\cdot)$ is the density of the multivariate normal distribution and $IG(\cdot)$ is the density of the inverse gamma distribution. Substituting Equation 6-12 back to Equation 6-11, we now have the posterior distribution of β and σ^2:

$$ln\ f(\beta,\sigma^2 \mid Y, X) \propto$$

$$-(v + 1 + T/2) \ln \sigma^2 - \frac{\lambda}{\sigma^2} -$$

$$\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) - \frac{1}{2}(\beta - M)' V^{-1}(\beta - M) \quad (6\text{-}13)$$

The next question is how do we draw a sample from a density like Equation 6-13. Drawing β and σ^2 simultaneously may not be an attractive option since significant efforts are required to customize a sampler. An alternative is to take advantage of the existing structure of the posterior distribution. If we fix σ^2, then $\beta \mid \sigma^2, Y, X$ follows a multivariate normal distribution:

$$f(\beta \mid \sigma^2, Y, X) \propto \exp\{-\frac{1}{2}(\beta - M)' V^{-1}(\beta - M) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\}$$

$$\propto \exp\{-\frac{1}{2}(\beta - \beta_1)\Sigma_1^{-1}(\beta - \beta_1)\}$$

where

$$\beta_1 = (V^{-1} + \sigma^{-2} X'X)^{-1}(V^{-1} M + \sigma^{-2} X'Y)$$

$$\Sigma_1 = (V^{-1} + \sigma^{-2} X'X)^{-1}$$

That is:

$$\beta \mid \sigma^2, Y, X : N(\beta_1, \Sigma_1)$$

Similarly, fixing β, $\sigma^2 \mid \beta, Y, X$ follows an inverse gamma distribution:

$$f(\sigma^2 \mid \beta, Y, X) \propto (\frac{1}{\sigma^2})^{v+T/2+1} \exp\{\frac{-1}{\sigma^2}(\lambda + \frac{e'e}{2})\}$$

That is:

$$\sigma^2 \mid \beta, Y, X : IG(v', \lambda')$$

$$v' = v + T/2$$

$$\lambda' = \lambda + (e'e)/2$$

Since both the multivariate normal and the inverse gamma distribution are well-known and have samplers readily available, our effort can be greatly reduced if we can construct the sampler of $\{\beta, \sigma^2\}$ from individual samplers. It turns out that we can construct the joint posterior of $\{\beta, \sigma^2\} \mid Y, X$ from two conditional posteriors: $\beta \mid \sigma^2, Y, X$ and $\sigma^2 \mid \beta, Y, X$. This procedure is the so-called Gibbs sampler.

The procedure works as follows. We first pick initial values $\beta^{(0)}$ and $\sigma^{2(0)}$ for β and σ^2. We then update β and σ^2 iteratively using the two conditional posteriors. Specifically, for iteration i:

1. Draw $\beta^{(i)}$ from $\beta \mid \sigma^{2(i-1)}, Y, X$.
2. Draw $\sigma^{2(i)}$ from $\sigma^2 \mid \beta^{(i)}, Y, X$.
3. Record $\beta^{(i)}$ and $\sigma^{2(i)}$.

The procedure is repeated for I iterations and we only collect random variables generated after B iterations ($0 < B < I$). Random variables generated from the first B iterations are discarded to minimize the impact of initial values. The sample $\{\beta^{(i)}, \sigma^{2(i)}\}_{i=B+1}^{I}$ then can be used to compute point estimators and confidence intervals.

A common practice is to use the posterior means as the point estimators:

$$\hat{\beta} = \frac{\sum_{i=B+1}^{I} \beta^{(i)}}{I - B + 1} \quad (6\text{-}14)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=B+1}^{I} \sigma^{2(i)}}{I - B + 1} \quad (6\text{-}15)$$

The confidence intervals can be computed directly from the percentile of the sample. For example, the 95% confidence interval of $\hat{\beta}$ can be constructed from the 2.5% and 97.5% percentile of $\beta^{(i)}$ using the sample $\{\beta^{(i)}, \sigma^{2(i)}\}_{i=B+1}^{I}$.

For a simple model like regression, Bayesian inference gives similar results to classical likelihood maximization. However, when a complicated model is involved, Bayesian inference may be more attractive because of the nature of Gibbs sampler: a complicated posterior distribution can be constructed from a collection of simpler conditional posteriors. The other advantage of Bayesian inference is that it can provide exact inference when the number of observations is small. In many cases, likelihood maximization computes confidence intervals based on the central limit theory. The approximation may be bad when the number of observations is small. Bayesian inference computes the confidence intervals directly using the sample drawn from the posterior and can provide more accurate confidence intervals when the sample size is small.

The prior distribution is also a valuable asset when a large number of regressors (independent variables) is involved. One can impose a normal prior for regression coefficients with mean zero and reasonable variances. Bayesian inference then will automatically assign coefficients close to zero to independent variables that do not have explanatory power to the dependent variable.

3.1.3 A Numerical Example

We provide a numerical example of Bayesian inference and likelihood maximization for linear regression. Simulated data was generated from a linear regression model. Gibbs sampler, introduced above, was used to conduct Bayesian inference. We report estimation results using different sample sizes.

Data Generating Process

We generate the data using the following model:

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_1 0 x_{i,10} + e_i \tag{6-16}$$

$$e_i \sim N(0, \sigma^2)$$

The regressors $x_i \equiv (x_{i,1}, x_{i,2}, \ldots, x_{i,10})$ was generated using a multivariate normal distribution. The variance of $x_{i,d}$ is 1 for $d = 1,2,\ldots,10$. Each of the last five variables was not correlated with the other nine variables. The correlation coefficients among the first five variables $(x_{i,1}, x_{i,2}, \ldots, x_{i,5})$ were 0.5. The true parameters used to generate y_i can be found in the first column of Table 6-2.

Table 6-2. Estimation results.

True value	MLE	Gibbs sampler	MLE	Gibbs sampler
	(15 Obs.)	(15 Obs.)	(30 Obs.)	(30 Obs.)
b_0 = 1.5	2.01 (1.23, 2.80)	1.53 (0.44, 2.42)	1.69 (1.15, 2.23)	1.56 (0.95, 2.11)
b_1 = 1	0.62 (−0.92, 2.16)	0.75 (−0.48, 1.95)	0.65 (0.12, 1.18)	0.68 (0.14, 1.25)
b_2 = 1	2.26 (1.08, 3.44)	1.46 (0.21, 2.56)	1.63 (1.02, 2.24)	1.36 (0.70, 1.98)
b_3 = 0	1.37 (−0.03, 2.78)	0.22 (−1.01, 1.39)	0.25 (−0.28, 0.77)	0.19 (−0.39, 0.75)
b_4 = −1	−2.60 (−4.02, −1.18)	−1.12 (−2.34, 0.24)	−1.63 (−2.23, −1.02)	−1.36 (−2.00, −0.71)
b_5 = −1	−1.89 (−2.89, −0.90)	−1.18 (−2.24, 0.03)	−1.63 (−2.16, −1.10)	−1.46 (−2.02, −0.90)
b_6 = 1	2.10 (1.34, 2.86)	1.29 (0.32, 2.15)	0.84 (0.45, 1.23)	0.85 (0.43, 1.28)
b_7 = 1	−0.02 (−1.06, 1.02)	0.45 (−0.63, 1.56)	0.39 (−0.07, 0.85)	0.31 (−0.19, 0.78)
b_8 = 0	0.32 (−0.61, 1.24)	0.25 (−0.74, 1.22)	−0.37 (−0.82, 0.09)	−0.26 (−0.75, 0.24)
b_9 = −1	−0.27 (−1.21, 0.66)	−0.05 (−1.04, 0.95)	−0.89 (−1.52, −0.27)	−0.90 (−1.56, −0.25)
b_{10} = −1	−0.68 (−1.85, 0.49)	−0.65 (−1.59, 0.41)	−1.36 (−1.76, −0.96)	−1.32 (−1.76, −0.90)
$\sigma^2 = 2.25$	0.90	2.54	0.87	1.25
RMSE	3.07	2.03	1.98	1.86

True value	MLE	Gibbs sampler	MLE	Gibbs sampler
	(300 Obs.)	(300 Obs.)	(3,000 Obs.)	(3,000 Obs.)
b_0 = 1.5	1.45 (1.28, 1.63)	1.44 (1.27, 1.62)	1.51 (1.45, 1.56)	1.51 (1.45, 1.56)
b_1 = 1	0.97 (0.74, 1.20)	0.95 (0.73, 1.19)	0.98 (0.91, 1.06)	0.98 (0.91, 1.06)
b_2 = 1	0.97 (0.73, 1.21)	0.96 (0.72, 1.20)	0.97 (0.90, 1.04)	0.97 (0.89, 1.04)
b_3 = 0	0.06 (−0.18, 0.29)	0.05 (−0.19, 0.29)	0.04 (−0.03, 0.11)	0.04 (−0.03, 0.11)
b_4 = −1	−1.06 (−1.28, −0.84)	−1.05 (−1.27, −0.83)	−1.06 (−1.13, −0.99)	−1.05 (−1.12, −0.98)
b_5 = −1	−0.75 (−0.98, −0.52)	−0.74 (−0.96, −0.51)	−0.97 (−1.05, −0.90)	−0.97 (−1.05, −0.90)
b_6 = 1	1.11 (0.94, 1.28)	1.10 (0.94, 1.28)	1.01 (0.95, 1.07)	1.01 (0.95, 1.07)
b_7 = 1	1.07 (0.90, 1.24)	1.06 (0.89, 1.22)	1.02 (0.96, 1.07)	1.02 (0.96, 1.07)
b_8 = 0	−0.02 (−0.20, 0.16)	−0.02 (−0.20, 0.16)	−0.02 (−0.07, 0.03)	−0.02 (−0.07, 0.03)
b_9 = −1	−0.96 (−1.14, −0.79)	−0.96 (−1.13, −0.78)	−0.99 (−1.05, −0.94)	−0.99 (−1.05, −0.93)
b_{10} = −1	−1.07 (−1.24, −0.90)	−1.06 (−1.24, −0.90)	−1.04 (−1.10, −0.98)	−1.04 (−1.10, −0.98)
$\sigma^2 = 2.25$	2.18	2.25	2.38	2.39
RMSE	1.56	1.56	1.53	1.53

We generated 3,000 pairs of y_i and x_i. The regression parameters then were estimated using the first 15, 30, 300, and 3,000 pairs of observations. We generated an additional 3,000 pairs of observations to compute prediction errors using estimated parameters. The additional 3,000 pairs of observations are referred to as the testing dataset.

Likelihood Maximization

The MLE estimators were calculated using Equations 6-8 to 6-10.

Bayesian Inference Using Gibbs Sampler

We conducted Bayesian inference using Gibbs sampler as described in Sect. 3.1.2. We chose $I = 3,000$ and $B = 1,000$. That is, only random numbers drawn during the last 2,000 iterations of the total 3,000 iterations were used to compute the estimated values. The prior distribution for β is a multivariate normal distribution with mean zero and variance 1. The prior distribution for σ^2 is an inverse gamma distribution with $v = 3$ and $\lambda = 1$. The initial values were set to the mean of the prior distributions. All regression coefficients had initial values equal to 0. The initial value for σ^2 is 0.5.

Estimation Results

We first take a look at the converging process of the Gibbs sampler. Before the sample constructed using the Gibbs sampler can be used for model estimation, we must make sure that the Markov process involved had indeed converged. Figure 6-2 plots the sampling values of b_2 for the first 300 iterations. It is clear that Gibbs sampler converged to the true value of b_2 quickly after a few iterations.

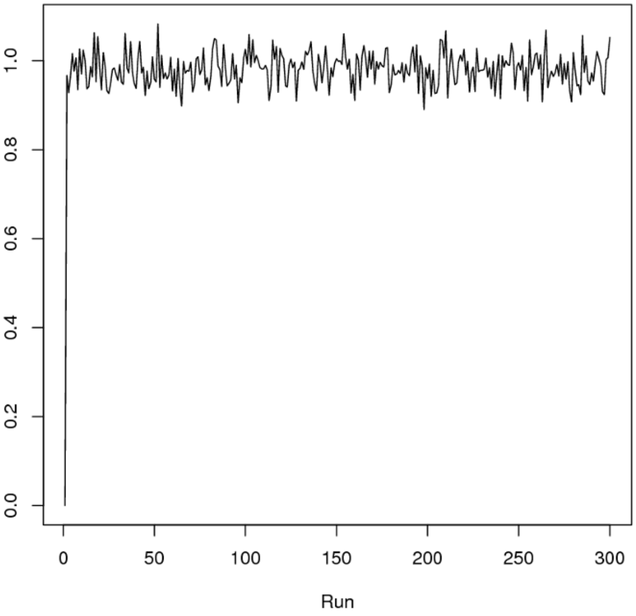


Figure 6-2. Sampling value of b_2 (sample size = 3,000).

Table 6-2 summarizes the estimation results. The second and third columns on the top panel report the estimation results of MLE and Gibbs sampler using the first 15 pairs of observations. Compared to MLE, Gibbs sampler gave less extreme values. For example, the estimations for b_2, b_4, and b_6 were 2.26, -2.60, and -1.89 using MLE. Gibbs sampler, on the other hand, gave more reasonable estimations (1.46, −1.12, and −1.18). Similar results can be observed when the first 30 pairs of observations were used (the fourth and fifth columns). When a large number of observations were used (reported on the lower panel), Gibbs sampler and MLE gave similar outcomes. Note that if we compare the the root mean square error (RMSE) using the testing dataset, Gibbs sampler gave smaller RMSE then that of MLE when the sample sizes were small (15 and 30 observations). The RMSE for large sample sizes were the same.

3.2 Markov Chain Monte Carlo and Gibbs Sampler

The MCMC method constructs the full posterior distribution by a collection of conditional posterior distributions with lower dimensions. Consider the full posterior distribution $f(\Theta | Y)$, where $\Theta = (\Theta_1, \Theta_2, ..., \Theta_k)$ is a vector of parameters and hidden state variables. The vector Y is the observed time series. Each component Θ_i could be a singleton or a vector. Without loss of generality, we assume Θ_i is a singleton in the following discussion. We can generate a Markov chain with $f(\Theta | Y)$ as its invariant distribution from the following conditional densities:

$$f(\Theta_1 | \Theta_2, \Theta_3, ..., \Theta_k, Y)$$

$$f(\Theta_2 | \Theta_1, \Theta_3, ..., \Theta_k, Y)$$

$$\vdots$$

$$f(\Theta_k | \Theta_1, \Theta_2, ..., \Theta_{k-1}, Y)$$

Given initial values $\Theta^{(0)} = (\Theta_1^{(0)}, \Theta_2^{(0)}, ..., \Theta_k^{(0)})$, the MCMC method generates Θ by updating each of the k elements sequentially. The first element $\Theta_1^{(1)}$ is drawn from $f(\Theta_1 | \Theta_2^{(0)}, \Theta_3^{(0)}, ..., \Theta_k^{(0)}, Y)$. Then, $\Theta_2^{(1)}$ is drawn from $f(\Theta_2 | \Theta_1^{(1)}, \Theta_3^{(0)}, ..., \Theta_k^{(0)}, Y)$. Note that Θ_1 has been updated with the latest sampling result. The process continues until all k elements have been updated. Repeating the above process creates a Markov chain with the following transition density:

$$T(\Theta^{(i)}, \Theta^{(i+1)}) = \prod_{z=1}^{k} f(\Theta_z^{(i+1)} | \Theta_1^{(i+1)}, \Theta_2^{(i+1)}, ..., \Theta_{z-1}^{(i+1)}, \Theta_{z+1}^{(i)}, ..., \Theta_k^{(i)}, Y)$$

If we can show that the invariant distribution of the transition density T is $f(\Theta \mid Y)$, then we can use the above procedure to break down a high-dimensional density into a collection of lower-dimensional densities. Specifically, we want to establish time-reversibility, which will lead us to the invariant distribution. The time-reversibility condition can be written as:

$$f(\Theta^{(i)} \mid Y)T(\Theta^{(i)},\Theta^{(i+1)}) = f(\Theta^{(i+1)} \mid Y)T(\Theta^{(i+1)},\Theta^{(i)})$$

To prove the time-reversibility condition, first note that we can write $f(\Theta_1^{(i)},\Theta_2^{(i)},...,\Theta_k^{(i)} \mid Y) = f(\Theta_1^{(i)} \mid \Theta_2^{(i)},...,\Theta_k^{(i)},Y) f(\Theta_2^{(i)},...,\Theta_k^{(i)} \mid Y)$. Following Besag (Besag, 1974), we can multiply and divide $f(\Theta_1^{(i+1)},\Theta_2^{(i)},...,\Theta_k^{(i)} \mid Y)$ at the right hand side and make the necessary rearrangement:

$$f(\Theta_1^{(i)},\Theta_2^{(i)},...,\Theta_k^{(i)} \mid Y) = \frac{f(\Theta_1^{(i)} \mid \Theta_2^{(i)},...,\Theta_k^{(i)},Y)}{f(\Theta_1^{(i+1)} \mid \Theta_2^{(i)},...,\Theta_k^{(i)},Y)} f(\Theta_1^{(i+1)}\Theta_2^{(i)},...,\Theta_k^{(i)} \mid Y)$$

Repeating the procedure on $\Theta_2^{(i)}$ for the last term at the right hand side, the equation becomes:

$$f(\Theta_1^{(i)},\Theta_2^{(i)},...,\Theta_k^{(i)} \mid Y) = \frac{f(\Theta_1^{(i)} \mid \Theta_2^{(i)},...,\Theta_k^{(i)},Y)}{f(\Theta_1^{(i+1)} \mid \Theta_2^{(i)},...,\Theta_k^{(i)},Y)} \frac{f(\Theta_2^{(i)} \mid \Theta_1^{(i+1)},\Theta_3^{(i)},...,\Theta_k^{(i)},Y)}{f(\Theta_2^{(i+1)} \mid \Theta_1^{(i+1)},\Theta_3^{(i)},...,\Theta_k^{(i)},Y)}$$
$$f(\Theta_1^{(i+1)}\Theta_2^{(i+1)},\Theta_3^{(i)}...,\Theta_k^{(i)} \mid Y)$$

Continuing for all $\Theta_z^{(i)}$, we have:

$$\frac{f(\Theta^{(i)} \mid Y)}{f(\Theta^{(i+1)} \mid Y)} = \prod_{z=1}^{k} \frac{f(\Theta_z^{(i)} \mid \Theta_1^{(i)},\Theta_2^{(i)},...,\Theta_{z-1}^{(i)},\Theta_{z+1}^{(i+1)},\Theta_k^{(i+1)},Y)}{f(\Theta_z^{(i+1)} \mid \Theta_1^{(i+1)},\Theta_2^{(i+1)},...,\Theta_{z-1}^{(i+1)},\Theta_{z+1}^{(i)},\Theta_k^{(i)},Y)} = \frac{T(\Theta^{(i+1)},\Theta^{(i)})}{T(\Theta^{(i)},\Theta^{(i+1)})}$$

which is exactly the time-reversibility condition. It is easy to verify that $f(\Theta \mid Y)$ is the invariant distribution of the Markov chain. When direct sampling of all $\Theta_z^{(i)}$ is possible, the above sampling procedure is called the Gibbs sampler (Geman and Geman, 1984).

4. CONDITIONAL POSTERIOR DISTRIBUTIONS OF THE MARKOV SWITCHING MODELS

The discussion of the MCMC and Gibbs sampler suggests that we can construct the joint posterior distribution $f(\Theta \mid Y)$ by a set of conditional distributions with lower dimensions. The strategy is to take advantage of the model structure and simplify the derivation. For the sake of convenience, the Markov switching model is restated here:

$$y_t = a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1} + e_t \tag{6-17}$$

$$s_t \in \{0,1\} \tag{6-18}$$

$$P(s_t = j \mid s_{t-1} = i) = p_{ij} \tag{6-19}$$

$$e_t : N(0,\sigma^2) \tag{6-20}$$

A few useful notes should be mentioned regarding the equations. First, if the state vector $S = (s_1, s_2, ..., s_T)$ is known, then the standard Bayesian linear model analysis technique can be used to compute the posterior distribution of $a_{0,0}$, $a_{0,1}$, $a_{1,0}$, $a_{1,1}$, σ^2. Second, the posterior distribution of the transition probability depends on the state vector S only. Finally, derivation of the posterior distribution of S can be simplified by taking parameters in A as fixed numbers.

The tactic is clear from the above discussion. We can divide the parameters and hidden state variables into three sets. The first set is regression-related parameters $a_{0,0}$, $a_{0,1}$, $a_{1,0}$, $a_{1,1}$, and σ^2. The second set is transition-related parameters p_{00} and p_{11}. The third set is the hidden state variable S. The conditional posterior distribution of one set of random variables given the other two sets can be derived. We can then iterate through these three sets of conditional distributions to generate the joint posterior distribution $f(\Theta \mid Y)$.

4.1 Conditional Posterior Distributions of Regression Parameters

To derive the posterior distribution of regression parameters, we first rewrite Equation 6-17 in matrix representation:

$$Y = X\beta + e, e \sim N(0, \sigma^2 I) \tag{6-21}$$

where $\beta = (a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1})$; X is the collection of row vector $x_t = (1, s_t, y_{t-1}, s_t y_{t-1})$; $e = (e_1, e_2, ..., e_T)'$.

The regression parameters are further divided into two sets: β and σ^2. We present the derivation of these two conditional posterior distributions in sequence:

- $f(\beta \mid \sigma^2, Y, S, p_{ij})$
- $f(\sigma^2 \mid \beta, Y, S, p_{ij})$

4.1.1 Deriving $(\beta \mid \sigma^2, Y, S, p_{ij})$

Using the Bayes theory, the conditional posterior of β is the complete likelihood function times the prior distribution of β:

$$f(\beta \mid Y,S,p_{ij}) \propto f(Y \mid \beta,S,p_{ij}) prior(\beta) \tag{6-22}$$

We choose to use the conjugate prior for β and all other parameters. Conjugate priors are very popular in Bayesian inference because they are considered to be the "natural" choice of the prior distribution given the likelihood function.

For β, the conjugate prior is a multivariate normal distribution:

$$\beta \sim N(\beta_0,\Sigma_0)$$

where β_0 and Σ_0 are given. That is:

$$prior(\beta) = (2\pi)^{-\frac{K}{2}} \mid \Sigma_0 \mid \exp\{-\frac{1}{2}(\beta-\beta_0)\Sigma_0^{-1}(\beta-\beta_0)\}$$

$$\propto \exp\{-\frac{1}{2}(\beta-\beta_0)\Sigma_0^{-1}(\beta-\beta_0)\} \tag{6-23}$$

where K is the length of β and $(2\pi)^{-\frac{K}{2}} \mid \Sigma_0 \mid$ is a known constant.

From Equation 6-21, the likelihood function is:

$$f(Y \mid \beta,S,p_{ij}) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\}$$

$$\propto \exp\{-\frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)\} \tag{6-24}$$

where the term $(2\pi\sigma^2)^{-\frac{T}{2}}$ is treated as a constant because all other parameters except β are considered given.

Substituting Equations 6-23 and 6-24 into Equation 6-22, we get the following posterior distribution of β:

$$f(\beta \mid Y,S,p_{ij}) \propto \exp\{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)-\frac{1}{2}(Y-X\beta)'(Y-X\beta)\}$$

$$\propto \exp\{-\frac{1}{2}(\beta-\beta_1)\Sigma_1^{-1}(\beta-\beta_1)\}$$

where

$$\beta_1 = (\Sigma_0^{-1}+\sigma^{-2}X'X)^{-1}(\Sigma_0^{-1}\beta_0-\sigma^{-2}X'Y)$$

$$\Sigma_1 = (\Sigma_0^{-1}+\sigma^{-2}X'X)^{-1}$$

Clearly, the posterior distribution of β follows a normal distribution:

$$\beta \mid \sigma^2, Y, S, p_{ij} \sim N(\beta_1, \Sigma_1)$$

4.1.2 Deriving $(\sigma^2 \mid \beta, Y, S, p_{ij})$

Again, using the Bayes theory:

$$f(\sigma^2 \mid \beta, Y, S, p_{ij}) \propto f(Y \mid \beta, \sigma^2, Y, S, p_{ij}) prior(\sigma^2) \qquad (6\text{-}25)$$

The conjugate prior of σ^2 is the inverse gamma (IG) distribution:

$$\sigma^2 \sim IG(v_{g_0}, \lambda_{g_0})$$

That is:

$$prior(\sigma^2) = \frac{\lambda_{g_0}^{v_{g_0}}}{\Gamma(v_{g_0})}(1/\sigma^2)^{v_{g_0}+1} \exp\left(-\lambda_{g_0}/\sigma^2\right)$$

$$\propto (1/\sigma^2)^{v_{g_0}+1} \exp\left(-\lambda_{g_0}/\sigma^2\right) \qquad (6\text{-}26)$$

The likelihood function is:

$$f(Y \mid \beta, S, p_{ij}) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\}$$

$$\propto (1/\sigma^2)^{\frac{T}{2}} \exp\{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\}$$

$$\propto (1/\sigma^2)^{\frac{T}{2}} \exp\{-\frac{1}{2\sigma^2}e'e\} \qquad (6\text{-}27)$$

where $e \equiv Y - X\beta$. Note that compared to Equation 6-24, we treat β as a constant vector and arrive at a different likelihood function.

Substituting Equations 6-27 and 6-26 back to Equation 6-25, we have the posterior distribution of σ^2:

$$f(\sigma^2 \mid \beta, Y, S, p_{ij}) \propto (\frac{1}{\sigma^2})^{v_{g_0}+T/2+1} \exp\{\frac{-1}{\sigma^2}(\lambda_{g_0} + \frac{e'e}{2})\}$$

The posterior distribution of σ^2 follows an inverse gamma distribution:

$$\sigma^2 \mid \beta, Y, S, p_{ij} \sim IG(v_g, \lambda_g)$$

$$v_g = v_{g0} + T/2$$

$$\lambda_g = \lambda_{g0} + (e'e)/2$$

4.2 Conditional Posterior Distributions of Transition Probability

We begin our discussion from the likelihood function of S. Since the evolution of S follows a Markov process, we know that the probability of s_t only depends on the value of s_{t-1}. That is:

$$f(S \mid p_{ij}) = \Pi_{t=2}^{T} f(s_t \mid s_{t-1}, p_{ij}) = p_{11}^{n_{11}}(1-p_{11})^{n_{10}}\, p_{00}^{n_{00}}(1-p_{00})^{n_{01}} \qquad (6\text{-}28)$$

where n_{ij} is the count of transitions from state i to j. The count n_{ij} can be calculated directly from S.

The conjugate priors of p_{00} and p_{11} are beta distributions:

$$p_{00} \sim beta(u_{00}, u_{01})$$

$$p_{11} \sim beta(u_{11}, u_{10})$$

where u_{ij}, $i, j \in \{0,1\}$ are constants chosen by domain experts. The densities of the prior distributions are:

$$prior(p_{00}) \propto p_{00}^{u_{00}-1}(1-p_{00})^{u_{01}-1} \qquad (6\text{-}29)$$

$$prior(p_{11}) \propto p_{11}^{u_{11}-1}(1-p_{11})^{u_{10}-1} \qquad (6\text{-}30)$$

Combining the likelihood function (Equation 6-28) and the prior distributions (Equations 6-29 and 6-30), we have the posterior of p_{00} and p_{11}:

$$p_{00} \mid S \sim beta(u_{00} + n_{00}, u_{01} + n_{01})$$

$$p_{11} \mid S \sim beta(u_{11} + n_{11}, u_{10} + n_{10})$$

4.3 Conditional Posterior Distributions of Hidden States

Deriving the conditional posterior distributions of S is slightly more complicated. Our target is $f(S \mid Y, \beta, \sigma^2, p_{ij})$. A simpler but less efficient way is making use of $f(s_t \mid s_{-t}, Y, \beta, \sigma^2, p_{ij})$, where $s_{-t} = (s_1, s_2, ..., s_{t-1}, s_{t+1}, ..., s_T)$. The posterior distribution $f(S \mid Y, \beta, \sigma^2, p_{ij})$ then can be approximated by combining individual $f(s_t \mid s_{-t}, Y, \beta, \sigma^2, p_{ij})$. This method is usually referred to as the single-move Gibbs sampler. The method is less efficient compared to the multi-move Gibbs sampler that we will introduce below due to the dependency among s_t.

The multi-move Gibbs sampler draws S from $f(S \mid Y, \beta, \sigma^2, p_{ij})$ at once and usually leads to faster convergence. In the following discussion, we suppress β , σ^2 , and p_{ij} since they play no roles in the derivation. Also, we use the following notations:

$$Y^t = (y_1, y_2, ..., y_t)$$

$$S^t = (s_1, s_2, ..., s_t)$$

Before getting into the details of the multi-move Gibbs sampler, there are several useful properties that are worth mentioning. First, because of the Markov property of s_t , s_{t+2} contains no information about s_t given s_{t+1} :

$$f(s_t \mid s_{t+1}, s_{t+2}) = \frac{f(s_t, s_{t+1}, s_{t+2})}{f(s_{t+1}, s_{t+2})}$$

$$= \frac{f(s_t, s_{t+2} \mid s_{t+1}) f(s_{t+1})}{f(s_{t+1}, s_{t+2})}$$

$$= \frac{f(s_t \mid s_{t+1}) f(s_{t+2} \mid s_{t+1}, s_t) f(s_{t+1})}{f(s_{t+1}, s_{t+2})}$$

$$= \frac{f(s_t \mid s_{t+1}) f(s_{t+2} \mid s_{t+1}) f(s_{t+1})}{f(s_{t+1}, s_{t+2})}$$

$$= f(s_t \mid s_{t+1})$$

It can be generalized to arbitrary future hidden states:

$$f(s_t \mid s_{t+1}, s_{t+2}, s_{t+3}, ..., s_T) = f(s_t \mid s_{t+1})$$

A similar relation exists for future observations and states:

$$f(s_t \mid s_{t+1}, y_t, y_{t+1}) = \frac{f(s_t, s_{t+1}, y_t, y_{t+1})}{f(s_{t+1}, y_t, y_{t+1})}$$

$$= \frac{f(s_t, y_{t+1} \mid s_{t+1}, y_t) f(s_{t+1}, y_t)}{f(s_{t+1}, y_t, y_{t+1})}$$

$$= \frac{f(s_t \mid s_{t+1}, y_t) f(y_{t+1} \mid s_t, s_{t+1}, y_t) f(s_{t+1}, y_t)}{f(s_{t+1}, y_t, y_{t+1})}$$

$$= \frac{f(s_t \mid s_{t+1}, y_t) f(y_{t+1} \mid s_{t+1}, y_t) f(s_{t+1}, y_t)}{f(s_{t+1}, y_t, y_{t+1})}$$

$$= f(s_t \mid s_{t+1}, y_t)$$

It can be verified that, given s_{t+1}, future observations (y_{t+1}, y_{t+2},…) and future states (s_{t+2}, s_{t+3},…) contain no information about s_t:

$$f(s_t \mid s_{t+1}, s_{t+2}, ..., s_T, y_t, y_{t+1}, ..., y_T) = f(s_t \mid s_{t+1}, y_t)$$

Using the above relations, we have:

$$f(S^T \mid Y^T)$$

$$= f(s_1, s_2, ..., s_T \mid Y^T)$$

$$= f(s_T \mid Y^T) f(s_{T-1} \mid s_T, Y^{T-1}, y_T) f(s_{T-2} \mid s_{T-1}, s_T, Y^{T-2}, y_{T-1}, y_T)...$$

$$f(s_1 \mid s_2, s_3, ..., s_T, y_1, y_2, ..., y_T)$$

$$= f(s_T \mid Y^T) f(s_{T-1} \mid s_T, Y^{T-1}) f(s_{T-2} \mid s_{T-1}, Y^{T-2})...f(s_1 \mid s_2, x_1)$$

$$= f(s_T \mid Y^T) \prod_{t=1}^{T-1} f(s_t \mid s_{t+1}, Y^t) \tag{6-31}$$

Equation 6-31 suggests that we can sample S^T at once by first drawing s_T from $f(s_T \mid Y^T)$, and then s_{T-1} from $f(s_{T-1} \mid s_T, Y^{T-1})$,…, and so on. By Bayes theory:

$$f(s_t \mid Y^t, s_{t+1}) \propto f(s_{t+1} \mid s_t) f(s_t \mid Y^t) \tag{6-32}$$

We only need to derive $f(s_t \mid Y^t)$, $t = 1,2,...,T$ for the multi-move Gibbs sampler. Following Kim (Kim and Nelson, 1999), $f(s_t \mid Y^t)$ can be derived recursively by the following steps:

(0) Given $f(s_{t-1} \mid Y^{t-1})$

(1) One-step ahead prediction of s_t:

$$f(s_t = l \mid Y^{t-1}) = \sum_{k=0}^{1} p_{kl} f(s_{t-1} = l \mid Y^{t-1})$$

(2) Filtering for s_t

$$f(s_t = l \mid Y^t) = \frac{f(y_t \mid s_t = l, Y^{t-1}) f(s_t = l \mid Y^{t-1})}{f(y_t \mid Y^{t-1})}$$

where

$$f(y_t \mid Y^{t-1}) = \sum_{k=0}^{1} f(y_t \mid S_t = k, Y^{t-1}) f(s_t = k \mid y^{t-1})$$

(3) Computing the target $f(s_t \mid s_{t+1}, Y^t)$:

$$f(s_t = l \mid s_{t+1} = k, Y^t) = \frac{p_{lk} f(s_t = l \mid Y^t)}{\sum_{j=0}^{1} p_{jk} f(s_t = j \mid Y^t)}$$

At time $t = 1$, the iteration start with $f(s_0 \mid Y^0) = f(s_0 \mid p_{00}, p_{11})$, the unconditional probability of the state given current parameters. A reasonable choice is the steady-state probabilities:

$$\pi_0 = f(s_0 = 0 \mid p_{00}, p_{11}) = \frac{1 - p_{11}}{2 - p_{11} - p_{00}}$$

$$\pi_1 = f(s_0 = 1 \mid p_{00}, p_{11}) = \frac{1 - p_{00}}{2 - p_{11} - p_{00}}$$

A more general discussion of the steady-state probabilities involving more than two states can be found elsewhere (Frihwirth-Schnatter, 2006).

4.4 Estimating Markov Switching Models via the Gibbs Sampler

As mentioned at the beginning of this section, our goal is to construct the joint posterior distribution $f(\beta, \sigma^2, S, p_{ij} \mid Y)$. Following the general procedure of MCMC, the joint posterior distribution can be constructed using the following conditional posterior distributions:

$$f(\beta \mid Y, \sigma^2, S, p_{ij}) = f(\beta \mid Y, \sigma^2, S)$$

$$f(\sigma^2 \mid Y, \beta, S, p_{ij}) = f(\sigma^2 \mid Y, \beta, S)$$

$$f(p_{00} \mid Y, \beta, \sigma^2, p_{11}, S) = f(p_{00} \mid S)$$

$$f(p_{11} \mid Y, \beta, \sigma^2, p_{00}, S) = f(p_{11} \mid S)$$

$$f(S \mid Y, \beta, \sigma^2, p_{ij}) = f(S \mid Y, p_{ij})$$

The functions on the left hand side are the conditional distributions required by MCMC. The functions on the right hand side are the conditional distributions we are actually dealing with considering the special dependent structure of the Markov switching model. For example, the conditional posterior of the hidden state S does not depend on β or σ^2 given Y and p_{ij}.

Also, we use conjugate prior distributions to compute the posterior distributions listed above:

$$\beta \sim N(\beta_0, \Sigma_0)$$

$$\sigma^2 \sim IG(v_{g_0}, \lambda_{g_0})$$

$$p_{00} \sim beta(u_{00}, u_{01})$$

$$p_{11} \sim beta(u_{11}, u_{10})$$

A general rule to determine parameters for the prior distributions is that we want to provide as little information as possible. As a result, Σ_0 should have large diagonal elements; v_{g_0} should be small and λ_{g_0} should be large. However, we also want to have the posterior distributions with finite means and variances so that the estimation results have meaningful interpretation. This imposes additional constraints on choosing the parameters for the prior distributions. A simple rule is to choose the parameters to ensure that the prior distributions have finite second moments. One may argue that the constraints may be too strong in the sense that the posterior distribution may still have a finite second moment even when the prior distribution does not. In practice, we find that imposing the additional constraint has little effect in restraining the estimation results but does ensure that the confidence intervals computed from the estimation results are meaningful.

To perform estimation using MCMC, a set of initial values are chosen: $(\beta^{(0)}, \sigma^{2(0)}, p_{00}^{(0)}, p_{11}^{(0)}, S^{(0)})$. For iteration i, we repeat the following steps:

1. Draw $\beta^{(i)}$ from $f(\beta \mid Y, S^{(i-1)}, \sigma^{2(i-1)})$.
2. Draw $\sigma^{2(i)}$ from $f(\sigma^2 \mid Y, S^{(i-1)}, \beta^{(i)})$.
3. Draw $S^{(i)}$ from $f(S \mid Y, p_{00}^{(i-1)}, p_{11}^{(i-1)})$.
4. Draw $p_{00}^{(i)}$ from $f(p_{00} \mid S^{(i)})$.
5. Draw $p_{11}^{(i)}$ from $f(p_{11} \mid S^{(i)})$.
6. Record $(\beta^{(i)}, \sigma^{2(i)}, p_{00}^{(i)}, p_{11}^{(i)}, S^{(i)})$.

The process is repeated for I times, where I is a predefined number. The number of iteration I can be determined iteratively by checking whether the Markov chain has converged. Some useful discussion can be found elsewhere (Cowles and Carlin, 1996). To ensure that the results are not influenced by the initial values, the initial B iterations are discarded before computing the statistics. The posterior means of parameters are the average of the sample generated by the MCMC method. For example, the posterior mean of σ^2 is:

$$\hat{\sigma}^2 = \sum_{i=B+1}^{I} \frac{\sigma^{2(i)}}{I - B + 1}$$

The confidence intervals can also be computed directly from the corresponding percentile of $\{\sigma^{2(i)}\}$.

5. CASE STUDY

The BioPortal Outbreak Detection System (BiODS) detects outbreaks in time series based on the Markov switching models. The BiODS is one of the subsystems of the BioPortal project (Zeng et al., 2005), which is aimed at developing an integrated, cross-jurisdiction infectious disease information infrastructure.

To demonstrate the application of the Markov switching models using a real-world dataset, we obtained a collection of chief complaints (CCs) for the time period June 30, 2000 to April 27, 2003 from a hospital in Taiwan. There were 368,151 CCs in our dataset. About a quarter of them contained Chinese characters. The BioPortal Multilingual Chief Complaint classifier (Lu et al., 2007, 2008) was used to classify CCs into nine syndromic categories: Botulism-Like, Constitutional, Gastrointestinal (GI), Hemorrhagic, Neurological, Rash, Respiratory (RESP), Fever, and Other. The daily count of each syndrome then was analyzed using the Markov switching model. We present the estimation results of the GI syndrome below.

Figure 6-3 plots the time series of the GI syndrome. It seems that there is no clear seasonal patterns. However, sporadic spikes are clearly visible. From the autocorrelation function shown in Figure 6-4, we can observe a very strong day-of-week effect as the autocorrelation coefficient peaks for lags that are multiples of 7. The other weaker pattern is a cyclical pattern of about 110 days. For the sake of clarify, we chose to incorporate the day-of-week effect and ignore other autocorrelation patterns. The day-of-week effect was modeled using six dummy variables for Monday, Wednesday, Thursday, Friday, Saturday, and Sunday. The coefficients of these dummy variables were directly related to the average differences between a particular day-of-week and a Tuesday. Our Markov switching model was:

$$y_t = a_{0,0} + a_{0,1}s_t + (a_{1,0} + a_{1,1}s_t)y_{t-1} + \sum_{i=1}^{6} w_i d_i + e_t \qquad (6\text{-}33)$$

$$s_t \in \{0,1\} \qquad (6\text{-}34)$$

$$P(s_t = j \mid s_{t-1} = i) = p_{ij} \tag{6-35}$$

$$e_t \sim N(0, \sigma^2) \tag{6-36}$$

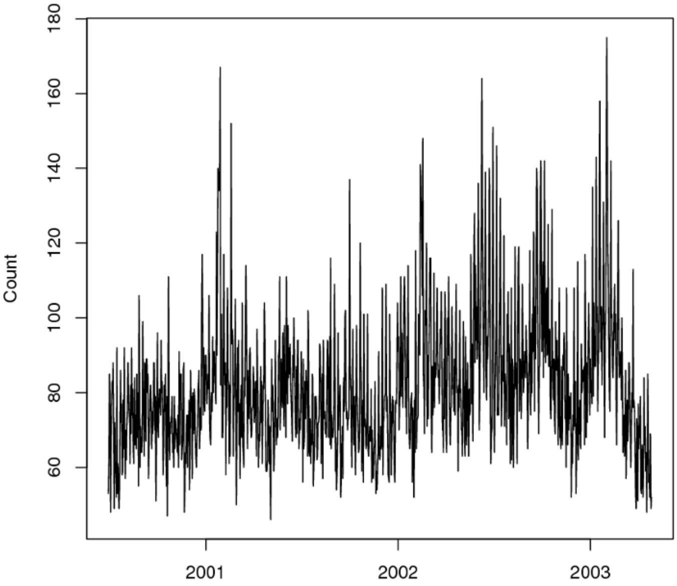


Figure 6-3. Gastrointestinal syndrome count time series.

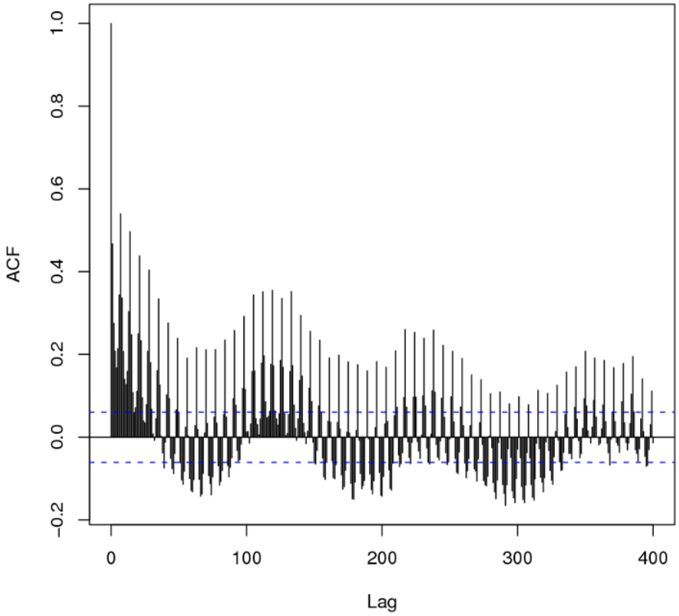


Figure 6-4. Autocorrelation function of the gastrointestinal syndrome count time series.

The derivation procedure for the posterior distributions remained unchanged. The additional dummy variables were included as additional independent variables and the same procedure applied. The posterior distribution of transition probability remained unchanged. The conditional posterior distribution of hidden state variables needed minor modifications.

We used Gibbs sampler to estimate the Markov switching model. The parameters and confidence intervals were computed from the last 2,000 of 2,500 iterations. Figure 6-5 plots the autocorrelation function of the residuals. Most of the short-term autocorrelations were removed. It is not surprising to see that the long-term autocorrelation pattern of about 110 days still exists because we chose not to handle this pattern. The autocorrelation function is considered "clean" since most lags fall within the 95% band.

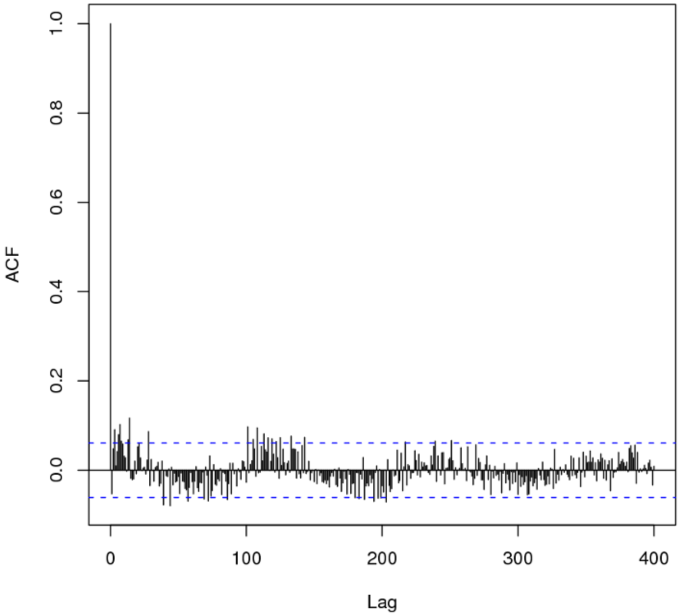


Figure 6-5. Autocorrelation function of the residual.

Table 6-3 summarizes the parameter estimation results. Both posterior means and 95% confidence intervals (CI) are listed. The constant term a_{00} was 39.77. It increased to 45.34 ($a_{00} + a_{11}$) when an outbreak occurred. The coefficient of the lagged term y_{t-1} was 0.4 when there was no outbreaks. The change of the coefficient was small (0.05) when outbreaks occurred. These estimated coefficients (a_{ij}) corresponded to the long-term mean of non-outbreak and outbreak states, which are reported in the last two rows. The long-term mean of the non-outbreak state was 65.69, while the long-term mean of the outbreak state was 82.35. Note that these values should be interpreted as the count of an "average Tuesday." For the average value of another day-of-week, one can compute the value by adding the contribution of w_i. Note that similar to the computation of unconditional mean (Equation 6-5), the contribution of the dummy needs to be adjusted using the coefficient of y_{t-1}. For example, the contribution of w_1 (Monday) to the long-term mean during the non-outbreak period is $w_1/(1-a_{1,0})$. As a result, on a non-outbreak Sunday, the average count was $65.69+27.16/(1-0.4)=110.96$. The transition probabilities p_{00} and p_{11} were 0.993 and 0.988, respectively.

Table 6-3. Estimation results of the Markov switching model.

Coefficient	Mean (95% CI)
a_{00}	39.77 (33.65, 46.13)
a_{01}	5.57 (–1.09, 12.74)
a_{10}	0.40 (0.31, 0.47)
a_{11}	0.05 (–0.03, 0.14)
w_1 (Mon)	–1.70 (–4.55, 1.11)
w_2 (Wed)	1.77 (–0.69, 4.23)
w_3 (Thu)	1.86 (–0.58, 4.24)
w_4 (Fri)	4.00 (1.50, 6.44)
w_5 (Sat)	7.08 (4.65, 9.50)
w_6 (Sun)	27.16 (24.55, 29.62)
σ^2	153.74 (137.00, 170.06)
p_{00}	0.993 (0.987, 0.997)
p_{11}	0.988 (0.976, 0.996)
Non-outbreak average	65.69 (62.16, 68.89)
Outbreak average	82.35 (78.53, 86.36)

Figure 6-6 plots the GI time series and the estimated outbreak states. The lower panel plots $f(S\mid Y)$, the posterior outbreak probability given the observed time series. The posterior outbreak probability at each period is constrained between 0 and 1. As a result, the interpretation is more intuitive compared to other outbreak detection methods such as CUSUM and EWMA.

The Markov switching model identified three major outbreak periods. The first period was during the end of 2000 and the beginning of 2001. The second outbreak period was between the first and third quarters of 2002. The third outbreak period was at the beginning of 2003. The outbreak periods were identified solely based on the specification of the model. As a result, the identified outbreak periods can only be interpreted as having higher time series values compared to the rest of the time series. The reason behind this needs to be further investigated. The state estimation results can guide the investigation efforts so that additional evidence can be collected to explain the observed time series dynamics.

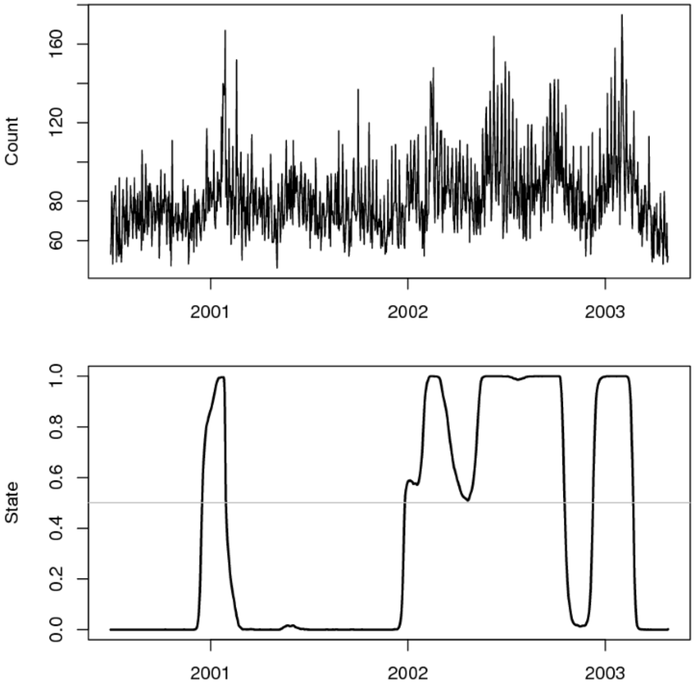


Figure 6-6. Estimated outbreak states.

For researchers working on developing novel outbreak detection algorithms, the estimation results can provide valuable information for their gold standard development. For practitioners using existing outbreak detection methods, the estimation results can guide them to choose proper training periods. For example, based on the estimation results, the first half of the time series is a better candidate then the second half. It may be worth the effort to investigate whether the outbreak identified during the end of 2000 to the beginning of 2001 is related to disease outbreaks before the first half of the time series is used for model training.

ACKNOWLEDGMENTS

This work is supported in part by the U.S. National Science Foundation through Grant #IIS-0428241. We would like to thank Dr. Chwan-Chuen King and Dr. Fuh-Yuan Shih from the National Taiwan University for providing research datasets.

QUESTIONS FOR DISCUSSION

1. What are the advantages of the Markov switching model compared to traditional statistical surveillance methods? What are the disadvantages? Discuss.
2. What approaches can be used to estimate the Markov switching models? What are their advantages and disadvantages?
3. What is a conjugate prior? Provide examples of conjugate priors in standard linear regression models and the Markov switching models.

REFERENCES

Albert, J.H., Chib, S., 1993. Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. Journal of Business & Economic Statistics 11, 1–15.

Baum, L.E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematics and Statistics 37, 1554–1563.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) 36, 192–236.

Carter, C.K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553.

Chapman, W.W., Christensen, L.M., Wagner, M.M., Haug, P.J., Ivanov, O., Dowling, J.N., Olszewski, R.T., 2005a. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artificial Intelligence in Medicine 33, 31–40.

Chapman, W.W., Dowling, J.N., Wagner, M.M., 2005b. Generating a reliable reference standard set for syndromic case classification. Journal of the American Medical Informatics Association 12, 618–629.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician 49, 327–335.

Clements, M.P., Hendry, D.F., 2006. Forecasting with breaks In: Elliot G, Granger CWJ & Timmermann A (eds.) Handbook of Economic Forecasting. Elsevier, Amsterdam, pp. 605–657.

Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91, 883–904.

Dahlquist, M., Gray, S.F., 2000. Regime-switching and interest rates in the European monetary system. Journal of International Economics 50, 399–419.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

Espino, J.U., Wagner, M., 2001. The accuracy of ICD-9 coded chief complaints for detection of acute respiratory illness. In: Proceedings of the AMIA Annual Symposium, pp. 164–168.

Frihwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, New York.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357–384.

Harvey, A.C., 1989. Forecasting, Structural Time Series Models and Kalman Filter. Cambridge University Press, Cambridge.

Ivanov, O., Wagner, M.M., Chapman, W.W., Olszewski, R.T., 2002. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. In: Proceedings of the AMIA Symposium, pp. 345–349.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Transactions ASME Journal of Basic Engineering D82, 35–45.

Kim, C.-J., Nelson, C.R., 1999. State-Space Models with Regime Switching. MIT Press, Cambridge, MA.

Lu, H.-M., King, C.-C., Wu, T.-S., Shih, F.-Y., Hsiao, J.-Y., Zeng, D., Chen, H., 2007. Chinese chief complaint classification for syndromic surveillance. In: Intelligence and Security Informatics: Biosurveillance. Springer, New Brunswick, NJ.

Lu, H.-M., Zeng, D., Trujillo, L., Komatsu, K., Chen, H., 2008. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. Journal of Biomedical Informatics 41, 340–356.

Montgomery, D.C., 2005. Introduction to Statistical Quality Control. Wiley, New York.

Page, E.S., 1954. Continuous inspection schemes. Biometrika 41, 100–115.

Reis, B.Y., Mandl, K.D., 2003. Time series modeling for syndromic surveillance. BMC Medical Informatics and Decision Making 3, 2.

Reis, B.Y., Pagano, M., Mandl, K.D., 2003. Using temporal context to improve biosurveillance. Proceedings of the National Academy of Sciences of the United States of America 100, 1961–1965.

Shewhart, W.A., 1939. Statistical Method from the Viewpoint of Quality Control. The Graduate School, The Department of Agriculture, Washington.

Zeng, D., Chen, H., Tseng, C., Larson, C., Chang, W., Eidson, M., Gotham, I., Lynch, C., Ascher, M., 2005. BioPortal: Sharing and Analyzing Infectious Disease Information. Springer, Berlin.

SUGGESTED READING

Kim, C.-J., Nelson, C.R., 1999. State-Space Models with Regime Switching. MIT Press, Cambridge, MA. 297 pages.

This book covers important topics of state-space models in general and Markov switching models in specific. Both classical estimation methods and Gibbs sampler are discussed in detail. The authors also provide sample programs that implement the algorithms discussed in the book.

ONLINE RESOURCES

The R Project provides a cross-platform computational environment that is suitable to implement the Markov switching models. The project website can be found at http://www.r-project.org/

The BioPortal project's homepage is at http://bioportal.eller.arizona.edu

Chapter 7

DETECTION OF EVENTS IN MULTIPLE STREAMS OF SURVEILLANCE DATA

Multivariate, Multi-stream and Multi-dimensional Approaches

ARTUR DUBRAWSKI*

CHAPTER OVERVIEW

Simultaneous monitoring of multiple streams of data that carry corroborating evidence can be beneficial in many event detection applications. This chapter reviews analytic approaches that can be employed in such scenarios. We cover established statistical algorithms of multivariate time series baseline estimation and forecasting. They are relevant when multiple streams of data can be modeled jointly. We then present more recent methods which do not have to rely on such an assumption. We separately address techniques that deal with data in a specific form of a record of transactions annotated with multiple descriptors, often encountered in the practice of health surveillance. Future event detection algorithms will benefit from incorporation of machine learning methodology. This will enable adaptability, utilization of human feedback, and building reliable detectors using some examples of events of interest. That will lead to highly scalable and economical multi-stream event detection systems.

Keywords: Event detection; Analysis of multivariate time series; Health surveillance

* *Auton Lab, Carnegie Mellon University, Newell Simon Hall 3128, 5000 Forbes Ave, Pittsburgh PA 15213-3815, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_7,

1. INTRODUCTION

Simultaneous monitoring of signals coming from distinct sources, or surveying diverse aspects of data even if it comes from a single source, can yield improvements in accuracy, sensitivity, specificity and timeliness of event detection over more traditional univariate analyses. Suppose for instance that today's sales of cough and cold medications are insignificantly higher than normal, absenteeism among school children in the district is up by one and a half standard deviation estimated from normal historical data, and a slightly higher than usual number of respiratory patients report to the region's emergency rooms. None of these individual signals are substantial enough to cause an alert on their own, but when looked at in conjunction, they may raise concern. Or suppose that in monitoring sales of non-prescription medicines, a particular region reports an unexpected increase in sales of pain relief and anti-fever drugs. However, it appears to be aligned with an unexpected rise of sales of non-medication groceries. This may indicate that the population of the region has increased (e.g., due to a sunny winter weekend in a mountain town). That in turn may sufficiently explain the initial observation about pain killers and dismiss the concern of a possible outbreak of infectious disease based on the single-stream analysis. Or, suppose that among the chief complaints of two cases from two different hospitals in the same city on the same date there was mention of bloody stools in pediatric patients. The multiple mentions of "bloody stools" or "pediatric" is not a surprise but the tying together of these factors is sufficiently rare that seeing even just a few such cases is of interest. This is precisely the evidence (that was spotted manually, not by automated surveillance) that was the first warning of the infamous Walkerton, Canada, outbreak of waterborne gastroenteritis in May 2000 (CMAJ, 2000).

There has been a great deal of work in recent years on univariate event detection algorithms and on their applications to biosurveillance. The methods proposed and tested include control charts, regression methods derived from Serfling's original work on seasonality modeling, wavelet approaches, moving average and its many variants (including exponentially weighted moving average), and the very popular cumulative sum (CUSUM) algorithm, to name just a few. The wide variety of methods reflects the wide variety of issues that occur in the variety of data sources that have been used. Issues that are faced in univariate time series analysis also affect multivariate approaches. New challenges, such as increased complexity of data and models, as well as new opportunities, such as the availability of independent sources of corroborating evidence, make multi-stream surveillance an attractive domain for ongoing research and applications. Researchers in the field of public health and biodefense have already attempted to exploit the benefits of simultaneous tracking of multiple sources of complementary evidence and multiple facets of the accessible data. However, this domain, being comparatively less explored, may also be relatively new to many practitioners of biosafety. This chapter aims at reviewing fundamental methods of multi-stream surveillance and it presents a few representative examples of its use.

Note that, in the context of multi-stream surveillance, the notion of multivariability may in practice imply one or more of the following meanings:

- Combination of data sourced from multiple distinct streams (for example, emergency department visits and over-the-counter (OTC) drug sales; or animal health examination records vs. microbial test results of food samples taken at slaughter houses vs. complaints submitted by food consumers). This is what we call the multi-stream data.
- Use of data records, perhaps originating from a single source, with multiple attribute fields (dimensions) such as spatial location and patient attributes such as age, gender, and symptoms. This data will be called multi-dimensional.
- Consideration and evaluation of multiple possible causes for an observed bioevent. This scenario involves data with multi-dimensional outcomes, sometimes also called multi-focus data.

Surveillance of multiple streams of data can often be tackled using the well known methods of multivariate time series analysis. This may make very good sense if the data from distinct streams is being supplied synchronously and if the probabilistic characteristics of the processes generating the data are coherent across all streams. In those cases, it may be appropriate to treat such data as a single multivariate stream coming from a joint source. This assumption allows for building detectors of anomalies based on multivariate statistical models involving all component streams. These models can exploit correlations between observations recorded in different streams. This is the topic of the next section in which we focus on methods of multivariate data analysis. First, we quickly review popular approaches to modeling and forecasting multivariate baselines, and then we move on to review approaches to detecting significant departures of the multivariate data from their expected settings. In other cases, it may be more practical to simply set up separate detectors for each individual stream of data (note that, in general, each of them may be multivariate on their own accord) and to aggregate their indications in a separate procedure. That topic is discussed in the subsequent section. We also dedicate a section to cover the methods specifically designed to deal with multi-dimensional data due to the abundance of such data sources in the practice of biosurveillance.

2. MULTIVARIATE ANALYSIS

A typical approach to anomaly detection in temporal domains is to first forecast the expected values of the monitored variables, and then to compare the actual current observations against their expected values, triggering an alert whenever the discrepancy is sufficiently large. Reliability of this process heavily relies on the quality of the forecast of the baseline. That may be a tough task. Its result depends on the level of understanding of the processes which generate the observations (which impacts the attainable accuracy of their models), as well as on the accuracy, relevance and completeness of the available data.

In many event detection approaches, the phases of baseline forecasting and scoring of anomalousness are integrated, but for clarity of presentation we first focus on the first stage of the process: baseline modeling and forecasting.

2.1 Modeling and Forecasting of Multivariate Baselines

If the observed time series are stationary and their mutual dependencies do not change over time (ideally, if they are statistically independent), one may consider using multiple regression. In its simplest form, multiple regression produces a linear model of relationships between *m* independent variables (observations, e.g., records of daily sales of *m* different types of non-prescription medicines), x_i, and one dependent variable (outcome, e.g., percentage of the local population affected by flu-like symptoms), y:

$$\hat{y} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_m \cdot x_m$$

Here, $\hat{y}$ denotes the predicted value of the independent variable y, corresponding to the specific set of independent observations $\{x_1, x_2, \ldots, x_m\}$, and β_i are the model parameters estimated from the available reference data (usually using the least squares method). If the above conditions are met, and if we have training data which accurately represents historical correlations between the observations and outcomes, we can build and use the above model to predict the outcomes based on actual observations.

Multiple regression is rather naïve in the context of surveillance in that it ignores a likely possibility of the observations to depend on time. This caveat is targeted by the autoregressive (AR) models which represent the *m*-dimensional vector $\mathbf{X}_t$ of current observations as a linear combination of observations made at *k* preceding time steps, $\mathbf{X}_{t-i}$:

$$\mathbf{X}_t = \mathbf{A}_1\mathbf{X}_{t-1} + \mathbf{A}_2\mathbf{X}_{t-2} + \ldots + \mathbf{A}_k\mathbf{X}_{t-k} + \mathbf{E}_t$$

Here, $\mathbf{A}_j$ are $m \times m$ matrices of autocorrelation coefficients estimated from training data, while $\mathbf{E}_t$ represents a vector of Gaussian noise. Multivariate autoregression can be directly applied to forecast expected values of the modeled time series based on the pattern of their past behavior. It can capture linear dependencies between the individual variables and their changes over the specific time horizon of analysis (*k* steps back), as well as linear relationships between the current and past observations in the series.

The AR models are vulnerable to noise in data. That can be to some extent alleviated by adding smoothing components to the regression equation, obtained, for instance, by implementing a moving average (MA) procedure. The resulting multivariate ARMA model can then be represented by the equation below. $\mathbf{B}_j$ and $\mathbf{E}_{t-j}$ denote respectively the $m \times m$ matrix of moving average coefficients and the *m*-dimensional vector of residuals observed at the *j*th time step before *t*. Typically, the orders *k* and *l* as well as the components of matrices $\mathbf{A}_j$ and $\mathbf{B}_j$ are determined using the Box–Jenkins approach (Box et al., 1994):

$$\mathbf{X}_t = \mathbf{A}_1\mathbf{X}_{t-1} + \mathbf{A}_2\mathbf{X}_{t-2} + \ldots + \mathbf{A}_k\mathbf{X}_{t-k} + \mathbf{E}_t - \mathbf{B}_1\mathbf{E}_{t-1} - \mathbf{B}_2\mathbf{E}_{t-2} \ldots - \mathbf{B}_l\mathbf{E}_{t-l}$$

Note that ARMA models rely on stationarity of the modeled time series. If this assumption is not met, the standard approach is to differenciate the non-stationary series. The result is the integrated ARMA, also known as the ARIMA model.

Another way of predicting multivariate time series is provided by an extension of the popular general purpose forecasting method: Exponentially Weighted Moving Average (EWMA) (Lowry et al., 1992). The multivariate EWMA (MEWMA) equation can be written as follows:

$$\mathbf{Z}_t = \mathbf{\Lambda}\mathbf{X}_t + (\mathbf{I} - \mathbf{\Lambda})\mathbf{Z}_{t-1}$$

Here, $\mathbf{X}_t$ denotes the vector of current observations, $\mathbf{Z}_t$ is the vector of current forecast and $\mathbf{Z}_{t-1}$ is the vector of the previous step forecast (all these vectors are *m*-dimensional). $\mathbf{\Lambda}$ is a $m \times m$ diagonal matrix of smoothing factors, and $\mathbf{I}$ represents the identity matrix of the same dimensions.

Traditional multivariate forecasting methods considered so far have been already widely used in the context of public health and biodefense applications (reviews can be found in Sonneson and Frisen, 2005 and Rolka et al., 2007). These popular models do not explicitly account for seasonality. Instead, either seasonal effects are filtered out from the raw data before modeling, or the regression equations are complemented with components representing factors such as day of the week or season of the year. Holt-Winter's method, commonly used in a variety of univariate forecasting applications, which separately identifies level, trend and seasonal components of time series, and whose multivariate extensions exist, has not made it, as of yet, to the mainstream of biosurveillance applications. This also seems to be the case of the forecasting methods based on the wavelet transform, which are well designed to account for non-stationarity of the modeled series (Lotze et al., 2006).

Many biosurveillance datasets consist of time series of counts. In such cases the standard Gaussian assumption, typically made in the context of basic multivariate regression, may not be appropriate. Held et al. (2005) present a regression model that works with Poisson and negative binomial observation models and its extension to multivariate disease surveillance.

If the observed component time series are correlated, and if their temporal evolution can be explained by the dynamics of an underlying low-dimensional process, it is possible to apply the concept of Linear Dynamical Systems (LDS, also known as state-space models, or Kalman filters). Siddiqi et al. (2007) proposed an application of LDS to forecast baselines of multivariate biosurveillance data. Their method uses a learned from historical data low-dimensional latent space model to represent dynamics of the set of more than 20-dimensional series of daily counts of sales of distinct categories of medications. If the series are at least partially correlated, and if the latent model evolves linearly under Markov assumptions over a sequence of discrete time steps, it is possible to reliably generate a stable and long sequence of the expected future multivariate counts.

Bayesian Networks (BN) provide a compact way to model probabilistic relationships among multiple variables of data. They became popular in the 1990s as tools to support medical and equipment fault diagnoses, but since then they gained universal acclaim in a range of application domains. They also have found uses in public health and biodefense contexts, especially – but not exclusively – in setups where the involved data is categorical. One of their typical uses is to learn the historical baseline interactions among variables. A trained BN can then be queried for the likelihood of the submitted vector of observations, providing grounds for scoring anomalies. It could also be used to generate synthetic data representing null distribution in detection tasks (Wong et al., 2005b). Other usage models involve inference about the state of the monitored population (Cooper et al., 2004), or testing hypotheses about common causes of the observed events (Dubrawski et al., 2006).

2.2 Detection of Events in Multivariate Time Series

As soon as the characteristics of the reference data have been reliably captured by the baseline model, reliable predictions of future observations can be made as long as the data generating processes are stationary. These predictions can then be compared against the actual observations, and the extent of discrepancy can be used to generate alerts.

Statistical Process Control (SPC) is a popular methodology of monitoring stability of processes over time (Montgomery, 2000). SPC uses control charts to determine whether a possible departure of the monitored process from its normal setting warrants an alert. The simplest (and most common) multivariate version of a control chart that takes advantage of possible correlations between involved variables is the Hotelling method (Hotelling, 1947). It learns from historical data the joint distribution of a set of signals. For example, when monitoring cough sales and nasal spray sales in a region's pharmacies, that distribution could be modeled as a two-dimensional Gaussian with a mean daily cough sale count, mean nasal spray count and covariance matrix derived from data over the previous year. If any given day has counts that fall outside, say, the 99% confidence ellipse of the covariance matrix, an alarm is sounded. This could happen for three reasons: (1) cough sales are surprisingly low or high; (2) nasal sales are surprisingly low or high; or (3) neither set of sales are abnormal by themselves, but the ratio of sales is abnormal. Hotelling's T^2 statistic measures the distance of the vector of observations from its expectation:

$$T^2 = n(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1} (\overline{\mathbf{X}} - \boldsymbol{\mu}_0)$$

Here, n denotes the number of current observations under consideration, $\overline{\mathbf{X}}$ is the m-dimensional vector of means computed for the m monitored time series over the n observations, $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix estimated (using the same n observations), $\boldsymbol{\mu}_0$ is the $m \times m$ matrix of m vectors of time series means computed using the historical baseline data, and the symbol T denotes transposition. Note than in the formula above, the covariance matrix is based on the observed sample. In practice, especially if the size of the sample of current observations is small, it is often replaced by the covariance derived from the historical baseline. Alternatively, to account for non-stationarity, it may be estimated from the batch of multivariate forecasts produced using one of the methods described before. In either case, the constant term n may need to be adjusted to reflect the actual size of the set of observations or forecasts used to estimate the covariance. Under null hypothesis of no difference between $\overline{\mathbf{X}}$ and $\boldsymbol{\mu}_0$, T^2 is distributed as:

$$T^2 \sim \frac{m(n-1)}{n-m} F_{(m,n-m)}$$

where $F_{(m,n-m)}$ represents Fisher's F distribution with m degrees of freedom for the numerator and $n-m$ degrees of freedom for the denominator. Whenever a new set of n m-variate observations arrives, T^2 can be computed and compared against the critical T^2 for the suitably selected threshold of sensitivity α in order to determine whether an alert is in order.

Hotelling's approach has been successfully applied in multivariate biosurveillance, including the ESSENCE framework (Burkom et al., 2004, 2005). Its appeal is primarily derived from its simplicity; however there are a few inconveniences involved. Unlike in the univariate case, the scale of the tracked statistic may be unrelated to the scales of any of the monitored variables. Also, when an alert is raised, it may not be immediately clear which variables have prompted it. A partial remedy is to run univariate control charts in parallel to the multivariate chart, so that the individual culprits can be identified. But that approach cannot explain alerts generated due to the joint contribution of individually sub-critical observations. Hotelling's method has been found to perform poorly at distinguishing location shifts (changes in mean of the joint distribution) from scale shifts (changes of the structure of correlations between variables). Some of those deficiencies can be addressed by using other control charts, such as multivariate CUSUM (Crosier, 1988; Pignatiello and Runger, 1990; Fricker, 2007) which can detect effects over multiple time scales. Another possibility is to preprocess data using, e.g., Principal Component Analysis (PCA) in order to remove correlations between variables and execute a set of univariate tests on a few individual, mutually independent principal components (Mohtashemi et al., 2007).

In general, anomaly detection motivated by the SPC line of thought does not consider specific signatures of the events of interest. Therefore, in general, it cannot be effectively used to distinguish between potential causes of a detected outbreak (i.e., anthrax vs. influenza) or to explain deviations caused by non-outbreak factors such as fluctuations in population or availability. In other words, it can successfully be used to answer the important question "Is there anything unexpected going on today?" but it falls short of explaining what may be causing the observed effects. For those reasons, multivariate control charts can be labeled as non-specific detectors.

The Bayesian Network (BN) framework is especially convenient for developing specific detectors in multivariate scenarios. Arguably one of the largest ever evaluated Bayesian Networks is the core of the Population-wide Anomaly Detection and Assessment (PANDA) system introduced by Cooper et al. (2004). Every person residing in a region is modeled with about 20 nodes, indicating whether or not they are infected by anthrax, what observed symptoms they have, and whether they present themselves at an Emergency Department (ED). Given a city of a million people, there is thus a network with 20 million nodes. Whether or not a person is infected with anthrax is related to whether there has been a recent anthrax attack on the city, and whether the attack was in the same postal code area as the person's location. The diagnosis problem is to take the very few observed nodes of the city on any given day (the symptoms, demographics and home zip codes of patients who actually reported to an ED) and infer the probability distribution over whether there was an attack and, if so, which area was affected. By partitioning city residents into a set of equivalence classes, and iterating over all possible attack zip codes and times, inference was made tractable and specific detections possible. Further development of PANDA led to enabling multi-stream monitoring and detection using aggregated regional counts for OTC drug sales and multivariate records of ED visits for individual patients (Wong et al., 2005a). More recently, the method was extended to concurrently model multiple diseases which belong to the CDC Category A. The resulting PANDA+ system takes as input a time series of emergency department chief complaints, and it produces the posterior probability of each CDC Category A disease and several additional diseases, namely, influenza, crytosporidiosis, hepatitis A and asthma (Cooper et al., 2006). Shen and Cooper (2007) introduced a Bayesian approach for detecting both specific and non-disease-specific outbreaks.

Another example of a successful application of the Bayesian approach to event detection in multivariate surveillance data is the Emerging Patterns in Food Complaints (EPFC) system (Dubrawski et al., 2006). It forms the analytic core of the Consumer Complaint Monitoring System II which helps the U.S. Department of Agriculture (USDA) monitor incoming reports on adverse effects of USDA-regulated food products on their consumers. These reports contain multi-dimensional, heterogeneous and sparse snippets of specific information about consumer demographics, the kinds, brands and sources of the food involved, symptoms of possible sickness, characteristics of foreign objects which could have been found in food, locations and times of occurrences, etc. The system uses probabilistic models of food safety problem scenarios, which are partly derived from small amounts of the available data, and partly crafted by hand under a close consult from domain experts. EPFC estimates how likely it is for a newly reported complaint case to be a close copy of some other case in the past data, if both have been generated by the same specific underlying cause, such as, for instance, malicious contamination of raw food at a plant. Each new case is evaluated against a range of potentially relevant past cases and against a number of predefined plausible food safety failure scenarios. The top matches are reported to human analysts for further investigation. A unique feature of EPFC is the ability to remain sensitive to signals supported by very little data – significant alerts can be raised on the basis of a very few complaints from consumers, provided that the few complaints are consequences of significantly similar and explicable root causes.

Bayesian Networks belong to the broader Machine Learning family of Graphical Models. Another prominent part of that family is made of Hidden Markov Models (HMMs). The key assumption leading to the use of HMMs is that complete knowledge of the current state of the world is not directly observable. Therefore, HMMs attempt to model how the measurable signals are affected by presence or absence of the considered diseases (states) and the environmental conditions (in that, they are very similar to Bayesian Networks). In addition, HMMs try to capture how the state of the world changes over time. The objective is to estimate which of the multiple discrete states best explains a set of recent observations (that is, for example, whether the particular disease outbreak is ongoing, in what phase it is, or whether there are no identifiable outbreaks going on). The model of temporal variability is plausibly simplified with the Markov assumption: the current state does not depend on any other previous states than the one immediately preceding it. This key statement makes the underlying mathematics tractable, while not limiting the utility of the attainable models in a range of practical situations. HMMs yield themselves naturally to multivariate observation scenarios with multi-dimensional outcomes (multiple states), where specificity of detection is required. Typically, the structure of an HMM is dictated by the formulation of the modeling problem. Its parameters can be either estimated from historical data, or extracted from human expertise, or they could result from a combined approach. Madigan (2005) provides a very good overview of HMMs approaches to biosurveillance. In addition, it points out the utility of these models in handling asynchronous observations arising in scenarios where data points taken from different sources come at various frequencies or at independent and random paces. This is a very useful feature, given that the standard regression models described above are not designed to handle such cases well.

3. MULTI-STREAM ANALYSIS

If a strong model of informative relationships between the multiple data streams is available, and if the statistical characteristics of these streams are coherent, then building joint multivariate models is the strategy of choice. In practice, the level of understanding of inter-stream relationships, as well the amount of the available data needed to build reliable multivariate models, may be too limited for the standard approach.

If the streams can be treated as independent of each other, they may be fitted with individual detectors and the system would raise an alert whenever any of the streams gets out of control. Such a parallel approach requires special attention to multiple testing phenomena as the probability of at least one stream causing an alert quickly increases with the number of independently monitored streams. For m streams and expected per-stream false alarm rate α, this probability equals $1 - (1 - \alpha)^m$. A typical approach to trim the risk of false alerts due to multiple testing is to decrease sensitivity of the individual detectors. Often, a very conservative Bonferroni correction or a more balanced False Discovery Rate method is used (Wasserman, 2004). Reduction of sensitivity decreases the number of alerts generated by the component detectors, but it may also adversely affect detectability of events and timeliness of detection. In addition, in a parallel approach, a potentially useful interplay between the streams is completely ignored.

The alternative is to derive an algorithm which would aggregate the output of the independently constructed single-stream detectors. Typically, the aggregate detector is made sensitive to cases when more than one stream is out of control or many of them are near critical. In a sense, these individual detectors must agree to cause the system-wide alert, and therefore such methods are often labeled as based on consensus.

3.1 Consensus Approach

Roure et al. (2007) discuss an example of the consensus approach. It considers three streams of independently collected food and agriculture safety data involving records of daily counts of condemned and healthy cattle, counts of positive and negative microbial tests of food samples, and counts of passed and failed sanitary inspections conducted at the U.S. Department of Agriculture-regulated slaughter houses. Roure et al. use a temporal scan algorithm as their basic detection tool, although any single-stream probabilistic anomaly detection method could be used in its place. It slides a fixed-width time window along the temporal dimension and compares the positive and negative counts inside of it against the aggregated counts observed during the outside period of reference, and – depending on the sizes of the involved samples – it applies either Chi-square or Fisher's exact test of independence to the obtained two-by-two contingency table. The more the observed counts of positives and their proportion to negatives inside the time window differ from the expectation based on the counts aggregated during the reference interval, the lower the p-value resulting from the test.

A parallel surveillance system would monitor the individual streams of p-values and trigger an alert if one or more of them went below a pre-set threshold, say $\alpha = 0.05$. It would have lower than attainable detection power if the available streams of data actually carried corroborating evidence. In the consensual approach, the anomaly detection algorithm monitors the aggregate of the component p-values which represents the consensus estimate of strangeness. At first glance, it might be tempting to apply Bonferroni's method for correction against effects of multiple testing and signal an alarm whenever the smallest p-value passes the corrected test. That, in terms of detection power, would bring results equivalent to applying *Min* (minimum) function to the set of p-values. In most cases, it makes more practical sense to use one of the techniques of combining p-values such as Fisher's (Fisher, 1948) or Edgington's (Edgington, 1972).

Since p-values follow a uniform distribution under the null hypothesis, the Fisher statistic (the doubled sum of natural logarithms of the *m* independent p-values) has a χ^2 distribution with 2*m* degrees of freedom. Conceptually, this approach is easier to understand as computing an aggregate statistic which is the product of the component p-values. It also turns out (Jost) that there exists a closed form solution for the combined p-value, p_F:

$$p_F = k\sum_{i=0}^{m-1}\frac{(-\ln k)^i}{i!} \text{, where } \quad k = \prod_{i=0}^{m-1} p_i$$

Edgington's aggregation is based on an additive model for p-values. If *S* denotes the sum of the *m* component p-values, then Edgington's aggregate can be computed as follows:

$$p_E = \sum_{j=0,1,\ldots}\binom{m}{2j}\frac{(S-2j)^m}{m!} - \sum_{j=0,1,\ldots}\binom{m}{2j+1}\frac{(S-(2j+1))^m}{m!}$$

The summations stop as soon as $S \le 1$.

Figure 7-1 illustrates the effects of Fisher's and Edgington's aggregations of p-values computed separately for two independent streams of data. These p-values, labeled P1 and P2, correspond to the horizontal and the vertical axes of the graph, respectively. If simply the individual streams are taken into consideration, and if the critical sensitivity is set to $\alpha = 0.05$, then the null hypothesis will be rejected if either P1 < 0.05 or P2 < 0.05, and the corresponding rejection regions in the graph would be rectangular in shape. The *Min* aggregation will result in concatenation of these two rectangles, as the null hypothesis will be rejected whenever at least one of the p-values is lower than critical. Fisher's and Edgington's methods add to the rejection region the instances where both component p-values are just slightly greater than critical. That enables flagging cases in which the individual streams are of marginal interest on their own, but they appear unusual when the corresponding pieces of evidence are combined. Both aggregates are, on the other hand, more conservative than the *Min* algorithm when either of the component p-values is substantially greater than critical, which makes sense in many practical situations. Fisher's method is multiplicative and as such it is sensitive to small numbers of low p-values. Additive by design, Edgington's method is slightly more sensitive to cases with multiple borderline p-values. This distinction determines in practice the choice between these two approaches.

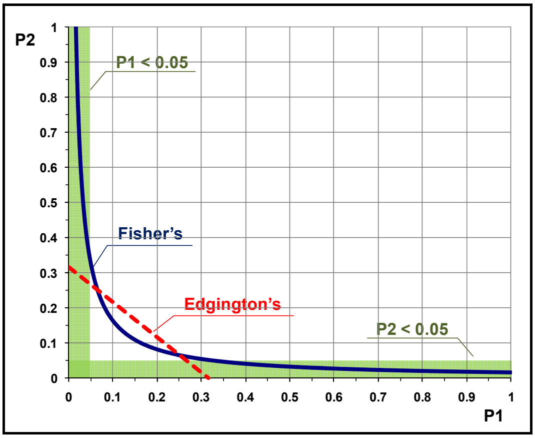


Figure 7-1. Effects of aggregation of two independent p-values using $\alpha = 0.05$. *Min* criterion leads to rejecting null hypothesis if the individual p-values fall inside either of the shaded rectangular areas. The boundaries of the rejection regions for Fisher's and Edgington's methods are shown with the solid blue and the dashed red line, respectively.

An alternative heuristic approach to combining signals produced by a number of independent univariate control charts has been proposed in (Yahav and Shmueli, 2007). It uses either a majority vote or an "M + n" rule to aggregate indications of the individual charts. The majority vote triggers an alert if more than half of the component detectors signal alerts. The "M + n" rule distinguishes a specific subset of M detectors and requires that all items in that set, as well as at least n other detectors, must be activated for the combined system to raise an alert. The method was originally tested on single-stream data and used a set of different control charts concurrently monitoring the same stream, but a similar approach may potentially be applicable to multi-stream scenarios.

3.1.1 Handcrafting Specific Detectors

Fisher's or Edgington's methods can be referred to as non-specific consensual detectors because they are treating each of the component streams equally in their targeting of departures from the joint null distribution. That can be useful when no information about the particular signatures of the patterns of interest is available. Otherwise, tweaking the basic model may produce better results. For example, in (Roure et al., 2007) a hypothetical outbreak scenario is used in which a ramp up in positive observations occurs simultaneously in all (three) data streams and a handcrafted extension to Fisher's method is designed to benefit from the additional piece of structural information about the targeted pattern. The modified detector considers the same combined p-value as the original, but then it chooses not to signal an alarm if fewer than two of the uncombined p-values are below a set threshold. Those are the cases where the evidence definitely does not match the scenario of interest, since at most one stream departs from its null distribution. The specific detector constructed this way reduces the number of false positives and that in turn allows for an increase of sensitivity and, consequently, enables earlier detections. Figure 7-2 presents a set of Activity Monitoring Operating Characteristic curves obtained for the temporal scan detectors set to work on the individual streams of data (labeled A, B and C), plain Fisher's aggregation method (F), and the specific, modified Fisher's detector explained above (F+). As anticipated, Fisher's aggregation substantially improves detection power (timeliness as well as alert frequency) over the results obtained using the individual streams, and the handcrafted specific multi-stream detector slightly outperforms the basic Fisher's approach.

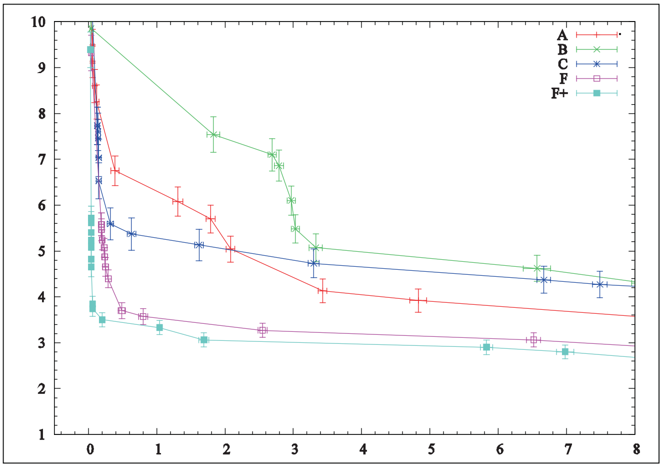


Figure 7-2. AMOC curves for univariate detectors (A, B, and C), non-specific Fisher's method (F), and manually designed specific detector (F+). The horizontal axis of the graph corresponds to the number of detects outside of the period of the injected synthetic outbreaks, the vertical axis denotes the time to detection in days from the first day of the outbreak.

3.1.2 Learning Specific Detectors from Data

Manual design of specific detectors may be impractical due to the underlying complexity of the processes represented in data. Also, maintaining their relevance and high detection power when facing changing environments can become a serious technical and organizational challenge (Dubrawski et al., 2006). Machine learning provides an appealing framework for using the labeled historical data to automatically train classifiers capable of discriminating periods of time that may belong to a specific outbreak type, from periods during which no such outbreaks occur. Following the idea behind the handcrafted specific detector, machine learning methods can be used to support filtering out false alarms from the stream of aggregated p-values produced by one of the consensual techniques (Figure 7-3).

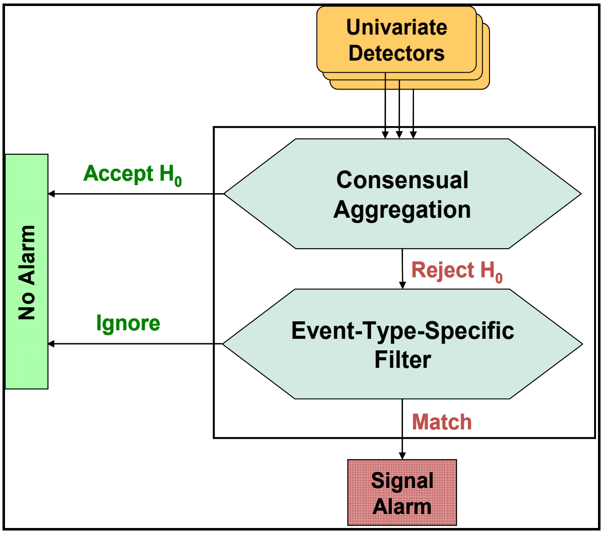


Figure 7-3. Schematic diagram of the idea of the filtering approach to constructing specific multi-stream detectors. P-values of single-stream tests are first processed by the consensual aggregation filter which rejects the null hypothesis whenever the combined p-value is lower than critical. If that happens, the candidate alert is checked by the event-type-specific filter which only allows raising alerts when the characteristics of the statistically significant events match the predefined patterns of interest, otherwise the candidate alerts are ignored.

That would be achievable if we could use the available labeled training data to train classifiers which could reliably tell apart potentially false detections from true ones. In (Roure et al., 2009), a classifier is trained from a real multi-stream data with injected synthetic outbreaks. Days belonging to the period of known outbreaks are labeled as positive training examples, while the days following the outbreak periods are used as negatives. The features prepared for training are derived from the p-values obtained for the individual streams, as well as their Fisher aggregates. The vector of features can be composed of those sets of p-values extracted for a few consecutive days ending at the day of the analysis and a few different widths of the temporal scan windows. The results presented below use three consecutive days and three widths of the observation interval (1-, 2- and 3-day wide). Therefore, the input space consists of 36 real-valued features corresponding to the Cartesian product of four streams of p-values, three temporal scan window widths and 3 days.

An important practical limitation of machine learning approaches is that in real-world scenarios it is hard to expect availability of many labeled events in data. The process of labeling can be difficult, time consuming and sometimes expensive, and the frequency of occurrence of actual, documented and attributed adverse events is typically (and luckily) quite low. However, the availability of identified negative examples is in practice usually much better. In such cases, a density model can replace the classifier. If only negative examples are available (the non-outbreaks), one can build a model of their distribution in the feature space similar to that described above, and query it whenever a candidate alert is raised by the non-specific detector. The alert can then be lifted if the query returns a sufficiently high likelihood value. That would happen whenever the considered case looks similar to previously recorded negative examples. Roure et al. use a non-parametric method of Kernel Density Estimation to construct such a filter. They also propose to independently construct density models for positive examples if some are available. Then, each candidate alert can be checked against historical distributions of positive and negative examples, and the final decision can be determined using, e.g., Bayes' rule. Note that this approach can also handle multi-focus scenarios with a number of distinct types of outbreaks: each type can be tackled by a separate density model.

Figure 7-4 depicts the AMOC curves for the Fisher-based non-specific detector (F), manually designed specific detector (F+), and the classifier-based specific detector (L). The classifier clearly beats all of the previously discussed detectors. This is not surprising since it has access to all of the

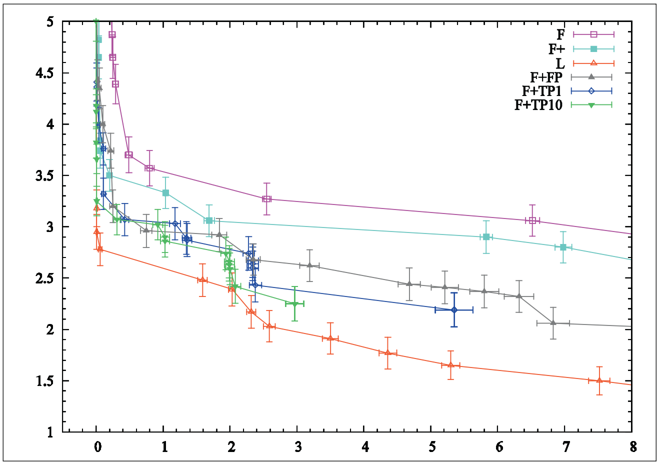


Figure 7-4. AMOC curves for Fisher-based detector (F), handcrafted specific detector (F+), the classifier-based specific detector (L), and the density-based with negative labels only (F+FP), labels from one (F+TP1) and ten (F+TP10) outbreaks.

information accessible to its competitors, in addition to 100 example outbreaks. These labeled examples allow the classifier to identify the winning combination of features and the kinds of patterns it seeks to match. Figure 7-4 also includes three curves for density-based detectors: one using only negative labels (F+FP), another using positive labels from just one outbreak (F+TP1) and one that uses positive labels from ten independent outbreaks (F+TP10). The observed characteristics of density-based detectors fall between the handcrafted and the classification-based detector characteristics: they significantly outperform the first and perform worse than the latter. This result follows intuition since the classification-based detector has access to much more information than the density-based models. In general, as expected, the more informed the detectors, the greater their power. For most parts of the AMOC curves, F+TP10 outperforms F+TP1, which in turn outperforms F+FP.

The utility of machine learning as a pragmatic approach to data-driven event detection has also been advocated by others (Neill, 2007). We expect it to make a substantial impact on the methodology of surveillance. It has already been demonstrated that it is possible to learn efficient detectors of specific types of events from labeled multi-stream data. The automatically built detectors are sometimes able to match and outperform the manually designed alternatives in terms of both speed of detection and accuracy. That particular finding is very interesting as it could lead to substantial reductions in costs of development and maintenance of future event detection systems. Recent advances in machine learning research, especially related to semi-supervised and active learning methods, will boost development of cost-effective multi-stream surveillance systems. They will enable efficient training of specific multi-focus detection models from sparsely labeled examples, and they will incorporate user's feedback for continued adaptation of the models to changing environments after their deployment, leading to smart and economical surveillance systems.

3.2 Multi-Stream Spatial Scan

Spatial scan statistic (Kulldorff, 1997) and space-time scan statistic (Kulldorff, 2001) are often used in the public health community to detect spatio-temporal concentrations of disease cases. These methods have a long and successful history and their fundamental variants typically monitor a single stream of spatially labeled data, detecting regions and times where the magnitude of signal is significantly higher than expected. The fundamental algorithms may become computationally infeasible in large-scale deployments. The Fast Spatial Scan algorithm (Neill and Moore, 2004) resolves this issue by using smart data structures and branch-and-bound search for the most unusual spatio-temporal sub-region.

A simple extension of the univariate spatial or space-time scan would follow the idea of parallelization of multi-stream detectors, but that would not use the information about the potential interplay between multiple streams to boost the detection power. Alternative approaches, presented in (Burkom, 2003) and (Kulldorff et al., 2007), characterize each stream with the log-likelihood ratios and assume that they are mutually independent. Then a region with the highest sum of log-likelihood ratios is flagged. These methods are non-specific in that they try to distinguish between the cases in which all streams are affected and those when no streams are indicative. In practice though, different types of events may have specific and varying effects on the individual streams. In addition, the individual streams are typically not independent of each other as they may be influenced by common confounding factors such as population size, availability constraints, seasonal effects, and the features of the ongoing outbreak.

The Multivariate Bayesian Spatial Scan (MBSS) algorithm (Neill and Cooper, 2008) attacks those challenges by modeling the joint distribution of all data streams given each possible hypothesis: either no outbreak is going on, or an outbreak of a specific type is occurring in the specific spatial region. Empirically, it achieves faster and more accurate detections of emerging outbreaks while modeling and distinguishing between different types of events. It was demonstrated to work well with uninformative priors, while higher detectability and discriminative power can be achieved for specific event types (e.g., a terrorist anthrax attack vs. seasonal flu). Detection models for MBSS can be learned from labeled training data, from expert knowledge, or from a combination of the two, even if the number of the available labeled events is relatively small.

4. MULTI-DIMENSIONAL ANALYSIS

In multivariate time series data, the atomic events of interest have already been aggregated into two or more time series, indexed usually by day, hour or week. Thus, at each time step, there might be one number characterizing daily volume for OTC cold medicines sales, one number for school absenteeism, one for the count of gastrointestinal emergency room admits, etc. In large surveillance projects there may be hundreds of thousands of such time series, with perhaps ten or twenty for each of the monitored geographic locations such as an individual zip code, for example. Such aggregated data yields itself to analysis using the methods described before.

However, often the source data used to obtain interval aggregates comes in the form of a record of transactions. Each entry in it corresponds to a unique discrete event such as an emergency room admission or a placed prescription order. Each of these events may involve many categorical descriptors (also referred to as attributes, features or dimensions) such as gender of a patient, syndromes, geographic location, age group, test results, diagnoses, etc. Performing analyses which account for these multiple descriptors requires methods specifically designed to deal with the multi-dimensional data.

The What's Strange About Recent Events (WSARE) algorithm detects anomalous patterns in discrete, multi-dimensional datasets with a temporal component (Wong et al., 2005b). WSARE requires records within a specified temporal period to be defined as recent. Records preceding the recent data in time are used to produce a baseline dataset that represents normal behavior. WSARE compares the recent data against the baseline data to find a combination of dimensions which corresponds to the most significant change in recent records. This change is described using a rule, which is composed of components in the form $x_i = v_i^j$, where x_i is the *i*th attribute and v_i^j is the *j*th value of that attribute. For example, the one-component rule *Gender = Male* characterizes the subset of the data involving males. Like SQL SELECT queries, rules in WSARE can consist of multiple conjunctive components (connected using the logical AND). For example, a two component rule could be *{Gender = Male} AND {Home Location = NW}*, which identifies the group consisting of males living in the northwest region of the city. WSARE identifies the rule which selects the group of records with the most significant change in its proportion between the recent dataset and the baseline dataset.

The key part of the WSARE algorithm is the synthesis of data to represent baseline distribution. One way of doing that is to use a Bayesian Network trained to model joint probability distribution of the values of the attributes of the reference data. The available dimensions can usually be divided into environmental attributes, such as the season and the day of week that may cause trends in the data, and response attributes, such as syndrome or gender of a patient reporting to an emergency room. During the BN structure learning phase, environmental attributes are prevented from having parents because, although we are not interested in predicting their distributions, we still want to use them to predict the response attributes. Once the BN structure is learned, the reference data is used to estimate conditional probability distributions which represent the baseline behavior given the environmental attributes observed on the current day. As an example, suppose we are monitoring emergency room data and that the environmental attributes Season, Day of Week, and Weather cause fluctuations in this data. Also, let the response attributes be $x_1, \ldots, x_m$. Assuming that today is a snowy winter Saturday, we can use the joint probability distribution captured by the Bayesian network to produce the conditional probability distribution

$P(x_1,\ldots,x_m \mid \mathit{Season} = \mathit{Winter},\ \mathit{Day\ of\ Week} = \mathit{Saturday},\ \mathit{Weather} = \mathit{Snow})$, which represents the baseline distribution given the conditions for the current day. The baseline dataset can then be produced by sampling a large number of records from this conditional probability distribution.

Once the baseline dataset is generated, a search for the highest scoring rule, which characterizes the set of attribute = value terms with the most unusual shift in proportions between the baseline and recent datasets, is conducted. Significance of the winning rule is measured using a randomization procedure consisting of several iterations. At each iteration, the dates are randomly swapped between records in the recent and the baseline datasets to produce a synthetic set of data. Then, the best scoring rule is found in it. At the end, ranking of the original highest scoring rule with respect to the best rules derived from multiple synthetic sets is determined. That rank, divided by the number of randomization trials, defines the empirical p-value. An alert is triggered whenever this p-value is lower than a threshold, for example, 0.01.

The Israel Center for Disease Control evaluated WSARE retrospectively using an unusual outbreak of influenza among school children. The algorithm monitored patient visits to community clinics over 18 days in spring 2004. The considered dimensions included the visit date, area code, ICD-9 code, age category, and day of week (which was used as the only environmental attribute). Two of the five flagged anomalous patterns corresponded to the actual influenza outbreak in the data. The rules that characterized the anomalous patterns consisted of the same three attributes of ICD-9 code, area code and age category, indicating involvement of children aged 6–14 having viral symptoms within a specific geographic area. WSARE made that detection on the second day from the onset. Similarly, in a retrospective analysis of the Walkerton outbreak, it was determined that WSARE would have detected the outbreak one day before a boil-water advisory was released if its alarm threshold was set to the level that permitted two false positive alerts per year.

An important feature of WSARE is its ability to identify data dimensions which are responsible for the alert. It does so by exhaustively searching for the most surprising pattern across all combinations of attribute = value pairs up to a certain size, and reporting the identified set of such pairs. An important practical implication is that the users do not have to specify which dimensions to monitor since every conceivable query is taken under consideration. Therefore no important event can be missed, as long as it is represented in data and if it makes it to the top of the significance rankings.

Multivariate detections provided by WSARE come at a substantial computational expense, mostly due to the costs of randomization tests. Facing datasets of sizes and dimensionalities typically found in biosurveillance applications, WSARE is able to produce results in reasonable times when aiming at rules up to the size of about 3. In most practical cases that is perfectly acceptable, but it also illustrates a challenge of scalability commonly encountered when analyzing highly multi-dimensional data.

A substantial part of the computational effort in such scenarios can often be attributed to the retrieval of time series of counts of events which match a specific query. Online Analytical Processing (OLAP) is a category of software techniques that enables fast response to certain aggregation queries against very large datasets. It pre-computes multiple views of selected data by aggregating values across all possible attribute combinations (a "group-by" operation in database terminology). The resulting data structures (data cubes) can then be used to dramatically accelerate visualization and query tasks associated with large datasets. Normally, the data cube also includes aggregation at different levels within a dimension (e.g., levels of state, city, and school district within area dimension) to allow the multi-dimensional viewing of data at different granularity (in support of drill-down and roll-up functions). Typical aggregation functions include, but are not limited to, sum, average, percentage of total, ranking (topN), and time to date (specific to time dimension). It normally takes hours or even longer to create a data cube (including aggregation and indexing) and at least minutes to update it, while it takes only seconds to query it for a pre-summarized parameter set. However, when responding to an ad hoc query (which does not have the corresponding aggregation existing in the cube), the retrieval procedure has to sift through the raw data and do the "group by"-like database query "on the fly." This task inevitably takes longer. In many biosurveillance applications, the advanced time series queries more often take the form of ad hoc rather than pre-determinable queries. This is particularly true if we give the users the ability to request on the fly analyses such as spatial scan, anomaly detection or outbreak diagnosis against dynamically selectable or exhaustively monitored geographical regions and demographic subsets of data.

A solution is provided by the T-Cube data structure (Sabhnani et al., 2007a). T-Cube is an in-memory cached sufficient statistic equivalent to the data cubes known in OLAP applications. It generalizes over the idea of AD-trees (Moore and Lee, 1998) which allows highly compressed main-memory storage and very fast retrieval of aggregate statistics such as counts, means and variances from large multi-dimensional datasets. AD-trees have been very successful in speeding up statistical algorithms, such as Bayesian Network learning, association rule learning, or decision tree learning, that need to search over many conjunctive queries, many contingency tables or many conditional probability tables of large datasets.

The essential property of the T-Cube is that once built, time series for any query (in a general class including conjunctions of disjunctions) can be obtained in constant time, which does not depend on the number of records in the raw dataset. One example of such a query is "retrieve the series of daily counts of emergency department visits by all males in postal codes 15213, 15217 and 15206, excluding children, which were specific to gastrointestinal or respiratory syndromes." An example given in (Dubrawski et al., 2007) involves a dataset with about 25 dimensions of arities varying from 2 to 80, covering a record of about 12,000 transactions, spanning over more than 2,000 temporal intervals. The application required a search through all combinations of attribute = value pairs of sizes 1 and 2, with the total number of such combinations in excess of four million. The analysis involved expectation-based temporal scan executed to detect unusual short term increases in counts of specific aggregated time series. The total number of individual temporal scan tests for such a dataset exceeded nine billion. Each such test involves a Chi-square test of independence performed on a 2-by-2 contingency table formed by the counts corresponding to the time series of interest (one of the four million series) and the baseline counts, within the current temporal window of interest (one of 2,000+) and during the reference interval. The complete set of computations, including the time necessary to retrieve and aggregate all the involved time series, compute and store the test results, load source data and build the T-Cube structure, etc., took about 8 h of CPU time. Using a commercial database tool the average time to retrieve one of the involved queries on the same hardware approached 180 ms. Therefore, without the T-Cube, it would take about 9 days just to pull all the required time series from the database, not including any processing or execution of statistical tests.

T-Cube is not a fits-all replacement for on-disk OLAP data structures. It is specifically designed for very rapid searching of millions of combinations and time series aggregations of demographic, spatial, syndromic, and similar dimensions within probabilistic models (Sabhnani et al., 2007b). Rapid access to complex extracts of data not only makes the data-intensive analyses feasible, but it also enables the user-level data navigation (drill-downs, roll-ups, visualization) at interactive speeds (Ray et al., 2008).

5. CONCLUSION

Modern biosurveillance is set to realize benefits from simultaneous consideration of evidence originating from multiple corroborating sources of information. The most important of these benefits include improved power and specificity of detection. Already, several promising analytic algorithms have been proposed and a few of them have been successfully transitioned to practice. Nonetheless, further development is still required in order to fully exploit the opportunities. The key challenges include scalability of the algorithms and computational tractability of the corresponding datasets; ability to handle non-stationary processes with complex patterns of inter-stream relationships; reliability of detection even if the available data is scarce, incomplete and noisy in prediction as well as in the model estimation phase; and the ability to deal with heterogeneous types of data. Future multi-stream event detection systems will benefit from incorporation of machine learning. This will enable adaptability to changes in analytic environment, consideration of human feedback, and model estimation from sparsely labeled data. These algorithms will be capable of performing large-scale inferences in support of the analyst decisions and they will guide their users in support of data exploration, problem diagnosis and isolation. They will also provide predictive analytic capabilities for proactive mitigation of risk of adverse bioevents.

ACKNOWLEDGEMENTS

Many thanks to Andrew Moore, Jeff Schneider, Gregory Cooper, Daniel Neill, Weng-Keen Wong, Josep Roure and Maheshkumar Sabhnani for providing information used in the text. This work is partially supported by the Centers for Disease Control (award number R01-PH000028) and by the National Science Foundation (under grants number 0325581 and 0911032).

QUESTIONS FOR DISCUSSION

1. What are the fundamental differences between multivariate, multi-stream and multi-dimensional event detection problems?
2. What are the conditions of applicability of statistical multivariate baseline estimation and event detection approaches to multi-stream problems?
3. What is the impact of multiple hypotheses testing on reliability of multi-stream detection scenarios?
4. Why does a direct application of detection threshold adjustment techniques such as Bonferroni correction or False Discovery Rate not boost the power of the resulting aggregated multi-stream detector?
5. What are the key challenges in design and practical application of multi-stream event detection systems? How can some of them be addressed with the use of machine learning methodology?

REFERENCES

Box, G., Jenkins, G.M., and Reinsel, G. (1994) *Time Series Analysis: Forecasting and Control*. Englewood Cliffs, NJ: Prentice Hall.

Burkom, H.S. (2003). "Biosurveillance applying scan statistics with multiple, disparate data sources." *Journal of Urban Health* 80(2 Suppl 1):57–65.

Burkom, H.S., Elbert, Y., Feldman, A., and Lin, J. (2004) "The role of data aggregation in biosurveillance detection strategies with applications from ESSENCE." *Morbidity and Mortality Weekly Report (Supplement)* 53:67–73.

Burkom, H.S., Murphy, S., Coberly, J., and Hurt-Mullen, K. (2005). "Public health monitoring tools for multiple data streams." *Morbidity and Mortality Weekly Report (Supplement)* 54:55–62.

CMAJ (2000). "Leadership and fecal coliforms: Walkerton 2000." Editorial, *Canadian Medical Association Journal* 163(11):1417

Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W.K., Hogan, W.R., and Wagner, M.M. (2004). "Bayesian biosurveillance of disease outbreaks." In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, Banff, Canada, pp. 94–103.

Cooper, G.F., Dowling, J.N., Levander, J.D., and Sutovsky, P. (2006) "A Bayesian algorithm for detecting CDC category A outbreak diseases from emergency department chief complaints." *Advances in Disease Surveillance* 2:45.

Crosier, R.B. (1988). "Multivariate generalizations of cumulative sum quality-control schemes." *Technometrics* 30(3):291–303.

Dubrawski, A., Elenberg, K., Moore, A., and Sabhnani, R. (2006). "Monitoring food safety by detecting patterns in consumer complaints." In: *Proceedings of AAAI/IAAI'06*, Boston, USA.

Dubrawski, A., Sabhnani, M., Ray, S., Roure, J., and Baysek, M. (2007) "T-Cube as an enabling technology in surveillance applications." *Advances in Disease Surveillance* 4:6.

Edgington, E.S. (1972). "An additive method for combining probability values from independent experiments." *Journal of Psychology* 80:351–363.

Fisher, R. (1948). "Combining independent tests of significance." *Journal of American Statistics* 2(5):30.

Fricker, R.D. (2007). "Directionally sensitive multivariate statistical process control procedures with application to syndromic surveillance." *Advances in Disease Surveillance* 3(1):1–17.

Held, L., Höhle, M., and Hofmann, M. (2005). "A statistical framework for the analysis of multivariate infectious disease surveillance counts." *Statistics Modelling* 5:187–199.

Hotelling, H. (1947). "Multivariate Quality Control." In: C. Eisenhart, M.W. Hastay, and W.A. Wallis, editors. *Techniques of Statistical Analysis*. New York: McGraw-Hill.

Jost, L. "Combining Significance Levels from Multiple Experiments or Analyses." Available at http://www.loujost.com/Statistics%20and%20Physics/Significance%20Levels/CombiningPValues.htm.

Kulldorff, M. (1997). "A spatial scan statistic." *Communications in Statistics: Theory and Methods* 26(6):1481–1496.

Kulldorff, M. (2001). "Prospective time-periodic geographical disease surveillance using a scan statistic." *Journal of the Royal Statistical Society A* 164:61–72.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K., and Platt, R. (2007). "Multivariate spatial scan statistics for disease surveillance." *Statistics in Medicine* 26(8):1824–1833.

Lotze, T., Shmueli, G., Murphy, S., and Burkom, H. (2006). "A Wavelet-based Anomaly Detector for Early Detection of Disease Outbreaks." ICML Workshop on Machine Learning Algorithms for Surveillance and Event Detection, Pittsburgh, PA.

Lowry, C.A., Woodall, W.H., Champ, C.W., and Rigdon, S.E. (1992). "A multivariate exponentially weighted moving average chart." *Technometrics* 34:46–53.

Madigan, D. (2005). "Bayesian Data Mining for Health Surveillance." In: B. Lawson and K. Kleinman, editors, *Spatial and Syndromic Surveillance for Public Health*, New York: Wiley.

Mohtashemi, M., Kleinman, K., and Yih, W.K. (2007). "Multi-syndrome analysis of time series using PCA: a new concept for outbreak investigation." *Statistics in Medicine* 26:5203–5224.

Montgomery, D.C. (2000). *Introduction to Statistical Quality Control*, 4th ed. New York: Wiley.

Moore, A.W. and Lee, M.S. (1998) "Cached sufficient statistics for efficient machine learning with large datasets." *Journal of Artificial Intelligence Research* 8:67–91.

Neill, D.B. and Moore, A.W. (2004) "A fast multi-resolution method for detection of significant spatial disease clusters." In: *Advances in Neural Information Processing Systems*. Vol. 16, pp. 651–658.

Neill, D.B. (2007) "Incorporating learning into disease surveillance systems." *Advances in Disease Surveillance* 4:107.

Neill, D.B. and Cooper, G.F. (2008). "A multivariate Bayesian scan statistic for early event detection and characterization." *Machine Learning* 79:261–282.

Pignatiello, J.J. and Runger, G.C. (1990). "Comparisons of multivariate CUSUM charts." *Journal of Quality Technology* 22(3):173–186.

Ray, S., Michalska, A., Sabhnani, M., Dubrawski, A., Baysek, M., Chen, L., and Ostlund, J. "T-Cube Web Interface: A Tool for Immediate Visualization, Interactive Manipulation and Analysis of Large Sets of Multivariate Time Series." AMIA Annu Symp Proc. 2008 Nov 6:1106.

Rolka, H., Burkom, H., Cooper, G.F., Kulldorff, M., Madigan, D., and Wong, W.K. (2007). "Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs." *Statistics in Medicine* 26(8):1834–1856.

Roure, J., Dubrawski, A., and Schneider, J. (2007). "A study into detection of bio-events in multiple streams of surveillance data." In: *Intelligence and Security Informatics: Biosurveillance*, LNCS Vol. 4506, Berlin: Springer-Verlag.

Roure, J., Dubrawski, A., and Schneider, J. (2007). "Learning Specific Detectors of Adverse Events in Multivariate Time Series." Advances in Disease Surveillance 4:111.

Sabhnani, M., Moore, A.W., and Dubrawski, A.W. (2007a) "T-Cube: a data structure for fast extraction of time series from large datasets." Technical Report CMU-ML-07-114, Carnegie Mellon University.

Sabhnani, M., Dubrawski, A., and Schneider, J. (2007b). "Multivariate time series analyses using primitive univariate algorithms." *Advances in Disease Surveillance* 4:112.

Shen, Y. and Cooper, G.F. (2007). "A Bayesian biosurveillance method that models unknown outbreak diseases." *Intelligence and Security Informatics: Biosurveillance*, LNCS Vol. 4506, Berlin: Springer-Verlag.

Siddiqi, S.M., Boots, B., Gordon, G.J., and Dubrawski, A.W. (2007). "Learning stable multivariate baseline models for outbreak detection." *Advances in Disease Surveillance* 4:266.

Sonneson, C. and Frisén, M. (2005). "Multivariate Surveillance." In: B. Lawson and K. Kleinman, editors, *Spatial and Syndromic Surveillance for Public Health*, New York: Wiley.

Wasserman, L. (2004). *All of Statistics*. Berlin: Springer-Verlag.

Wong, W.K., Cooper, G.F., Dash, D.H., Levander, J.D., Dowling, J.N., Hogan, W.R., and Wagner, M.M. (2005a) "Bayesian biosurveillance using multiple data streams." *Morbidity and Mortality Weekly Report (Supplement)* 54:63–69.

Wong, W., Moore, A., Cooper, G., and Wagner, M. (2005b). "What's strange about recent events (WSARE): an algorithm for the early detection of disease outbreaks." *Journal of Machine Learning Research* 6:1961–1998.

Yahav, I. and Shmueli, G. (2007). "Algorithm Combination for Improved Performance in Biosurveillance Systems." In: *Intelligence and Security Informatics: Biosurveillance*, LNCS Vol. 4506, Berlin: Springer-Verlag.

SUGGESTED READING

Burkom, H.S., Murphy, S., Coberly, J., and Hurt-Mullen, K. (2005). "Public health monitoring tools for multiple data streams." *Morbidity and Mortality Weekly Report (Supplement)* 54:55–62.

Rolka, H., Burkom, H., Cooper, G.F., Kulldorff, M., Madigan, D., and Wong, W.K. (2007). "Issues in applied statistics for public health bioterrorism surveillance using multiple data streams: research needs." *Statistics in Medicine* 26(8):1834–1856.

Wagner, M.M., Moore, A.W., and Aryel, R.M. (2006) editors. *Handbook of Biosurveillance*. New York, NY: Academic Press.

Wong, W.K., Cooper, G.F., Dash, D.H., Levander, J.D., Dowling, J.N., Hogan, W.R., and Wagner, M.M. (2005a) "Bayesian biosurveillance using multiple data streams." *Morbidity and Mortality Weekly Report (Supplement)* 54:63–69.

ONLINE RESOURCES

- Engineering Statistics Handbook (http://www.itl.nist.gov/div898/handbook/index.htm):
  This interactive handbook developed and hosted by the National Institute of Standards and Technology is a rich source of information about fundamental methods of statistical analysis. It includes comprehensive descriptions of multivariate control charts which can be used to monitor multi-stream data.
- Dataplot (http://www.itl.nist.gov/div898/software/dataplot/homepage.htm):
  Dataplot is a free, public-domain, multi-platform software for scientific visualization, statistical analysis, and modeling developed at the National Institute of Standards and Technology. Its extensive functionality includes a set of statistical process control and time series analysis methods; it also supports process monitoring.

- SaTScan™
  (http://www.satscan.org/):
  SaTScan™ is a free software that analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. It can also scan multiple datasets simultaneously to look for clusters that occur in one or more of them.
- WSARE and Fast Spatial Scan
  (http://www.autonlab.org/autonweb/downloads/software.html):
  Free downloadable implementations of WSARE and scalable spatial scan are available from the Carnegie Mellon University Auton Lab web site.
- T-Cube
  (http://www.autonlab.org/T-Cube/):
  A public demo version of the T-Cube prototype web interface is available at the CMU Auton Lab web site. It includes massive screening and detection functions which can be executed against user-supplied or locally available example multi-dimensional data to demonstrate efficiency of the underlying data structure.
- WEKA
  (http://www.cs.waikato.ac.nz/ml/weka/):
  WEKA is a free Java language library including many machine learning algorithms which can be used to implement data-driven approaches to detection of events in multi-stream surveillance.

Chapter 8

ALGORITHM COMBINATION FOR IMPROVED PERFORMANCE IN BIOSURVEILLANCE

Univariate Monitoring

INBAL YAHAV[1,]*, THOMAS LOTZE[2], and GALIT SHMUELI[1]

CHAPTER OVERVIEW

This chapter proposes an enhancement to currently used algorithms for monitoring daily counts of pre-diagnostic data. Rather than use a single algorithm or apply multiple algorithms simultaneously, our approach is based on ensembles of algorithms. The ensembles lead to better performance in terms of higher true alert rates for a given false alert rate. Combinations can be employed at the data preprocessing step and/or at the monitoring step. We discuss the advantages of such an approach and illustrate its usefulness using authentic modern biosurveillance data.

Keywords: Control charts; Monitoring; Ensemble methods; Optimization; Pre-diagnostic data

1. INTRODUCTION

In 1918 one of the deadliest influenza pandemics in history erupted, called the Spanish Flu. Approximately 20–40% of the worldwide population fell ill and over 50 million people died. Outbreaks followed shipping routes

[1*] *Department of Decision, Operations & Information Technologies, Robert H Smith School of Business, University of Maryland, College Park, MD 20742, USA*

[2] *Applied Mathematics & Scientific Computation Program, University of Maryland, College Park, MD 20742, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_8,

from North America through Europe, Asia, Africa, Brazil and the South Pacific. The pandemic reached its peak after 5–6 months. Nearly 40 years later, in February 1957, the Asian influenza pandemic erupted in the Far East. Unlike the Spanish Flu, the Asian influenza pandemic virus was quickly identified and vaccines were available 6 months later. Approximately two million people died in this outbreak (compared to the 50 million in the Spanish Flu). Other known outbreaks in history, such as the Hong Kong Flu (1968–1969), the Avian Flu (1997) and SARS (2003) also resulted in high death tolls over the years. Unfortunately the threat of new pandemic outbreaks is still looming.

A major goal of public health is to figure out whether and how transmission of diseases can be diminished. Researchers at the Center for Humanitarian Logistics at Georgia Tech have shown that pandemic outbreak effects can be greatly reduced if quarantine is imposed at the early stages of the disease.[1] The U.S. Centers for Disease Control & Prevention (CDC) lay out guidelines and strategies for reducing disease transmission, including use of personal protective equipment (e.g., masks and gloves), hand hygiene, and safe work practices. The CDC also recommends actions to be taken during the earliest stage of a pandemic, when the first potential cases or disease clusters are detected. These include individual-level containment measures such as patient isolation and identification, monitoring, and quarantine of contacts.[2]

The early detection of disease outbreaks therefore plays a major role in preventing disease transmission and reducing the size of the affected population. In modern biosurveillance a wide range of pre-diagnostic and diagnostic daily counts are monitored for the purpose of alerting public health officials when there is early evidence of a disease outbreak. This is in contrast to traditional biosurveillance, where only diagnostic measures (such as mortality and lab reports) are examined, usually locally, and at aggregation levels such as weekly, monthly, or annually. Moreover, in modern biosurveillance the goal is prospective while traditional biosurveillance is more retrospective in nature. Although the tasks and data types and structures differ widely between traditional and modern biosurveillance, most monitoring algorithms have been migrated from traditional to modern systems. The result is that current modern biosurveillance detection methods suffer from multiple statistical and practical limitations that greatly deteriorate their ability to achieve their intended purpose. For a general overview of the statistical challenges that arise in biosurveillance see Shmueli and Burkom (2008). In particular, there is often a mismatch between the types of

[1] http://www.tli.gatech.edu/research/humanitarian/projects.php

[2] http://www.hhs.gov/pandemicflu/plan/appendixf.html.

algorithms being used and the data structure of modern biosurveillance data. This means that the assumptions behind those algorithms with regards to baseline behavior and outbreak nature are often violated in the modern biosurveillance context. Another important problem is that of multiplicity: when monitoring multiple data sources with multiple streams, using multiple monitoring algorithms, the unavoidable result is a large inflation in false alarm rate. In the case of biosurveillance the implication of excess false alerts outweighs the benefit of an early detection, as the practical result is that public health users ignore alerts altogether. An excess of false alerts usually leads to the ignoring of true alerts, and potentially to deletion of the alerting system altogether. An example is the rocket-alert system in the city of Ashkelon, Israel that was recently disconnected because of five false alarms in April 2008 which led to panic. Thus, when a rocket fell on a shopping mall in Ashkelon the following month, there was no early warning.[3]

In this chapter we focus on a solution to two important problems: that of multiplicity in monitoring algorithms and the unknown nature of the outbreak signature. We show that by combining results from multiple algorithms in a way that controls the overall false alert rate, we can actually improve overall performance. The remainder of the chapter is organized as follows. Section 2 describes control charts in general and the limitations of applying them directly to raw modern biosurveillance data. It then describes a preprocessing step that is needed before applying control charts. In Sect. 3 we describe an authentic modern biosurveillance dataset and the simulation of outbreak signatures. Section 4 introduces the notion of model combinations in terms of combining residuals and combining control chart output. Section 5 applies the different combination methods to our data, and we display results showing the improvement in detection performance due to method combination. Section 6 summarizes the main points and results and describes potential enhancements.

2. CONTROL CHARTS AND BIOSURVEILLANCE

Control charts (also referred to as monitoring charts) are used to monitor a process for some quality parameter in order to detect anomalies from desired behavior. In the context of modern biosurveillance, control charts are used to monitor aggregated daily counts of individual healthcare seeking behavior (such as daily arrivals to emergency departments or medication sales), for the purpose of early detection of shifts from expected baseline behavior. Three control charts are commonly used to monitor such pre-

[3] http://www.haaretz.com/hasen/spages/983479.html, accessed May23, 2008.

diagnostic daily data, and are implemented (with some variations) in the three main national biosurveillance systems in the U.S.: BioSense (by CDC), ESSENCE (by DoD), and RODS. The three charts are Shewhart, Cumulative Sum (CuSum) and Exponential Weighted Moving Average (EWMA) charts. These control charts are described in detail in Sect. 2.1.

Using control charts to monitor biosurveillance data has two major drawbacks. First, control charts assume that the monitored statistics follow an independent and identically-distributed (iid) normal distribution with constant mean and variance. Daily pre-diagnostic counts usually fail to meet this assumption. In reality time series of such daily counts often contain seasonal patterns, day-of-week effects, and holiday effects (see Figure 8-1 for illustration). Monitoring such data therefore requires an initial processing step where such explainable patterns are removed. Such methods are described in Sect. 2.2. For illustration, compare Figures 8-1 and 8-2 that show a series of daily military clinic visits before and after preprocessing. One explainable pattern that is removed is the day-of-week effect, which is clearly visible in Figure 8-1, but absent from Figure 8-2.

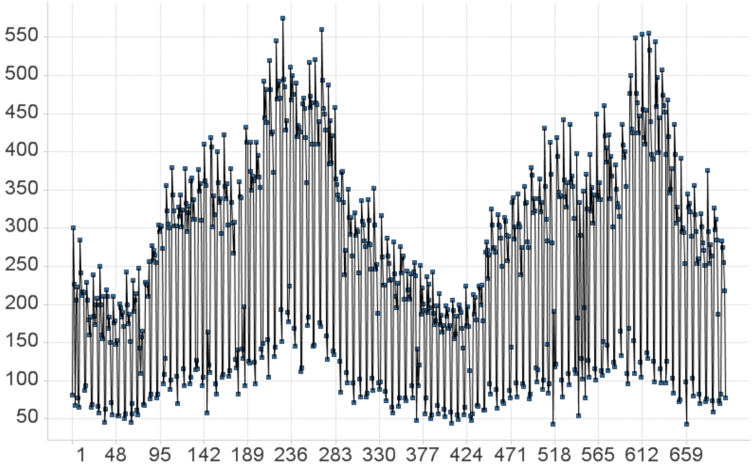


Figure 8-1. Raw series of number of daily military clinic visits with respiratory complaints.

The second challenge of applying control charts in the modern biosurveillance context is that each type of chart is most efficient at capturing a specific outbreak signature (Box and Luceno, 1997). Yet, in the context of biosurveillance the outbreak signature is unknown, and in fact the goal is to detect a wide range of signatures for a variety of disease outbreaks, contagious and non-contagious, both natural and bioterror-related. It is therefore unclear which method should be used to detect such a wide range of unspecified anomalies.

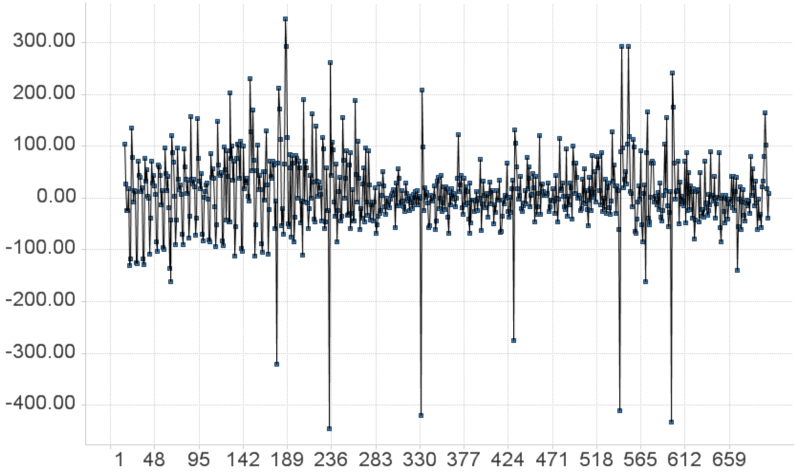


Figure 8-2. Daily military clinic visits series after removing explainable patterns.

2.1 Control Chart Overview

We briefly describe the most popular control charts in statistical quality control, which are widely used in modern biosurveillance systems:

Shewhart. The Shewhart chart is the most basic control chart. A daily sample statistic (such as a mean, proportion, or count) is compared against upper and/or lower control limits (UCL and LCL), and if the limit(s) are exceeded, an alarm is raised. The control limits are typically set as a multiple of standard deviations of the statistic from the target value (Montgomery, 1997). It is most efficient at detecting medium to large spike-type outbreaks.

CuSum. Cumulative Sum (CuSum) control charts monitor cumulative sums of the deviations of the sample statistic from the target value. CuSum is known to be efficient in detecting small step-function type changes in the target value (Box and Luceno, 1997).

EWMA. The Exponentially Weighted Moving Average (EWMA) chart monitors a weighted average of the sample statistics with exponentially decaying weights (NIST/SEMATECH Handbook). It is most efficient at detecting exponential changes in the target value and is widely used for detecting small sustainable changes in the target value.

Table 8-1 summarizes for each of the three charts its monitoring statistic (denoted Shewhart_t, EWMA_t and CuSum_t), the upper control limit (UCL) for alerting, the parameter value that yields a theoretical 5% false alert rate, and a binary output indicator that indicates whether an alert was triggered on day t (1) or not (0). Y_t denotes the raw daily count on day t. We consider one-sided control charts where an alert is triggered only when there is indication of an increase in mean (i.e., when the monitoring statistic exceeds the UCL). This is because only increases are meaningful in the context of healthcare seeking counts.

Table 8-1. Features of three main c*ontrol charts*.

	Shewhart	EWMA	CuSum
Monitored Statistic	Shewhart$_t$ = Y_t	EWMA$_t$ = $\lambda Y_t + (1 - \lambda)\text{EWMA}_{t-1}$	CuSum$_t$ = Max (0, CuSum$_{t-1}$ + $Y_t - \sigma/2$)
UCL	UCL = $\mu + k \times \sigma$	UCL = EWMA$_0$ + $k \times \sigma$ $s^2 = \lambda/(2 - \lambda) \times \sigma^2$	UCL = $\mu + h \times \sigma$
Theoretical 5% Threshold	k = 1.5	k = 1.5	h = 2.5
Output	S$_t$ = if [Shewhart$_t$>UCL]	E$_t$ = if [EWMA$_t$ >UCL]	C$_t$ = if [CuSum$_t$>UCL]

2.2 Preprocessing Methods

There are a variety of methods for removing explainable patterns from time series. Methods generally are either model-based or data-driven. Model-based methods remove a pattern by directly modeling the pattern via some specification. An example is a linear regression model with day-of-week indicators. Data-driven methods either suppress certain patterns (e.g., differencing at a certain lag) or "learn" patterns from the data (e.g., exponential smoothing). In the following we describe three methods that have been shown to be effective in removing the types of explainable effects that are often exhibited in pre-diagnostic daily count series (day-of-week, holiday, seasonal, and autocorrelation). For a more detailed discussion of preprocessing methods see Lotze et al. (2008) and Lotze and Shmueli (2008). In the following we describe three methods that produce next-day forecasts. The forecasts are then subtracted from the actual counts to produce residuals.

Method 1: Holt-Winters exponential smoothing, using smoothing parameter values $\alpha = 0.4$, $\beta = 0$, and $\gamma = 0.15$ as suggested in Burkom et al. (2007). In addition, we do not update the forecasting equation if the percentage difference between the actual and fitted values is greater than 0.5.

Method 2: 7-day differencing (residuals are equal to the difference between the values of the current day and the same day 1 week previous). Equivalently, forecasts are obtained by using the values from 1 week previous.

Method 3: Linear regression of daily counts, using as covariates sine and cosine yearly seasonal terms, six day-of-week dummy variables, and a linear trend term. Only data in the first year are used for parameter estimation (training data).

3. DATA AND OUTBREAKS

3.1 Data Description

Our data is a subset of the dataset used in the BioALIRT program conducted by the U.S. Defense Advanced Research Projects Agency (DARPA) (Siegrist and Pavlin, 2004). The data include six series from a single city, where three of the series are indicators of respiratory symptoms and the other three are indicators of gastrointestinal symptoms. The series come from three different data sources: military clinic visits, filled military prescriptions, and civilian physician office visits. Figures 8-3 and 8-4 display the six series of daily counts over a period of nearly 2 years. We illustrate the methods throughout this chapter using the series of respiratory symptoms collected from military clinic visits (top panel).

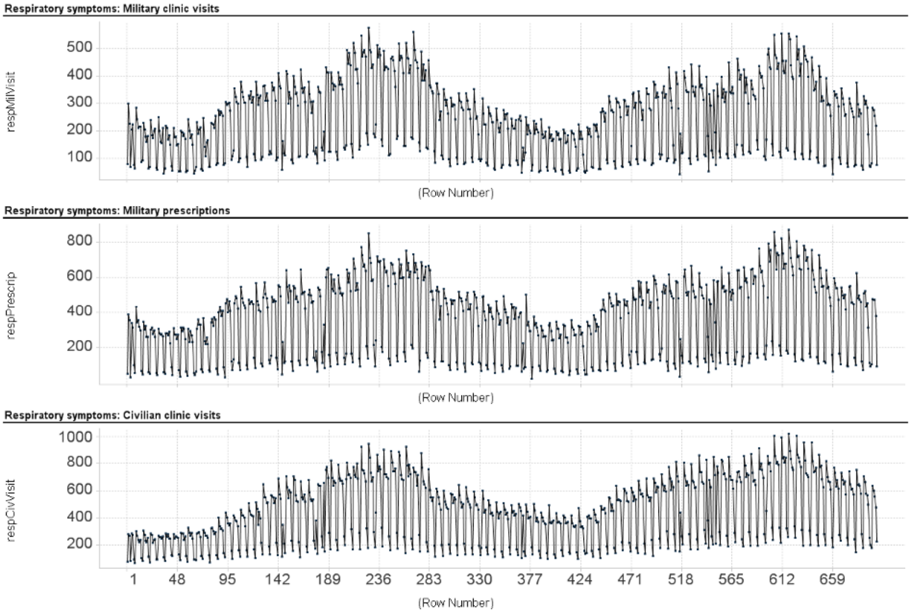


Figure 8-3. Daily counts of military clinic visits (*top*), military filled prescriptions (*middle*) and civilian clinic visits (*bottom*), all respiratory-related.

3.2 Outbreak Signatures

Before preprocessing the raw data, we inject into the raw data outbreak signatures. The insertion into the raw data means that we assume that effects such as day-of-week and holidays will also impact the additional counts due to an outbreak. We simulate two different outbreak signature shapes: a single-day spike and a multiple-day lognormal progression. We set the size of the affected population to be proportional to the variance of the data series (Lotze et al., 2007).

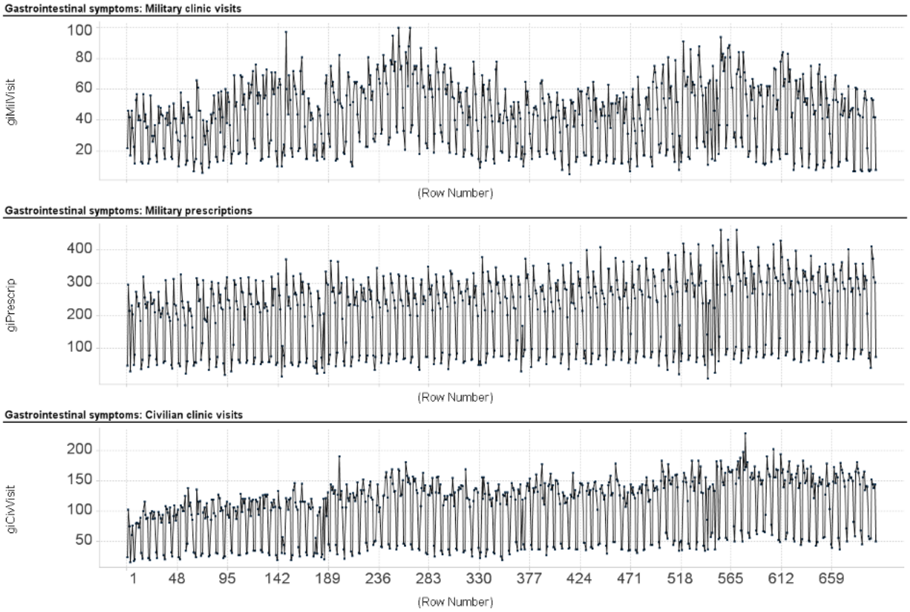


Figure 8-4. Daily counts of military clinic visits (*top*), military filled prescriptions (*middle*) and civilian clinic visits (*bottom*), all gastrointestinal-related.

For the single-day spike, we consider small to medium spike sizes, because biosurveillance systems are designed to detect early, more subtle indications of a disease outbreak. We also consider a lognormal progression signature, because incubation periods have been shown to follow a lognormal distribution with parameters dependent on the disease agent and route of infection. In order to generate a trimmed lognormal signature (Burkom, 2003), we set the mean of the lognormal distribution to 2.5 and the standard deviation to 1. We trim 30% of the long tail, limiting the outbreak horizon to approximately 20 days. This choice of parameters results in a gradually increasing outbreak with a slow fading rate (long tail). Figure 8-5 illustrates the process of injecting a lognormal outbreak into the raw data.

4. COMBINATION MODELS

We consider the problem of linearly combining residuals and/or control chart output vectors for improving the performance of automated biosurveillance algorithms. In order to better evaluate the contribution of each of the two levels of combination, we first examine residual combinations and control chart combinations separately: when combining residuals from different preprocessing techniques, we use a single control chart (see Figure 8-6); when combining control chart outputs we use a single preprocessing technique (see Figure 8-7). We then examine the additional improvement in performance from optimizing the complete process (combining both residuals and control charts).

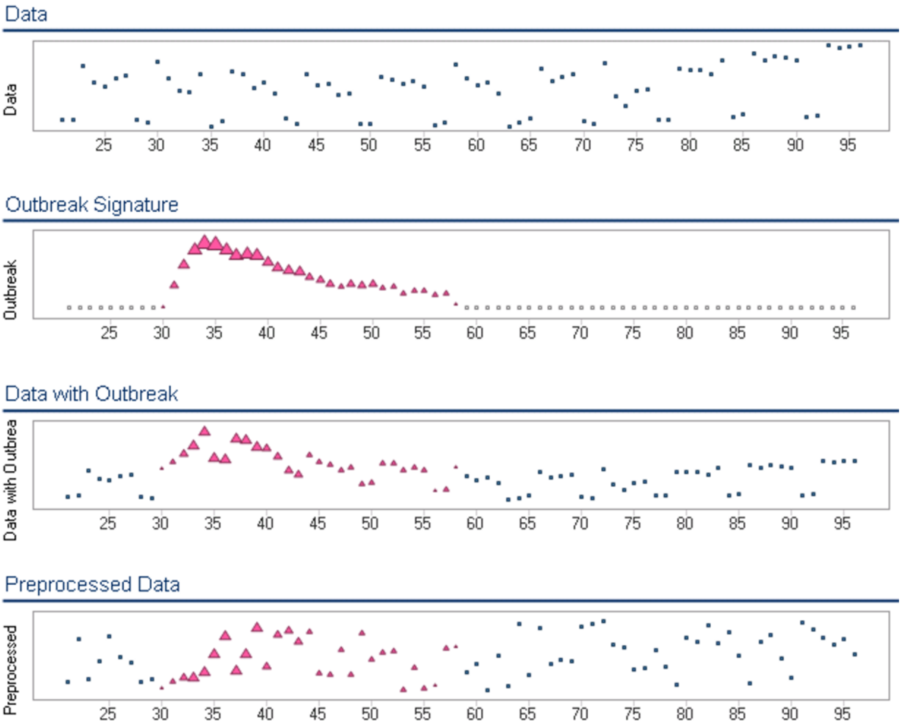


Figure 8-5. Injecting a lognormal outbreak signature into raw data, and preprocessing the resulting series.

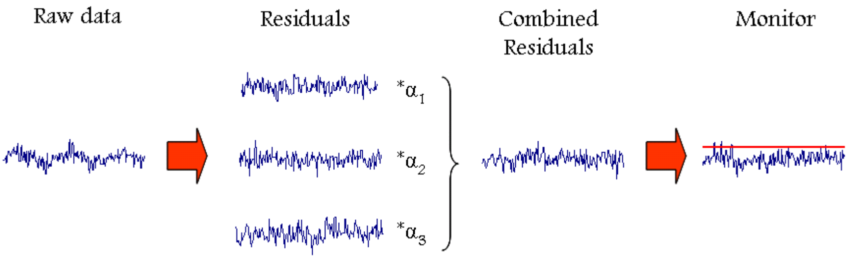


Figure 8-6. Combining residuals.

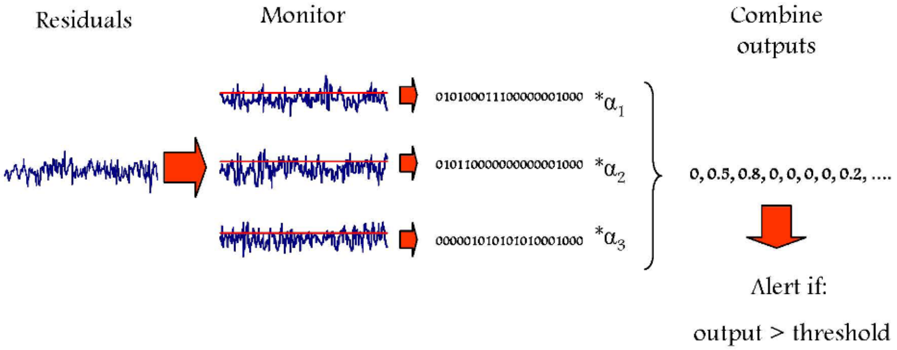

Residuals
Monitor
Combine outputs
01010001110000000001000 *α_1
01011000000000000001000 *α_2
00000101010101010001000 *α_3
0, 0.5, 0.8, 0, 0, 0, 0, 0.2, ...
Alert if:
output > threshold

Figure 8-7. Combining control chart outputs.

We assume that the data are associated with a label vector O_t, which denotes whether there is an actual outbreak at day *t*. We further assume a sufficient amount of training data. The labeled vector and sufficient training data are essential when seeking an optimal combination that increases the true alert rate while maintaining a manageable false alert rate.

4.1 Residual Combination

The idea of using an ensemble is inspired by machine learning classifier techniques, which have produced improved classification by combining multiple classifiers. We used a simple method as our main combination method: a linear combination of next-day forecasts that minimizes the mean squared errors of past data. The coefficients for this linear combination can be determined using linear regression, with the forecasters as predictors and the actual value as the dependent variable. We also compared combinations using a day-of-week modification: residual combinations are optimized separately using only past data from the same day of the week.

4.2 Control Chart Combination

In this section, we assume that the raw data have undergone a preprocessing step for removing explainable patterns. Thus, the input into the control charts is a series of residuals. We consider the three monitoring charts described in Sect. 2: Shewhart, EWMA and CuSum. We construct a linear combination of the monitoring binary output for the purpose of maximizing the true alert rate, while constraining the false alert rate to be below a specific threshold. This formulation yields the following mixed integer programming optimization problem:

$$\max \sum_{t=1}^{n} TA_t$$

$$s.t.$$

$$\begin{array}{ll}
(Bin:) & FA_t, MA_t \in \{0,1\} \\
(FA:) & (w_S \times S_t + w_E \times E_t + w_C \times C_t) \times (1 - O_t) - T < FA_t \\
(TA1) & [(w_S \times S_t + w_E \times E_t + w_C \times C_t) - T] \times O_t \le TA_t \times O_t \\
(TA2) & TA_t \times O_t \le (w_S \times S_t + w_E \times E_t + w_C \times C_t) \times O_t \\
(FA\_sum:) & \sum_{t=1}^{n} FA_t < \alpha \times n
\end{array}$$

where w_i is the weight of control chart *i*, and FA_t (TA_t) is an indicator for a false (true) alert on day *t*. The constraints can be interpreted as follows:

(Bin) restricts the false alert (*FA*) and true alert (*TA*) indicators on day *t* to be binary.

(FA) is a set of n (training horizon) constraints that determine whether the combined output $w_S \times S_t + w_E \times E_t + w_C \times C_t$ yields a false alert on day *t*:

If there is an outbreak on day t, then $1 - O_t = 0$ and the constraint is satisfied.

Otherwise ($1 - O_t = 1$) we compare the combined output with the threshold T = 1. If the combined output is greater than the threshold, we set FA_t to 1.

Similarly *(TA1 and TA2)* is a set of 2n constraints that determine whether the combined output $w_S \times S_t + w_E \times E_t + w_C \times C_t$ yields a true alert on day *t*.

Finally, we set the false alert rate to be less than α (*FA_sum*).

5. EMPIRICAL STUDY AND RESULTS

In this section we describe the results obtained from applying the combination methods to authentic pre-diagnostic data with simulated outbreaks. We start by describing the experimental design and then evaluate the methods' performance.

5.1 Experiment Design

We inject into the raw series 100 outbreak signatures, in random locations (every 10 weeks on average). Each outbreak signature is a spike of size 0.5 × σ (~60 cases), with probability 0.6, and a trimmed lognormal curve of height 5 × σ (~450 cases). The peak of the lognormal curve is typically on the fifth or sixth day. We inject a mixture of the two outbreak signatures to illustrate the robustness of the algorithm combination. We repeat this test setting 20 times.

When combining control charts, the desired false alert rate is varied in the range $\alpha \in \{0.01, 0.05, 0.1, 0.2\}$. We set the threshold of the monitoring charts to meet the desired overall false alert rate α, using 1 year of training data (referred to as the *experimental threshold*).

5.2 Results

5.2.1 Residuals Combination

In this section we compare four preprocessing methods, two simple methods and two combinations. The simple methods are Holt-Winters' exponential smoothing and linear regression. The two combination methods are a combination of Holt-Winters and linear regression residuals and a day-of-week variant of this combination. We then monitor each of the residual series with a Shewhart control chart. Because the preprocessing methods use different lengths of training data, we use the remaining dataset to compute the threshold that gives us the α = 0.05 false alert rate. The resulting true alert rate is shown in Figure 8-8, which displays the true alert rate distribution

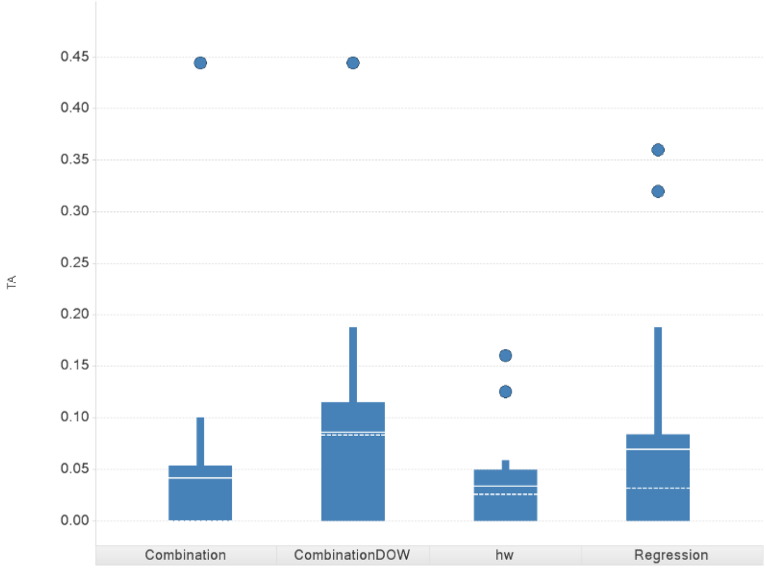


Figure 8-8. True alert rate distribution for Holt-Winters, linear regression, their combination, and a day-of-week combination. Means are marked by *solid white lines* and medians by *dashed white lines*.

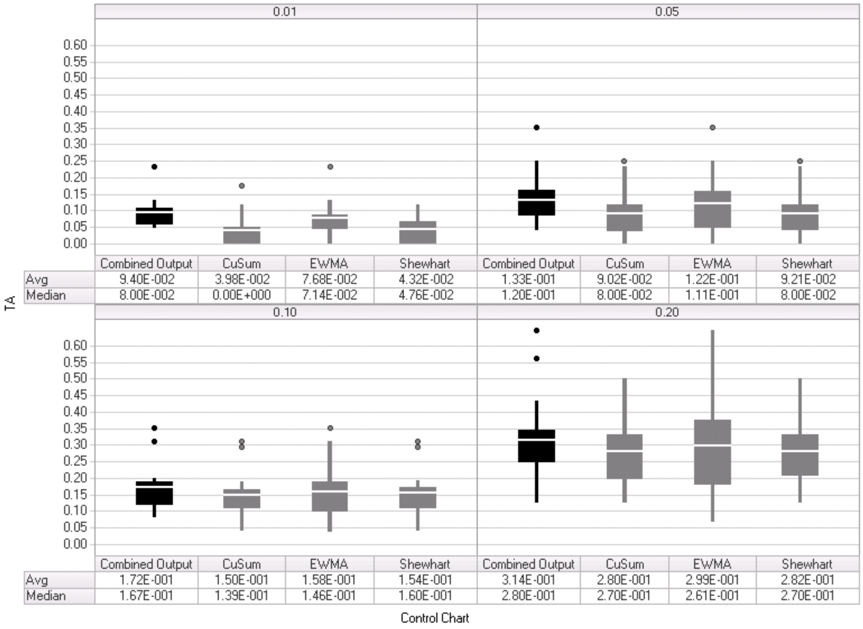
for each of the four methods. Means and medians are marked as solid and dashed white lines, boxes correspond to the inter-quartile range, and the lines extend between the fifth and 95th percentiles.

The figure depicts the advantage of the day-of-week combined preprocessing, which has the highest mean, median and 75th percentile.

5.2.2 Control Chart Combination

We start by preprocessing the raw series using Holt-Winters exponential smoothing. Control charts (Shewhart, EWMA, CuSum, and the combined output) are then used to monitor the series of residuals. Finally, we calculate the false and true alert rates produced by each method. For the lognormal outbreak signature, we consider a true alert only if the alert took place before the peak, because timely alerting plays an important role in diminishing the spread of a disease.

In the first experiment we optimize the control chart combination separately for each of the 20 tests. Figure 8-9 depicts the results of this experiment. The different panels correspond to different levels of experimental threshold α. Each panel shows the true alert rate distribution for each of the four methods. The results clearly show the advantage of the combined method in terms of both increasing true alert rate, subject to a given false alert rate, and in reducing the variance of the true alert rate.



	0.01				0.05			
	Combined Output	CuSum	EWMA	Shewhart	Combined Output	CuSum	EWMA	Shewhart
Avg	9.40E-002	3.98E-002	7.68E-002	4.32E-002	1.33E-001	9.02E-002	1.22E-001	9.21E-002
Median	8.00E-002	0.00E+000	7.14E-002	4.76E-002	1.20E-001	8.00E-002	1.11E-001	8.00E-002

	0.10				0.20			
	Combined Output	CuSum	EWMA	Shewhart	Combined Output	CuSum	EWMA	Shewhart
Avg	1.72E-001	1.50E-001	1.58E-001	1.54E-001	3.14E-001	2.80E-001	2.99E-001	2.82E-001
Median	1.67E-001	1.39E-001	1.46E-001	1.60E-001	2.80E-001	2.70E-001	2.61E-001	2.70E-001

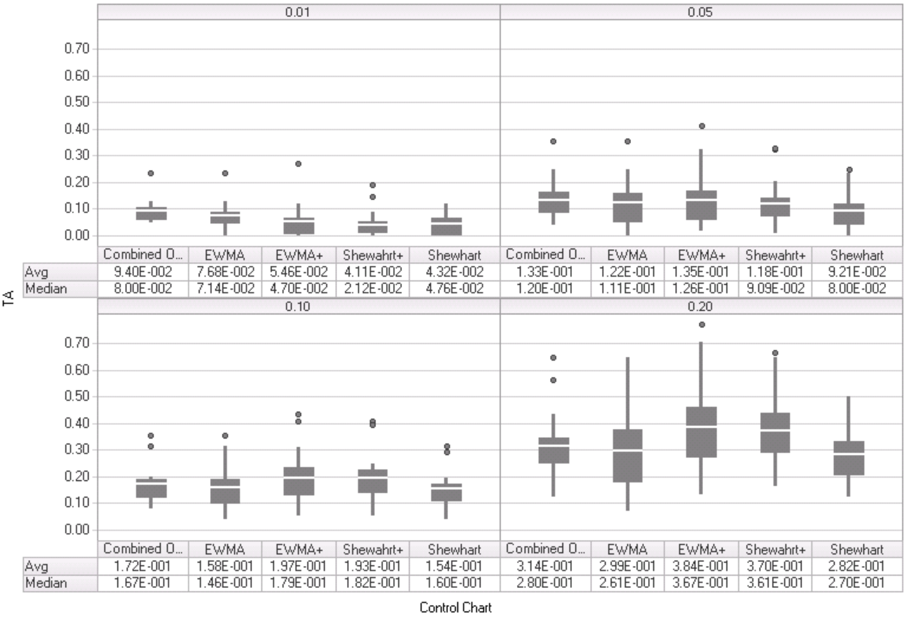
Figure 8-9. True alert rate distribution for three control charts and their combination, by false alert rate (α = 0.01, 0.05, 0.10, 0.20). Means are marked by *solid white lines*.

The main drawback of the first experiment is that the computation is very time consuming. Since achieving the optimal weights for the control charts is an NP complete problem, computation time increases exponentially in the length of the training data. Moreover, examining the actual weights shows that EWMA and Shewhart charts dominate the combination such that alerts are mostly determined by one of them (e.g., Shewhart) combined with an alert by one other method (e.g., either EWMA or CuSum). In an effort to reduce computation time, yet seek for good combinations, we take a hybrid approach: we choose among a small set of pre-determined combinations that appear to work well. This approach greatly reduces computation time and allows for real-time computation in actual settings.

Based on the general results found in the first experiment for the optimal weights, in the next experiment we chose two settings of pre-set weights:

1. Shewhart +: The algorithm signals an alert at time t if the Shewhart statistic signals an alert, *and* at least one other chart signals an alert.
2. EWMA +: The algorithm signals an alert at time t if the EWMA statistic signals an alert, *and* at least one other chart signals an alert.

The resulting true alert rates are shown in Figure 8-10. We observe that for a very low experimental false alert rate threshold (α = 0.01) the two new combination charts (Shewhart + and EWMA +) do not perform as well as the individual Shewhart and EWMA charts. However, when the experimental



	Combined O...	EWMA	EWMA+	Shewahrt+	Shewhart	Combined O...	EWMA	EWMA+	Shewahrt+	Shewhart
Avg	9.40E-002	7.68E-002	5.46E-002	4.11E-002	4.32E-002	1.33E-001	1.22E-001	1.35E-001	1.18E-001	9.21E-002
Median	8.00E-002	7.14E-002	4.70E-002	2.12E-002	4.76E-002	1.20E-001	1.11E-001	1.26E-001	9.09E-002	8.00E-002

	Combined O...	EWMA	EWMA+	Shewahrt+	Shewhart	Combined O...	EWMA	EWMA+	Shewahrt+	Shewhart
Avg	1.72E-001	1.58E-001	1.97E-001	1.93E-001	1.54E-001	3.14E-001	2.99E-001	3.84E-001	3.70E-001	2.82E-001
Median	1.67E-001	1.46E-001	1.79E-001	1.82E-001	1.60E-001	2.80E-001	2.61E-001	3.67E-001	3.61E-001	2.70E-001

Figure 8-10. True alert rate distribution for select combinations, by false alert rate (α = 0.01, 0.05, 0.10, 0.20).

false alert rate threshold is higher (α = 0.05) the new charts perform at least as well as the ordinary charts, and even outperform the optimal combination (based on training data) when α > 0.05. None of the methods violated the experimental false alert rate threshold by more than 10% when α = 0.01, and 3% when α ≥ 0.05.

5.2.3 Combining Residuals and Monitoring

After examining the combination of residuals separately from combining control chart outputs, we now examine the effect of using combined pre-processing methods monitored by combined control charts on detection performance. The false alert rate is set to α = 0.05 and we observe the resulting true alert rate. We compare the performance of the different pre-processing methods monitored by either Shewhart, or Shewhart+. Figure 8-11 presents the resulting true alert rate distributions for each of the different combinations. We see that using Shewhart+ increases the true alert rate by approximately 50% compared to Shewhart. Also, the day-of-week residual combination outperforms the alternative preprocessing methods. The best performance is obtained by using both the day-of-week residual combination and applying the Shewhart+ to monitor the series. Thus, method combination provides improved detection at each of the preprocessing and monitoring levels, as well as when they are combined.

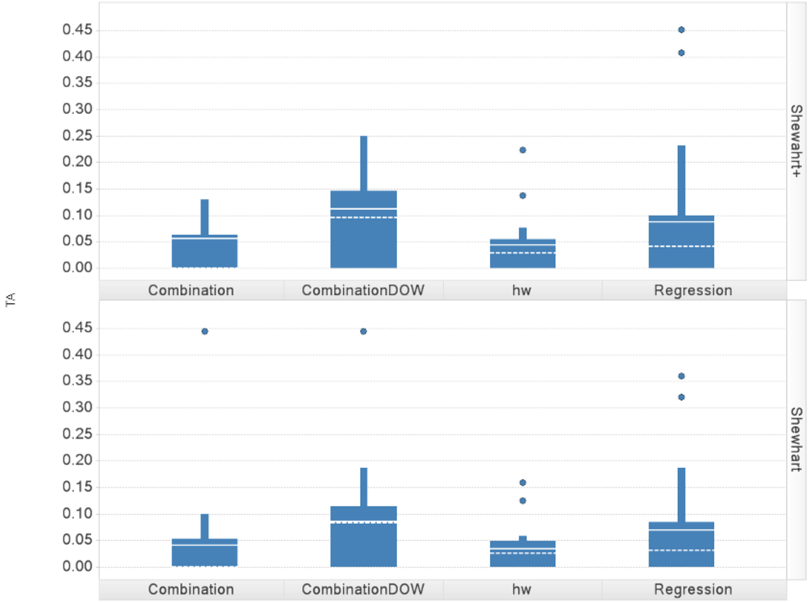


Figure 8-11. True alert rate distribution when combining residuals and control chart outputs (*top panel* is Shewhart +, *bottom panel* is Shewhart).

6. CONCLUSIONS

In this chapter we propose methods for improving the detection performance of univariate monitoring of non-stationary pre-diagnostic data by combining operations at each of the two stages of the outbreak detection task, data preprocessing and residual monitoring, for the purpose of increasing true alert rate for a given false alert rate.

Improved performance by combining control chart output is achieved by formulating the true alert/false alert tradeoff as a mixed integer programming problem (MIP). The MIP enables us to find the weights that optimize the combination method. To decrease computation time we take a hybrid approach where the weight optimization is carried out over a restricted set of combinations, which is obtained from a training stage. We show that the hybrid approach still provides improved performance. Our empirical experiments confirm the advantage of this portfolio approach in each of the stages (preprocessing and monitoring) separately and in the mixture of both.

Future extensions include adaptive combinations, where the weights of each method change dynamically over time, based on more current *history*. Another extension is using machine learning methods that automatically adjust combination weights based on current and recent *performance*, and on the most recent weight vector.

ACKNOWLEDGEMENTS

The research was partially funded by NIH grant RFA-PH-05-126. Permission to use the data was obtained through data use agreement #189 from TRICARE Management Activity. For the second author, this research was performed under an appointment to the U.S. Department of Homeland Security (DHS) Scholarship and Fellowship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. All opinions expressed in this chapter are the authors and do not necessarily reflect the policies and views of DHS, DOE, or ORAU/ORISE.

QUESTIONS FOR DISCUSSION

1. What is the role of data preprocessing in modern biosurveillance data? What are its limitations? How does preprocessing affect outbreak signatures in the data?

2. What are the requirements for combination methods?
3. How can the notion of combinations be extended for monitoring multiple time series? When would it be beneficial?
4. Discuss advantages and disadvantages of adaptive combinations, where the weights change over time.
5. What are other applications in which combination methods can be beneficial?

REFERENCES

Box, G., Luceno, A. (1997) *Statistical Control: By Monitoring and Feedback Adjustment.* 1st ed. Wiley-Interscience, London.

Burkom, H. S. (2003) "Development, adaptation and assessment of alerting algorithms for biosurveillance", *Johns Hopkins APL Technical Digest* 24(4), 335–342.

Burkom, H. S., Murphy, S. P., and Shmueli, G. (2007) "Automated time series forecasting for biosurveillance", *Statistics in Medicine* 26(22), 4202–4218.

Lotze, T. and Shmueli, G. (2008) "Ensemble forecasting for disease outbreak detection", *23rd AAAI Conference on Artificial Intelligence, Chicago July 08.*

Lotze, T., Shmueli G. and Yahav, I. (2007) Simulating Multivariate Syndromic Time Series and Outbreak Signatures, *Robert H. Smith School Research Paper No. RHS-06–054*, Available at SSRN: http://www.ssrn.com/abstract=990020.

Lotze, T., Murphy, S. P., and Shmueli, G. (2008) "Preparing biosurveillance data for classic monitoring", *Advances in Disease Surveillance* 6. http://www.isdsjournal.org/article/view/516

Montgomery, D.C. (1997) *Introduction to Statistical Quality Control*. 3rd ed. Wiley, London.

NIST/SEMATECH: E-handbook of statistical methods. http://www.itl.nist.gov/div898/handbook.

Siegrist, D. and Pavlin, J. (2004) Bio-ALIRT biosurveillance detection algorithm evaluation. *MMWR*, 53(Suppl), 152–158.

Shmueli, G. and Burkom, H. S. (2010) "Statistical Challenges Facing Early Outbreak Detection in Biosurveillance", *Technometrics* vol 52(1), 39–51.

ONLINE RESOURCES

http://www.projectmimic.com: R code for simulating or mimicking multivariate time series of biosurveillance data and outbreak signatures. Also includes semi-authentic data.

Chapter 9

MODELING IN SPACE AND TIME

A Framework for Visualization and Collaboration

DANIEL A. FORD[1,*], JAMES H. KAUFMAN[1], and YOSSI MESIKA[2]

CHAPTER OVERVIEW

This chapter describes the Spatiotemporal Epidemiological Modeler (STEM), now being developed as an open source computer software system for defining and visualizing simulations of the spread of infectious disease in space and time. Part of the Eclipse Technology Project, http://www.eclipse.org/stem, STEM is designed to offer the research community the power and extensibility to develop, validate, and share models on a common collaborative platform. Its innovations include a common representational framework that supports users in creating and configuring the components that constitute a model. This chapter defines modeling terms (canonical graph, decorators, etc.) and discusses key concepts (e.g., labels, disease model computations). Figures illustrate the types of visualizations STEM provides, including geographical views via GIS and Google Earth™ and report generated graphics.

Keywords: Open source tools; Modeling; Visualization; Infectious disease transmission; Collaboration

[1*] *Healthcare Informatics Research, Department of Computer Science, IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA*, daford@almaden.ibm.com

[2] *Healthcare and Life Sciences, IBM Haifa Research Lab, Haifa University Campus, Mount Carmel, Haifa 31905, Israel*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_9,

1. INTRODUCTION

In 2006, an article in *Science* concluded that "our knowledge of influenza transmission is incomplete, and more basic data are needed to make models accurate and to give them predictive weight" (Ash and Roberts, 2006). The "startling spread" of the Avian Influenza underscored the need to "build a long-lasting international infrastructure to monitor and thwart threats from such emerging infections." The development of the Spatiotemporal Epidemiological Modeler (STEM), then underway and in its second phase as of this writing, is an effort to create such an infrastructure.

A computer software system for defining and visualizing simulations of the spread of infectious disease in space and time, STEM can track the progress of specified infectors across designated populations. Its ultimate goal is to provide a common platform with the power and extensibility to develop, validate, and share models. As an open source system, STEM allows researchers to leverage the work of others by re-using components developed by their peers as well as offering their own work up for others to re-use. By virtue of its design and its unique framework, STEM can reduce duplication, allow models to be compared and refined, support collaborative efforts to understand disease transmission in a global society, and ultimately help protect the health of all populations around the world.

2. MODELING: AN OVERVIEW

In January 2006, we described how STEM could be used to track Avian Flu among humans and ducks in Thailand and display the progress of the disease in a series of maps (Ford et al., 2006). In that article, we set forth our vision of STEM as an open source framework to support sharing across the research and public health communities.

Interest in modeling has intensified since January 2006, when our article was published in the *International Journal of Health Geographics*. A literature search on PubMed in May 2008 listed 118 "related articles," 85 of them published in 2006 or later. Still, only a subset presented epidemiological models, and most of them tended to focus on a single population type affected by a single disease. Or they assumed that the population being affected by a disease was "well mixed" and either not distributed (zero dimensional simulation) or geographically distributed in a uniform manner. Such approaches represent complex realities only incompletely. Moreover, they do not purport to offer a common platform that allows researchers to share components for creative re-use, making it possible for them to collaborate in understanding and solving complex problems.

Today we remain committed to STEM as a framework upon which disease models can be developed, simulated, and shared by many different researchers. This chapter will focus on that framework and illustrate how STEM's capabilities allow users to develop models and visualize the spread of disease across time and space.

3. ABOUT STEM

STEM is being developed as an open source software system and is part of the Eclipse open source application development framework whose development is being managed by the Eclipse Foundation, http://www.eclipse.org. It is a subproject of the Eclipse Technology Project, http://www.eclipse.org/stem.

The Eclipse framework offers many advantages to Java™ based software development projects such as STEM. Eclipse is based on the industry standard Open Services Gateway Initiative (OSGi, http://www.osgi.org) software component architecture. In fact, Eclipse serves as the reference implementation of the standard. The OSGi standard partitions software into distinct components, known as *bundles* or *plug-ins.* These can be independently developed, deployed, and managed. These characteristics and the concept of declarative software extension points, also a feature of Eclipse, enable a software system like STEM to be easily extended by the addition of new plug-ins.

Designed from the start to be extensible, STEM's implementation is such that its main components, the core representational framework, graphical user interface (GUI), simulation engine, disease model computations, and the various data sets, are all partitioned into separate plug-ins. This makes it relatively straightforward for a third party to augment or replace any one of the system's components. For instance, STEM can be extended by supplying additional disease models, enhanced geographic or other visualizations, or alternative data sets. Another advantage of the Eclipse framework is that it offers true platform independence. Currently, versions of STEM are available for the Microsoft Windows, Apple Mac, and Linux operating systems; versions for other operating systems, supported by Eclipse, will be available as demand dictates.

As part of Eclipse, STEM is distributed under the Eclipse Public License (EPL), http://www.eclipse.org/legal/epl-v10.html. This license is attractive to the STEM project because it is fairly unrestrictive, allowing re-use of the code and even commercial exploitation. These traits lend themselves to furthering the altruistic goals of STEM by enabling others to exploit the code base for their own purposes. This in turn attracts a community of developers, users, and resources that support the system.

STEM includes all of the data sets, mathematics, and visualizations required to develop sophisticated simulations of disease spread over geographic regions (countries, continents) or even the entire planet. The *data sets* distributed with STEM include the geography, transportation systems (e.g., roads, air travel), and population for the 244 countries and dependent areas defined by the International Standards Organization.[1] The resolution of the data sets covers most countries, with some exceptions, to administrative level 2. In the United States, level 2 corresponds to counties.

The disease modeling *mathematics* built into STEM are standard textbook implementations of disease models: a SIR model, with states of *susceptible*, *infectious*, and *recovered*; a SEIR model, that adds an *exposed* state; and a SI model, with the *recovered* state removed. Difference equations compute the number of population members entering and leaving the specified disease states in a particular interval of time, using stochastic or deterministic computations. These models are provided both as examples and as base classes that can easily be extended to simplify the task of creating new state-of-the-art models. STEM also provides examples of how to add such extensions.

To support users in developing and refining models, STEM provides two main types of *visualizations*. Geographic displays represent disease spread on maps provided by STEM; alternately, output can be viewed on Google Earth™. Report generation displays plot simulation data values with respect to time or other data values.

3.1 A Common Collaborative Framework

As an epidemiological modeling system, STEM's main innovation is its framework that partitions a disease model into constituent components that represent different aspects of a disease model. For instance, there are components that represent disease model computations, geography, population data, and the passage of time. The key feature of STEM's innovation is that it includes a component that represents the grouping or aggregation of other components. This allows the basic individual components to be combined in creative ways to make useful groupings that can be shared and reused. This group component is called a *model* in STEM, and it can group other models as well as basic components.

The ability of one STEM model to contain another model is incredibly powerful. It allows for the creation of detailed subcomponents whose complexity is hidden within the confines of their encompassing models. For instance, a researcher can create a model that represents a particular country. This model may itself contain complex submodels that define such things as the country's geography and its transportation infrastructure. The transportation

[1] These are set forth in ISO 3166–1 Geographic Coding Standard: Codes for the Representation of Names of Countries and Their Subdivisions - Part 1.

infrastructure submodel in turn may contain complex submodels for different types of transportation systems (e.g., air travel, rail, road, etc.).

Using STEM, researchers modeling the spread of a disease in a particular country can create their own STEM model that combines the model of the country with population demographics, for instance. The details of the country model can be included in their simulations simply by referencing the country submodel. The design goal of STEM is that the country submodel will be a standardized community resource that is maintained and refined by many different contributors. Over time, the country model will become more accurate, detailed, and valuable. As its use and support increase within the community, it will become easier to compare and share different models because their underlying components will be the same.

As a common collaborative platform, STEM enables the import and export of models so that they can be easily shared among researchers. A researcher who has developed a detailed country model that includes population demographics can import a component with specialized disease mathematics from another researcher and combine the two. The new combination can then even be re-exported for still others to leverage. Because this type of collaboration is a key design point of STEM, the system attaches descriptive metadata to each component that conforms to the *Dublin Core Metadata* standard, http://www.dublincore.org, developed by library science researchers. This standard defines a set of attributes that specify such things as the title, description, and version of the component, as well as identifying who created it and when. This metadata allows the research community, and individual researchers, to identify, understand, and trust (or not) each of the components they are using.

3.2 A Common Representational Framework

To enable the goal of creating a common collaborative framework, the design of STEM needs to have a powerful representational framework. This framework must be such that instances can be assembled from disparate components and it must be able to represent the state of any type of arbitrary simulation one might envision. The representational framework used by STEM is a *graph*. A graph is a powerful mathematical abstraction for representing entities (i.e., things in the world) and their relationships (Widgren, 2004; Myers et al., 2003; Gross and Yellen, 2003). More formally, a graph is a set of *nodes*, *edges*, and *labels*. Nodes generally correspond to entities, while edges form a link between two nodes and represent some relationship between them. Labels are attached to either a node or an edge and represent some aspect of their host, such as the name of the entity or of the relationship. Each node may have more than one label, but each edge will only have one.

In STEM, nodes typically represent geographic regions while edges represent relationships between regions. There may be any number of edges between any two nodes, reflecting the fact that in the real world there may be any number of relationships between any two geographic locations. For instance, an edge between two nodes might represent a relationship such as the sharing of a common border (i.e., two regions are physically adjacent and could easily exchange population members); a different edge between the same two nodes might represent a road that connects the two locations. Completely different edges could also exist, for instance, ones that represent human air travel or even the flight paths of migratory birds. A label on a node might represent the physical area of the corresponding geographic region, the number of population members of a particular type who live there, or a mathematical representation of the state of a particular disease at that location. In case of a border edge, the label might record the length of the border between the two regions; in the case of the road edge, the label might indicate the type of road and how much traffic it carries between the two regions.

The graph that includes all of the nodes, edges, and labels necessary to represent the state of the simulation is called the *canonical graph*. It is the canonical graph that is constructed from the contributed components that are aggregated together by a STEM model. Each component is either a *graph* that contains a collection of nodes, edges, and/or labels, or is another submodel that produces its own canonical graph of its contents. The canonical graph of a model, thus, is the union of the canonical graphs of its submodels and the contents of any graphs it contains. When this collection is combined, references are resolved such that edges connect geographical regions with nodes and labels attach themselves to their targets. The composed model hierarchy has a single model at the root; it is the canonical graph of this model that is used as the canonical graph of the simulation, as illustrated in Figure 9-1.

In STEM, a *scenario* serves as the specification of a simulation. It references a single model which will be used to create a canonical graph and binds together information that specifies the time steps of the simulation and initialization. Figure 9-1 shows two scenarios referencing the same model instance. This illustrates the situation where two simulations are to be run using the same model, but with different initialization. For instance, the point of infection could be different between the two simulations; everything else would be the same.

In addition to containing graphs, which basically just hold data, and other submodels, a model in STEM can contain computational specifications as well. This is the mechanism through which disease model computations are included in a simulation. The specification of a "unit" of computation is called a *decorator*. There may be any number of decorators in a simulation, and they have two particular functions. The first is to initialize (or "decorate")

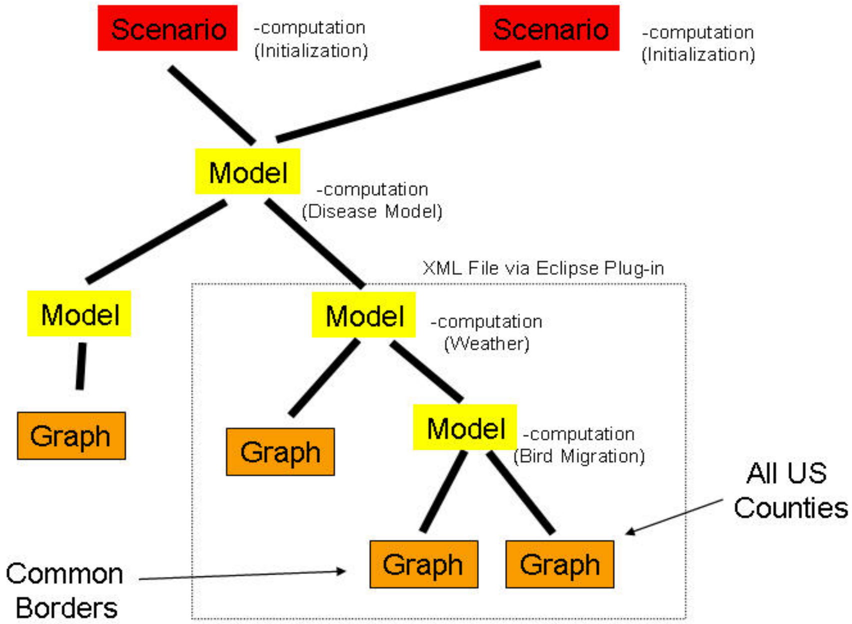


Figure 9-1. STEM components can be combined to compose a model that is used by two different scenarios.

the canonical graph; and the second is to compute the next state of the graph at each simulation cycle. A scenario may also contain decorators, the purpose of which is to perform scenario specific initialization, such as inserting infected individuals into the canonical graph. The order in which decorators are invoked to perform initialization and on each cycle strictly respects the model hierarchy, with the scenario decorators always being last. Thus, when a canonical graph is being initialized, the scenario decorators always have the last opportunity to perform modifications to complete initialization of the canonical graph.

A simulation in STEM begins with an *initialized* canonical graph and a starting *time*. The first step in a simulation is to determine the next point in time that will be used to update the state of the canonical graph. Once this value is determined, the internal STEM simulation engine invokes, in specific order, the decorators associated with the canonical graph. These computations take the time point as input and compute the next state of the graph as it will be at that future time. When these computations are complete, the state of the entire graph is changed to the next value just computed. This process continues until stopped by the user, or, if specified, a predetermined end-time is reached.

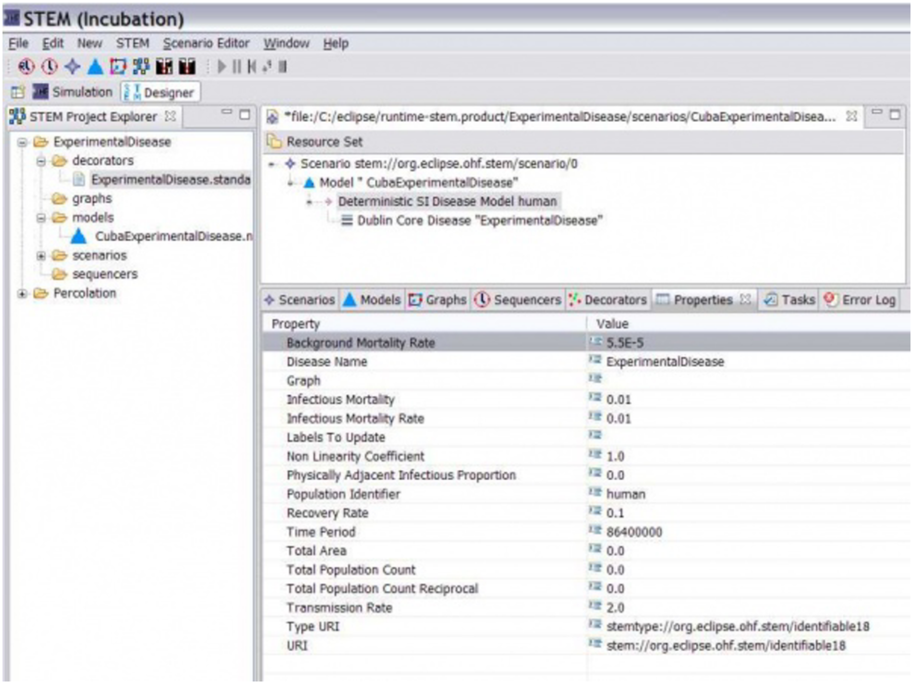
The sequence of time intervals that will be used in a simulation is specified in STEM by a *sequencer*. During a simulation, the sequencer instance associated with the simulation is referenced at the beginning of each cycle for the time to use for that cycle.

Each scenario references a single sequencer, a single model, and a set of decorators, and completely defines a simulation.

The last component of STEM to discuss is an *experiment*. An experiment is a specification of a collection of simulations that are based on a single specific *base* scenario that is modified slightly for each simulation. This allows a researcher to vary one or more parameters – for instance, the infectious rate of a disease model or the initial number of infectious individuals – and examine how the model is affected.

3.3 Creating and Configuring Components

STEM's user interface, shown in Figure 9-2, allows researchers to easily create, configure, and compose these components. To create a component, a user invokes a "wizard" that gives prompts for the details required for the



```
ExperimentalDisease
  decorators
    ExperimentalDisease.standa
  graphs
  models
    CubaExperimentalDisease.n
  scenarios
  sequencers
Percolation
```

```
*file:/C:/eclipse/runtime-stem.product/ExperimentalDisease/scenarios/CubaExperimentalDisea...
Resource Set
  Scenario stem://org.eclipse.ohf.stem/scenario/0
    Model " CubaExperimentalDisease"
      Deterministic SI Disease Model human
        Dublin Core Disease "ExperimentalDisease"
```

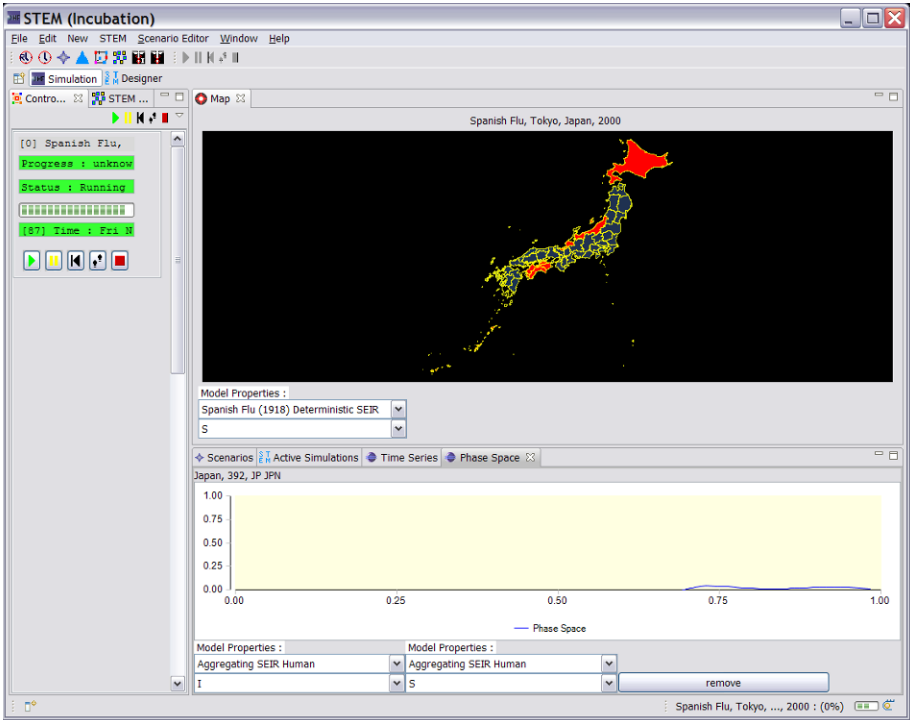
Property	Value
Background Mortality Rate	5.5E-5
Disease Name	ExperimentalDisease
Graph	
Infectious Mortality	0.01
Infectious Mortality Rate	0.01
Labels To Update	
Non Linearity Coefficient	1.0
Physically Adjacent Infectious Proportion	0.0
Population Identifier	human
Recovery Rate	0.1
Time Period	86400000
Total Area	0.0
Total Population Count	0.0
Total Population Count Reciprocal	0.0
Transmission Rate	2.0
Type URI	stemtype://org.eclipse.ohf.stem/identifiable18
URI	stem://org.eclipse.ohf.stem/identifiable18

Figure 9-2. The Designer Perspective of STEM is a drag and drop interface useful for composing the graph that defines a scenario. The screen shot shows a scenario for modeling an "experimental" disease model in Cuba. The model for Cuba as well as the disease has been added to the scenario. The property editor shows details about the disease model as well as the Dublin Core (not visible in the image).

component's creation. Existing components can be edited with customized editors that make it easy to modify the component's configuration (e.g., change a population value). Model composition is accomplished using the editors as well by "dragging and dropping" iconic component representations into the specific editor.

The configuration interface allows a user to specify a simulation scenario, selecting a geographic model, one or more population types, and one or more disease models. The disease models can be configured by having their parameters altered from their (reasonable) default values. The specification of initial conditions for the simulation such as the "seeding" of infectious population members at locations in the model is also part of the interface.

When the configuration is complete, the user moves on to the Simulation Perspective, a display that shows several views of the model and the state of the simulation. The distribution of a population is illustrated by highlighting a geographic region using intensity to indicate relative numbers and color to indicate disease state. Figure 9-3 illustrates how STEM represents a disease



[0] Spanish Flu,
Progress : unknow
Status : Running
[87] Time : Fri N

Spanish Flu, Tokyo, Japan, 2000

Model Properties :
Spanish Flu (1918) Deterministic SEIR
S

Japan, 392, JP JPN

Model Properties :
Aggregating SEIR Human
I

Model Properties :
Aggregating SEIR Human
S

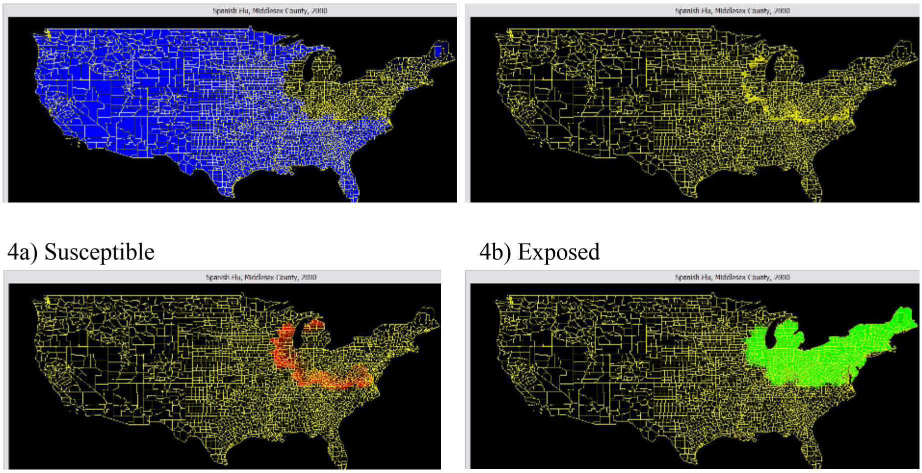
Spanish Flu, Tokyo, ..., 2000 : (0%)

Figure 9-3. The STEM simulation perspective showing a visualization of an infectious disease in Tokyo, Japan. This figure shows a map view representing the "current state" of an outbreak. This GIS view is the default view provided included in STEM.

state visually, using a GIS representation. This representation may be adjusted using either logarithmic or linear intensity scales using a varying gain factor. After the simulation is started, the simulation display updates itself after each time step and displays the current state information. Letting the simulation run animates the display so that the progression of a disease can be watched in "real time." Users can also run multiple simulations in batch mode as an experiment to compare results and gain insights into the impact of different variables on transmission.

3.3.1 Labels

In the compositional modeling framework, *labels* play a special role in that they can store two "state" values simultaneously. They have a "current" value which, collectively, records the current state of the graph. They can also have a "next" value which is used, collectively, to store the next state of the graph. For example, labels on nodes can represent different (diseased) states of a population. The fraction of people infectious or recovered from a disease might be represented by a dynamic label, the value of which changes in time. Figure 9-4 shows how the STEM map view can be used to visualize different disease states using dynamic labels.


Spanish Flu, Middlesex County, 2000
Spanish Flu, Middlesex County, 2000
Spanish Flu, Middlesex County, 2000
Spanish Flu, Middlesex County, 2000

4a) Susceptible

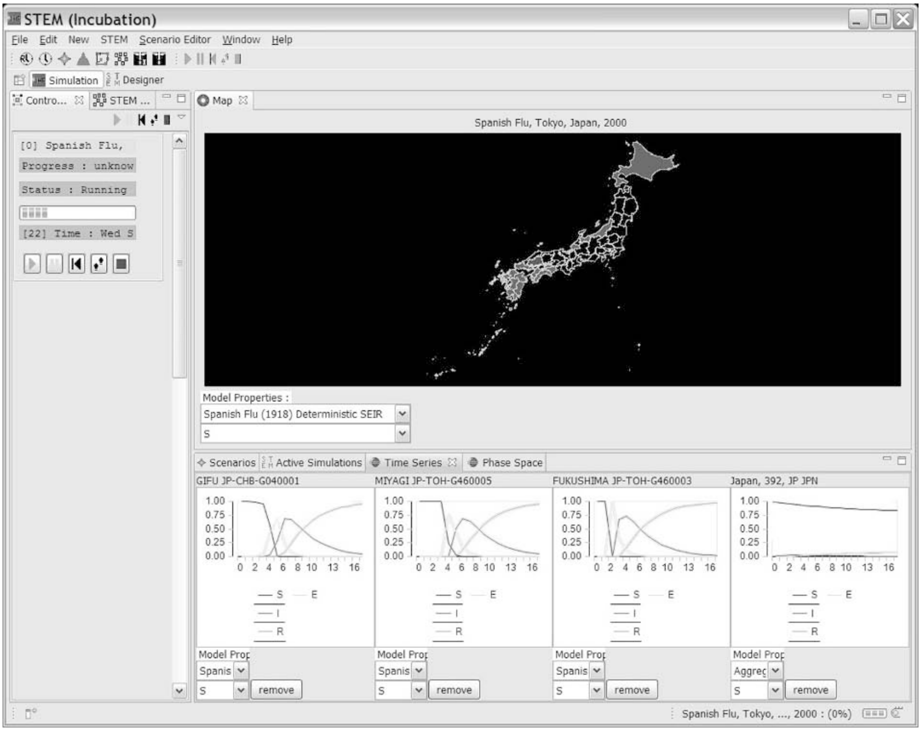
4b) Exposed

4c) Infectious

4d) Recovered

Figure 9-4. Visualizing labels. This figure represents the fraction of people in each of the Susceptible State (**a**), Exposed State (**b**), Infectious State (**c**), and Recovered State (**d**) at the same instant in time during a simulation of an infectious disease.

Other label values can be visualized and monitored in a similar way (or logged to a file). Since label values on nodes (and edges) may serve both as input to and output from a model, it is important to provide appropriate tools to visualize the inputs to and progress of a simulation. In addition to providing a visual map, it is also desirable to graph the dynamic label values and to do quantitative analysis. Figure 9-5 shows the same simulation of Japan shown in Figure 9-3 but now a time series view has been activated allowing users to monitor the disease state variables as a function of time. Other report views provide the trajectory of these dynamic variables in phase space (for example, plotting I(t) vs. S(t)).



[0] Spanish Flu,
Progress : unknow
Status : Running
[22] Time : Wed S

Spanish Flu, Tokyo, Japan, 2000

Model Properties : Spanish Flu (1918) Deterministic SEIR
S

GIFU JP-CHB-G040001 | MIYAGI JP-TOH-G460005 | FUKUSHIMA JP-TOH-G460003 | Japan, 392, JP JPN

Model Prop Spanis S | Model Prop Spanis S | Model Prop Spanis S | Model Prop Aggreg S

Spanish Flu, Tokyo, ..., 2000 : (0%)

Figure 9-5. STEM also provides report views allowing the user to graph dynamic label variables. In the plots above, four locations are being monitored; the state variables, S (*susceptible*), E (*exposed*), I (*infectious*), and R (*recovered*), are plotted as a function of time in separate graphs for each monitored location.

The visualization capabilities of STEM are themselves Eclipse plug-ins so users may easily add new components to visualize data. The built in components include the GIS Map View shown in Figures 9-3 and 9-4 as well as plug-ins that direct output to other viewers, for example, Google Earth™ as shown in Figure 9-6.

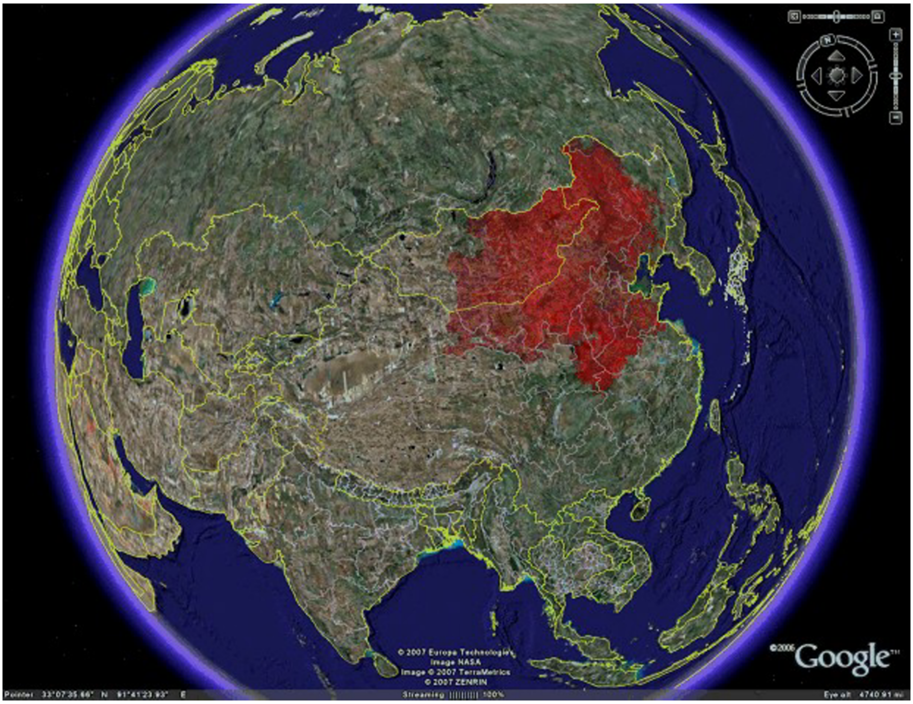

© 2007 Europa Technologies
Image NASA
Image © 2007 TerraMetrics
© 2007 ZENRIN
©2006 Google
Pointer 33°07'35.66" N 91°41'23.93" E
Streaming 100%
Eye alt 4740.91 mi

Figure 9-6. Google Earth™ as an Alternate View. In this figure, an infectious disease model (without air travel) is running in Asia and with the output directed to the Google Earth viewer™.

3.3.2 Disease Model Computations

The current version of STEM includes decorators for both deterministic and stochastic SI, SIR, and SEIR models based on finite difference equations (Schaffer and Bronnikova, 2001). The basic *SEIR* disease model assumes a uniform population at a single location and that the population members are well "mixed," meaning that they are equally likely to meet and infect each other. The built in decorators in STEM allow for either fixed or variable populations. For a normalized population, the model is defined by the four equations below:

- $\Delta s = \mu - \beta s\cdot i + \alpha r - \mu s$
- $\Delta e = \beta s\cdot i - \varepsilon e - \mu e$
- $\Delta i = \varepsilon e - \gamma i - \mu i$
- $\Delta r = \gamma i - \alpha r - \mu r$

Where:

- *s* is the normalized *Susceptible* population
- *i* is the normalized *Infectious* population

- μ is the *background mortality rate*, and, because it is assumed that the population was not growing or shrinking significantly before the onset of the disease, μ is also assumed to be the birth rate.
- β is the disease *transmission (infection) rate*. This coefficient determines the number of population members that become *Exposed* per population member in the *Infectious* state, assuming the entire population is in the *Susceptible* state.
- γ is the *Infectious* recovery rate. This coefficient determines the rate at which *Infectious* population members *Recover*.
- α is the *immunity loss rate*. This coefficient determines the rate at which *Recovered* population members lose their immunity to the disease and become *Susceptible* again.
- ε is the *incubation rate*. This coefficient determines the rate at which *Exposed* population members become *Infectious*.

In order to turn these zero dimensional equations into a spatiotemporal solution (Haberman, 1998), two transportation models are included in STEM. The first allows for mixing of populations that are connected by appropriate edges. A mixing model is appropriate, for example, when populations in adjacent connected regions are commuting between the regions on a time scale less than or of the same order as the simulation time step. A second model of transportation, also using the edges defined in STEM, model transportation as a packet model where the transportation packet itself is a node in which a disease state may evolve in time. The packet model is appropriate in ships or airplanes where the transportation time is longer than the finite difference time unit. STEM also provides examples of how a user can "extend" the built in decorator models and create their own advanced experimental models (Cummings et al., 2004). The built in "experimental example" shows how a user can extend the base models by adding nonlinearity to the infection term ($\beta s \cdot i$) defined above (Liu et al., 1987).

4. CONCLUSION

Today STEM is a work in progress. The various models and reference data included in STEM are provided as examples of how users can create and define their own models and data. While we have made, and continue to make, every effort to ensure the data reflects information available today in the public domain, we do not represent or guarantee the accuracy of data or of the original sources, including populations, area, population densities, and geographic information (precise latitudes and longitudes, etc.). In some cases, estimations were used to provide approximations for population densities where no public data could be found. We based the mathematical models in

STEM on standard textbook models (e.g., SIR, SEIR), but have not validated them independently. Accordingly, we give users the capability to adjust the rate constants in these mathematical models to create their own approximations or models of infectious disease. In addition to changing rate constants, advanced users will want to code their own mathematics extending the simple base models.

We do not claim that the simple base models included in STEM are optimal for any particular infectious disease. We recognize that the creation of a good model for real disease is part of modern epidemiological research. In developing STEM, we are looking to provide a framework to support researchers in creating, exchanging, and validating sound disease models. Part of our plan is to build tools into STEM to help users do their own validation studies of various models.

As a collaborative platform, STEM will support and be supported by the research community. Their involvement and their contributions will contribute to STEM as part of an infrastructure for global health.

ACKNOWLEDGEMENTS

This project is being developed under Contract Number FA7014-07-C0044 with the U.S. Air Force Surgeon General's Office (AF/SG) and administered by the Air Force District of Washington (AFDW). The Air Force has not accepted the products depicted and issuance of a contract does not constitute Federal endorsement of IBM Almaden Research Center.

The authors also acknowledge the Eclipse Foundation (http://www.eclipse.org) for support of the technology project and the open source technology that provide the foundation for STEM.

QUESTIONS FOR DISCUSSION

1. STEM is being developed as an open source computer software system. What advantages does this provide to the developers and to the researchers who will use it?
2. According to its developers, "STEM's main innovation is its framework that partitions a disease model into constituent components that represent different aspects of a disease model." Discuss.
3. What types of visualizations are available to users of STEM? Describe the geographic representations and report-generated graphics STEM provides.

REFERENCES

Ash, C., and Roberts, L. (2006). “Influenza: The state of our ignorance,” *Science*, 312 (April 21, 2006), 379. http://www.sciencemag.org.

Cummings, D.A., Irizarry, R.A., Endy, T.P., Nisalak, A., and Burke, D. (2004). “Travelling waves in dengue hemorrhagic fever incidence in Thailand,” *Nature*, 427:344–347.

Epstein, J.M., and Cummings, D. (2002). *Toward a Containment Strategy for Smallpox Bioterror: An Individual-Based Computational Approach,* CSED Working Paper 31 (December 2002). Washington, DC: Brookings Institution.

Ford, D.A., Kaufman, J.H., and Eiron, I. (2006). “An extensible spatial and temporal epidemiological modeling system,” *International Journal of Health Geographics*, 5(4) (January 17, 2006). http://www.ij-healthgeographic.

Gross, J.L., and Yellen, J. (2003). *Handbook of Graph Theory*. Boca Baton, FL: CRC Press.

Haberman, R. (1998). *Mathematical Models: Mechanical Vibrations, Population Dynamics, & Traffic Flow (Classics in Applied Mathematics)*. Philadelphia, PA: Society for Industrial & Applied Mathematics.

Liu, W-M., Hethcote, H.W., and Levin, S.A. (1987). “Dynamical behavior of epidemiological models with nonlinear incidence rates,” *Journal of Mathematical Biology*, 25:359–380.

Myers, L.A., Newman, M.E.J., Martin, M., and Schrag, S. (2003). “Applying network theory to epidemics: Control measures for mycoplasma pneumonia outbreaks.” *Emerging Infectious Diseases*, 9(2):204–210. (February 2003). http://www.cdc.gov/ncidod/EID/volno2/02–0188.

Schaffer, W.M., and Bronnikova, T.V. (2001). *Ecology/Mathematics 380: Modeling Microparasitic Infections*. See for example and references therein. Available at: http://www.bill.srnr.arizona.edu/classes/195b/195b.epmodel.

Widgren, S. (2004). *Graph Theory in Veterinary Epidemiology – Modelling an Outbreak of Classical Swine Fever*. Thesis. Institution for Ruminant Medicine and Veterinary Epidemiology, Swedish University of Agricultural Sciences.

SUGGESTED READING

Anderson, R.M. (1982). *Population Dynamics of Infectious Diseases: Theory and Applications*. New York: Chapman and Hall.

Banks, T. and Castillo-Chavez, C. (eds.). (2003). *Bioterrorism: Mathematical Modeling Applications in Homeland Security*. *SIAM Series Frontiers in Applied Mathematics*, Volume 28.

Barrett, C.L., Eubank, S.G., and Smith, J.P. (2005). “If smallpox strikes Portland,” *Scientific American*, 292:42–49, How modeling might be used to plan for and develop responses to epidemic events.

Enserink, M. (2003a). “SARS in China: China’s missed chance,” *Science*, 301:294–296.

Enserink, M. (2003b). “SARS in China: The big question now: Will it be back?,” *Science*, 301:299.

Gumel, A., Castillo-Chavez, C., Clemence, D.P., and Mickens, R.E. (2006). *Modeling The Dynamics of Human Diseases: Emerging Paradigms and Challenges*, American Mathematical Society, Volume 410.

Normile, D., Enserink, M. (2003c). “SARS in China: Tracking the roots of a killer,” *Science*, 301:297–299.

Rios-Doria, D., Chowell, G., Munayco-Escate, C., Whitthembury, A., and Castillo-Chavez, C. (2009). "Spatial and Temporal Dynamics of Rubella in Peru, 1997-2006: Geographic patterns, age at infection and estimation of transmissibility," In *Mathematical and Statistical Estimation Approaches in Epidemiology*. Edited by G. Chowell, J.M. Hyman, L.M.A. Bettencourt, and C. Castillo-Chavez, Springer.

ONLINE RESOURCES

Eclipse Platform: http://www.eclipse.org/platform/.
Eclipse Platform Technical Overview: http://www.eclipse.org/articles.
Open Services Gateway initiative (OSGi) Alliance: http://www.osgi.org.
Spatiotemporal Epidemiological Model (STEM): http://www.eclipse.org/stem.

Chapter 10

SURVEILLANCE AND EPIDEMIOLOGY OF INFECTIOUS DISEASES USING SPATIAL AND TEMPORAL CLUSTERING METHODS

TA-CHIEN CHAN and CHWAN-CHUEN KING*

CHAPTER OVERVIEW

In the control of infectious diseases, epidemiologic information and useful clustering algorithms can be integrated to garner key indicators from huge amounts of daily surveillance information for the need of early intervention. This chapter first introduces the temporal, spatial and spatio-temporal clustering algorithms commonly used in surveillance systems – the key concepts behind the algorithms and the criteria for appropriate use. This description is followed by an introduction to different statistical methods that can be used to analyze the clustering patterns which occur in different epidemics and epidemic stages. Research methods such as flexible analysis of irregular spatial and temporal clusters, adjustment of personal risk factors, and Bayesian approaches to disease mapping and better prediction all will be needed to understand the epidemiologic characteristics of infectious diseases in the future.

Keywords: Infectious disease epidemiology; Geographical information system; Spatial epidemiology; Tempo-spatial cluster methods; Dengue; Influenza; Emerging infectious disease; Taiwan

* *Institute of Epidemiology, College of Public Health, National Taiwan University, 17 Xu-Zhou Road, Taipei (100), Taiwan,* *chwanchuen@gmail.com*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_10,

1. INTRODUCTION

Spatial epidemiology has been commonly utilized to describe and to analyze the geographical distributions of diseases in recent decades. The distribution patterns of diseases are further investigated by several risk factors including demographic variables, levels of social economic status, environmental factors, genetic variations, exposure-related behaviors, contact patterns, specific niche of the etiologic agent, and modes of transmission [1, 2]. In general, descriptive epidemiologic studies present the mortality or incidence rate of an interesting disease by using thematic maps. The best example is John Snow's cholera map used in 1854 (Figure 10-1). Snow plotted all fatal cholera cases on the map to find that the contaminated pump was located on Broad Street in London, United Kingdom [http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm]. In recent decades, the geographic information system (GIS) has been applied to understand the epidemiology of infectious diseases, particularly the relationship among agent, host and environment [3, 4]. And it even helped to eliminate cholera outbreaks in Bangladesh [5].

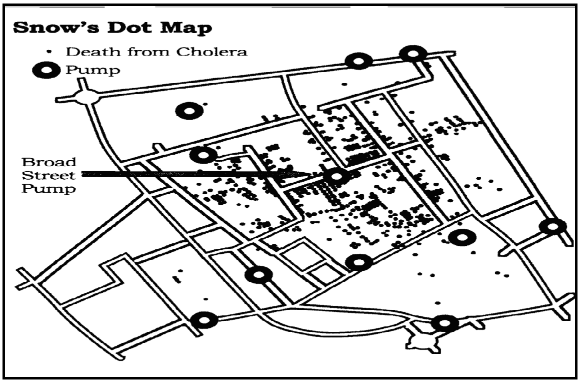


Figure 10-1. John Snow's dot map of cholera cases in 1854 (Source: http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm).

Surveillance, a public health endeavor to monitor health data regularly by searching for evidence of a change, is the most cost-effective way to provide early warning signals and then to prevent outbreaks of infectious diseases [6]. The traditional analysis of geographical distribution of disease cases is generally to mark darker colors in a choropleth map[1] with the location of cluster cases that can be identified visually. This approach is easily misled

[1] **Choropleth Map:** A thematic map in which areas are distinctly colored or shaded to represent classed values of a particular phenomenon.

by the misclassification of symbology[2] or by neglecting temporal factors. On the other hand, much progress has been made in spatial techniques, which are frequently used to indicate the extent of "clustering" across a map. The follow-up spatial analysis can determine whether the increase in each epidemiologic measure is localized or general and even where high risk areas are located with statistically significant increases [7]. Furthermore, development of spatial and temporal clustering methods may provide a more integrated picture of the dynamic diffusion of disease cases that could block further transmission more effectively. In other words, the combination of surveillance, spatial techniques, and statistical methods – particularly the methods developed for characterizing the spatial and temporal clustering, can not only improve the surveillance system but can also enhance the effectiveness of the surveillance system to reach public health goals.

2. CURRENT COMMONLY USED METHODS IN SPATIAL, TEMPORAL, AND TEMPO-SPATIAL CLUSTERING

Investigating disease clusters is an urgent task for public health authorities and professionals. If the disease happened non-randomly in temporal and spatial units, the clustering cases in time and place would be observed. Since outbreaks of emerging infectious diseases (EID) have been increasing rapidly in the past 2–3 decades, infectious disease surveillance becomes the most important task in public health. With the advances of information technology, electronic disease reporting systems have been established in many parts of the world. The real-time collection of disease information through the Internet is becoming more feasible [8]. However, numerous data need to be summarized. Therefore, the development of more convenient algorithms to detect temporal and spatial clustering is necessary to help public health staff with routine monitoring. In general, temporal clustering algorithms focus on setting up the baseline data for determining the threshold cut-off values. When the observation value exceeds the expected value, the alert signal will be triggered [8–10]. Spatial clustering algorithms test the null hypothesis, which assumes the disease is randomly distributed. If the null hypothesis is rejected by the predefined confidence level, the so-called "spatial clusters" would occur. Since time and place are the two most important epidemiological

[2] **Symbology:** The set of conventions, rules, or encoding systems that define how geographic features are represented with symbols on a map. A characteristic of a map feature may influence the size, color, and shape of the symbol used.

characteristics in infectious disease outbreaks, recently efforts have tried to consider both simultaneously.

2.1 Temporal Clustering Methods

Public health professionals have used three main methods – historical limit, cumulative sum (CUSUM), and time series to detect cases with temporal clustering.

2.1.1 Historical Limit, the Concept of Moving Average, and Scan Statistics

Historical Limit

Historical limit is a method that was frequently used to monitor infectious disease surveillance data in the United States before 2001 by requiring historical information – generally at least 5 years of background data – to serve as the upper baseline data for statistical aberration detection. If the observed value is higher than the 95% confidence limit of this upper baseline data, it is assumed that an outbreak would occur [10]. Therefore, the levels of baseline data in this method are easily influenced by the large-scale epidemic(s) of the past.

The Early Aberration Reporting System (EARS), developed by the Centers for Disease Control and Prevention in the United States of America (US-CDC), consists of a class of quality control (QC) charts, including the Shewhart chart (P-chart), moving average (MA), exponentially weighted moving average (EWMA), and variations of cumulative sum (CUSUM) [10]. In temporal analysis of syndromic surveillance data, a common approach is the use of a sample estimate for obtaining the baseline mean and standard deviation (SD) to circumvent the possible difficulties associated with the baseline trend that may be complicated by the seasonality and daily fluctuation of the syndromic data [9].

The Application of Moving Average

In 1989, Stroup et al. [11] used three simple moving average measures, moving average in mean, moving average in medium, and scan statistics, to implement historical limit methods on notifiable infectious diseases. The concept of analysis adopted the general form shown in Equation 10-1. The numerator X_0 was the observation value at the current time point (the temporal unit that can be daily or weekly or monthly, defined by the users). The denominator, serving as a baseline, was calculated as a mean or a median value of the same time period plus the pre- and post-periods within the past 5 years (e.g., total 15 time points) (Figure 10-2).

$$\frac{X_0}{Baseline} \tag{10-1}$$

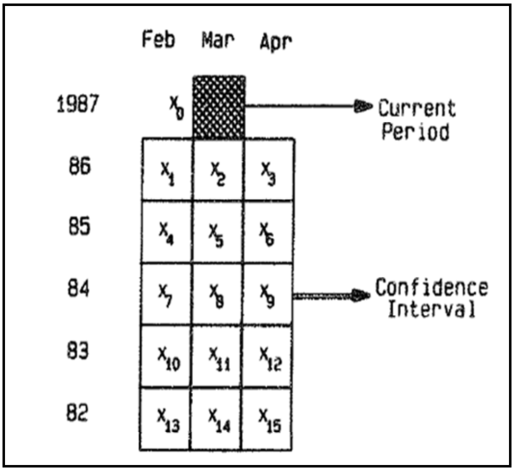


Figure 10-2. Baseline for comparison cases reported for March 1987 [11].

Since 1989, the historical limit method has been employed in the summarized surveillance results of the U.S.A. published in the Morbidity and Mortality Weekly Report (MMWR). The case numbers of a reported specific disease for a given health outcome in the three most recent time periods (pre-, current, post-) are compared with historical incidence data on the same health outcome from the same three time periods of the preceding 5 years. The results are shown by comparing the ratio of the current case numbers with the historical mean and SD. The historical mean and SD involve the 15 totals of the three time intervals, including the same previously mentioned three periods (the current period plus the preceding one period, and the subsequent one period over the preceding 5 years) as the historical data. For example, if we want to know whether the influenza-like illness (ILI) cases in September of 2008 are unusual or not, the ILI case numbers of August, September and October in each year from 2003 to 2007 need to be added up to obtain a mean or a median for comparison. For an infectious disease with a strong seasonality trend, the seasonally adjusted CUSUM can be applied. That is, the positive one-sided CUSUM where the count of interest is compared with the 5-year mean ± SD for that specific period. Similarly, Taiwan's emergency department syndromic surveillance system can track diseases with strong holiday, post-holiday, or weekend effect because closures of most local clinics occur on most holidays/weekends.

To verify the accuracy and sensitivity of the outbreak detection, the epidemiologic investigation has to be followed. In 1993, Stroup et al. [12] compared the historical limit method with three other methods (bootstrap, jackknife, delta) for estimating standard error to detect abnormal time clusters. The results showed that the values estimated by using the historical limit method and delta method were close to the true value. The variance values estimated by the two methods were under-estimated, which might result in over-alert. Therefore, using bootstrap in the historical limit method to obtain an estimated confidence interval is a good statistical approach.

2.1.2 Cumulative Sum

Cumulative sum (CUSUM), a method initially used in quality control, has recently been applied to surveillance [13]. The original idea of CUSUM was to set up a control limit under a steady period. The strength of CUSUM, similar to the exponentially weighted moving average control chart, is to detect small shifts in the process mean even without historical data for 3–5 years. Two important parameters are used in CUSUM. First, an appropriate value for the control limit, h, is based on the average run length (ARL) of the CUSUM, while the failure rate is acceptable within a time interval for quality control that can be regarded as the upper limit of failure rate in quality control or the case number of a studied disease in surveillance. The other parameter is k, which represents the minimum standardized difference from the running mean. The traditional CUSUM chart generally uses the sum of differences both above and below the mean to detect anomalies in either direction. For biosurveillance, an upper sum S_H is used to look only for excessive counts in which small differences are ignored and only differences of at least 2k standard deviations above the mean μt (mean of day/week/other time unit t) are counted. A common practice is to set k at 0.5 to detect a shift of one SD.

Since the anthrax attack [10] **that** occurred shortly after the September 11, 2001World Trade Center Attack, more interest has arisen in using public health approaches that could rapidly detect "unusual events" without requiring several years of background data. Therefore, newly developed non-historical aberration detection methods can analyze data as short as 1 week. To consider daily variation, the revised CUSUM method, using the measurements from C1, C2 to C3 to increase the sensitivity of the detection based on a positive one-sided CUSUM calculation from a week's information, was then developed [9]. For C1 and C2, the CUSUM threshold reduces to the mean plus 3 standard deviations (SD). The mean and SD for the C1 are based on the raw data from the past 7 days. The mean and SD for the C2 and C3 are based on the data from 7 days, ignoring the two most recent days to minimize the bias. For C1 and C2, the test statistic on day t was calculated as $St = \max [0, (X_t - (\mu t + k^*\sigma t))/\sigma t]$ where X_t is the count data (number of cases) on day t, k is the shift from the mean to be detected, and μt and σt are the mean and standard deviation of the counts during the baseline t time period. For C1, the baseline period is (t-7 to t-1); for C2 the baseline is (t-9 to t-3). The test statistic for C3 is the sum of St + St-1 + St-2 from the C2 algorithm. Using these C1, C2 and C3, outbreaks of any infectious disease with a strong seasonal or regular fluctuation trend can be easily detected. This is particularly useful for an agent such as influenza virus, in which different types or subtypes of the virus are dominant each year in addition to continuous antigenic drifts.

2.1.3 Time Series

Based on the epidemiologic characteristics of each infectious disease, certain diseases have trends in the periodicity of epidemics. Therefore, researchers could use time series models such as the autoregressive integrated moving average model (ARIMA) or the Serfling model to predict the possible time and wave of the outbreak. The ARIMA models fit better with time series data that can be applied to better understand the characteristics of epidemiologic data or to predict future time points in a series. The fine-tuning characteristics of ARIMA involve adding lags of the different series and/or considering time lags of the forecast errors to the prediction equation to better predict the temporal trend. The Serfling model uses regression by adding sine and cosine functions to adjust the periodicity. It is frequently used in the excess mortality data analysis of influenza or pneumonia and influenza. Using these methods, the training period of the dataset to be selected is very critical. The cyclical pattern of time intervals such as seasons or months or other time units should be represented in the training data. Then, the dynamic pattern would be updated and predicted by the latest data. This time series model has recently been used in predicting the impact of several infectious diseases related to climate changes.

2.2 Spatial Clustering Methods

To analyze spatial data, the characteristic of the data – pointed data or regional data – needs to be examined first. In general, northern, southern, central and eastern Taiwan regional data are frequently used in routine surveillance for monitoring possible changes of several important infectious diseases in different geographical areas. Once the outbreaks occur, point data will be gathered by collecting the geo-coding information of the cases' addresses or by using a Global Positioning System (GPS) to locate any important sites related to possible exposures of the cases for further detailed investigation.

The next step is to select appropriate methods for analysis of spatial clustering. Three methods of spatial clustering, including global cluster, local cluster and focused cluster, are frequently used for analyzing epidemiologic data [14]. Spatial autocorrelation, involving global indices and local indices as the degree of association between a factor of interest and its specific location, is a convenient approach to explore the degree of spatial clustering among cases and the possible associated spatial risk [14].

2.2.1 Global Clustering Test

Global cluster detection methods can help to determine whether or not spatial clustering exists in any place of the study period statistically [15]. Positive spatial autocorrelation reflects a “clustering” of points related to a particular variable of interest to be assessed. Negative spatial autocorrelation (e.g., spatial outliers) displays inverse correlation between the tested neighboring areas based on the attribute of interest. A zero spatial autocorrelation indicates a random distribution rather than a cluster or a dispersed distribution. This method is particularly useful if the source of infection is unknown or not easily identified. The limitation of this method is that it cannot identify the specific location(s) of spatial cluster(s).

Clustering tests involve four types – (1) area-based tests for global clustering, (2) point-based tests for global clustering, (3) area-based tests for local clustering, and (4) point-based tests for local clustering. Different statistical tests are used for each of these four types, depending on the type of data. Area data emphasize analysis on the relationship between the tested area and its neighboring area. Pointed data stress the distance between the two points. However, the central point of an area can be regarded as a point and then be tested in point data. Besides, both LISA and Moran’s I spatial autocorrelation tests in Table 10-1 can be applied to point or polygon data, depending on the definition of the spatial relationship. If public health authorities have pointed data, more hypotheses can be tested and better diffusion dynamics of cases can be described. To protect patients’ privacy, more area-type data are available than pointed-type data, particularly for those infectious diseases with higher social stigma.

Table 10-1 summarizes the clustering methods. The first two methods (Whittemore’s test and K nearest neighbors) are global tests for pointed data and the next three methods are local tests for pointed data. Whittemore’s test is to measure the mean distance of all cases divided by the mean distance of all individuals in that area. IF this ratio is less than 1, it reflects there is a cluster. In addition, the K nearest neighbor method assumes that the spatial

Table 10-1. Summary of the most commonly used spatial clustering algorithms.

Data format (point/polygon)	Type of method (global/local)	Statistical methods	Authors
Point	Global	Whittemore's test	Whittemore et al. (1987)
Point	Global	K nearest neighbors	Cuzick and Edwards (1990)
Point	Local	Geographical analysis machine (GAM)	Openshaw et al. (1987)
Point	Local	Besag and Newell test	Besag and Newell (1991)
Point	Local	Satscan	Kulldorff (1995)
Area	Global	Moran's I	Moran (1950)
Area	Global	Geary's C	Geary (1954)
Area	Local	Gi	Getis and Ord (1992)
Area	Local	Local indicator of spatial association (LISA)	Anselin (1995)

distribution of cases is random. If the observed value is higher than the expected value, it means a spatial cluster is present there. However, this test does not point out where the cluster is.

On the other hand, two global tests for area-type data are Moran's I and Geary's C tests. Moran's I statistic works by comparing the value at any one location with the value at all other locations. Moran's I is the most frequently used as the screening tool for clusters in global testing. It is generally used to reveal whether there is evidence of clustering or indication of the evidence of hot spots, shown by geographic boundary aggregated data. The results of Moran's I vary between –1.0 and +1.0. The Moran's $I > 0$, $=0$, and <0 indicate the positive spatial autocorrelation, random distribution, and negative correlation, respectively. If areas are close together with similar values, the Moran's I result is high. Geary's C statistic is used to describe differences at the local level by measuring the deviations in intensity values of each point with one another. The values of the C statistic vary between 0 and 2, where values equal to 1 represent spatial independence for each point, values less than 1 indicate evidence of positive spatial autocorrelation, and values between 1 and 2 indicate evidence of negative spatial autocorrelation.

2.2.2 Local Clustering Test

This method can provide definitive information on the specific location of clusters derived from local autocorrelation indices to evaluate clustering trends of an interested variable or factor, particularly under the condition with unidentified source of the infection, by determining whether the data are spatially similar or different at that specific area/site [16]. Practically, public health personnel can use this method to define the risk areas of a disease for further prevention and control efforts. For example, the epidemic of dengue in 2002 in Taiwan was fast spreading. First, we investigated whether clustering occurred using the "global cluster" test. Then, the boundary between Kaohsiung City and Kaohsiung County was identified by "local cluster" test, and prevention and control efforts were immediately implemented. In infectious disease epidemiology, the local clustering test is very useful in investigating not only the source of the infection but also potentially unidentified risk areas that might facilitate subsequent diffusion and further spread of cases.

Satscan is a point-based test for local clustering whereas Local Indicator of Spatial Autocorrelation (LISA) is an area-based test for local clustering. Both methods are very frequently used. LISA divides the significant areas into four categories: (1) high-high for central area is high and neighboring area is also high, (2) high-low, (3) low-high and (4) low-low. The other area-type data local test is Gi and its calculation is quite simple. High Gi value represents the presence of clusters in high Gi value areas whereas low Gi value indicates the existence of cluster in low Gi value areas, similar to high-high and low-low areas in LISA, but it does not involve the other two categories in the LISA method.

Scan Statistic

Spatial scan method, initially used to detect clusters in cancer epidemiology [17], has been applied to infectious diseases since 2000, such as bovine tuberculosis in Argentina, Toxoplasmosis in Southeast Asia, West Nile encephalitis in the United States, and human granulocytic ehrlichiosis near Lyme disease in Connecticut [18, 19]. The spatial scan can handle both point and area types of data, and it takes the central point of each polygon of the area-type data to be calculated. Nowadays, Satscan, which uses a circular window (circle centroid) to scan the entire study area to calculate the likelihood ratio, has become the most popular tool to detect diseases clusters. For any given location of the centroid, the radius of the screened window is continuously changed to take any value between zero to a certain upper limit.

The size of the circular window changes until the predefined population is screened. The maximum size of this circular window in the tested area has to be less than 1/2 of the target population to get a meaningful likelihood ratio in comparison with those in other areas and to avoid the overlap areas as well. After scanning the whole area, the area of the maximum likelihood ratio with statistical significance, so-called "clustering area," will be obtained.

However, the circular window in the spatial scan method is not the natural shape of most clusters. Therefore, an ellipse shape [20] and irregular shape [21] have been developed recently.

Local Indicator of Spatial Autocorrelation

In many studies, the LISA method has frequently served as the spatial risk index to identify both significant spatial clusters and outliers [22]. *Spatial outliers* present particular areas that have values of the tested variable opposite to its neighboring areas. The definition of a LISA index is:

$$I(i) = \frac{X_i - \overline{X}}{\delta} \times \sum_{j=1}^{n} W_{ij} \times \frac{X_j - \overline{X}}{\delta}$$

where I(i) = the LISA index for region i, W_{ij} = the proximity of region i to region j, where a value of 1 means the region i is next to the region j, and a value of 0 means the region i is far away from the region j, X_i = the value for the tested variable in region i, X_j = the value for the tested variable in region j, $\overline{X}$ = the mean value of the tested variable, δ = the standard deviation of X_i, and n = the total number of regions to be tested.

A positive *l(i)* value of the LISA index designates that a region and its neighboring areas tend toward local spatial dependency. In other words, area-specific cases of an interested infectious disease in the tested region and its neighboring areas approach homogeneity. In contrast, a negative *l(i)* value, which tends toward the opposite values between X_i and X_j (i.e., X_i = high, X_j = low or vice versa), specifies that the spatial dependency is negative, thereby suggesting that the region is a spatial outlier in relation to neighboring regions. In general, a Monte Carlo statistical test is used to evaluate the significance of spatial clusters and outliers [23]. Using LISA index values, risk areas of any infectious disease, such as dengue in southern Taiwan, have been classified into several different risk levels for implementing various control strategies to counteract outbreaks [24].

GAM and Besag and Newell Tests

The other two local clustering methods for pointed data are geographical analysis machine (GAM) and Besag and Newell tests. GAM is to test whether there is a statistically significant high disease rate by comparing each circle of the studied area with various radius values. The Besag and Newell test assumes that k is the minimum case number of the clustering area and then uses each case as a center to look for k-1 cases regarded as a cluster. In this way, the lacking neighboring cases force the investigator to look for further areas so that the case number divided by larger searching area becomes a smaller value implying "cases without cluster." Both of these two methods result in the overlaps of sub-clusters (circles with different radius values or different k-1 cases), in which case the GAM method offers higher repetition without independence that may provide more false positives.

2.2.3 Focused Clustering Test

The "focused clustering" test is to assess the clustering of the observed cases around a fixed point – the smallest scope that is different from "general clustering" or "local clustering" without having any prior information on the centre of clustering. Therefore, this test has been used to investigate raised incidence of disease, particularly the rare disease or the beginning period of an outbreak of infectious disease, in the vicinity of pre-specified putative sources of increased risk. In addition, the focused clustering method is applied to detect whether there is an excess risk or a cluster of cases of a disease around a putative source of the infection [25]. Stone's test [26] is a very popular method used in testing "focused clustering" since it is based on traditional epidemiological estimates after adjusting the important confounders – standardized mortality ratio (SMR) or standardized incidence ratio (SIR).

The following summary Table 10-1 helps readers to firstly assess which type of spatial data – pointed format or area format – are collected. Then, global clustering tests can be employed to examine the presence of clusters or not. If the answer is "yes," subsequent local clustering tests will be followed to indicate the exact location of the case clustering areas. All these methods can be found in GIS software or free statistic test R packages.[3] Different statistic tests using the same datasets can also be simultaneously compared and evaluated to find out which one offers the best power. In general, spatial scan statistic has good power in detecting hot spot clusters.

2.3 Spatial and Temporal Clustering Methods

In addition to a spatial clustering method, temporal factors must also be taken into consideration. Analysis of spatial clustering data is quite similar to the data analysis in cross-sectional study design in epidemiology. When the distribution of the cumulative cases is displayed, it only explains the results of the overall pattern without definite conclusions on causal inference. Once

[3] **The R Package for Multidimensional and Spatial Analysis:** This is a group of programs (Macintosh and VAX/VMS) that allows public health data analyzer to perform with ease various complex multidimensional and spatial analysis procedures (http://www.r-project.org).

the temporal factors are included into the analysis, the results can clearly show the different waves of the epidemics, the transmission patterns, and possible risk factors that are involved in different time periods. Then, the three most important epidemiologic characteristics – person, place and time – can be simultaneously integrated to obtain more insights than each characteristic alone. Here we briefly introduce two methods, namely Knox method and the space-time scan statistic, which integrate spatial and temporal factors.

2.3.1 Knox Method

The Knox method is to test for space-time interactions, particularly when there are different impacts of time factor on the studied population in various regions [27, 28]. The time and geographical location of each case is obtained. For each possible pair of cases, the distances between them are also calculated in time and space. If many of the cases that are "close" in time are also "close" in space or vice versa, then there is a space-time interaction. Users can predefine how close the time period and the geographical distance are to one another of those studied cases in temporal and spatial units, based on their research questions. Then for each space-time combination, expected values will be calculated by a 2 × 2 contingency table [29]. Cases are assumed to be rare, independent events, distributed as a Poisson variable. The significance of the departure of the observed number of close pairs (O) from the expected number (E) is tested using d, where:

$$d = \frac{(O-E)}{\sqrt{Var(O)}}$$

A d value greater than 1.96 indicates that there is a statistically significant cluster at p-value 0.05.

The Knox test is attractive in epidemiologic data analysis because it is simple and straightforward to calculate the test statistic without requiring the knowledge of controls. However, the Knox test can be biased if the population growth is not constant for different geographical areas (e.g., distribution does not meet Poisson distribution). For detecting an "early" outbreak of infectious disease, such bias is not a major problem to be considered.

2.3.2 Space-Time Scan Statistic

Space-Time Scan Statistic [16], an improved version of the purely spatial scan method, is defined by a cylindrical window with a circular geographic base and with height corresponding to time. The base will vary the radius continuously. The height reflects any possible time interval of less than or equal to half the total study period. The likelihood in each cylinder will be calculated. Using the cylinder with the maximum likelihood, and then selecting the tempo-spatial one with more than its expected number of cases, is denotes the most likely cluster.

Comparing the Knox and space-time distance methods, the Knox method categorizes the individual case's space-time distance into several groups and then uses a test similar to the Chi-square test. The temporal and spatial distance between the cases is determined by the user-based research questions. The space-time scan statistic, a purely spatial scan, uses the cylinder as the scanning window and the height is time. It scans over the study area by the different radii of the base to calculate the observed values in different areas. The expected value can be calculated by Monte Carlo simulation. Finally, the question on tempo-spatial clusters can be tested to determine whether the observed value exceeds the expected value. For example, if point data of individual cases from outbreaks of infectious diseases such as dengue or enterovirus-related cases are available, the Knox method is very suitable to apply. Alternatively, when an overall incidence or prevalence rate from different geographical regions rather than individual case data is available, the space-time scan method is more appropriate to use.

3. CASE STUDIES USING SPATIAL CLUSTERING METHODS IN INFECTIOUS DISEASE EPIDEMIOLOGY

The following sections introduce the application of the above spatial and spatio-temporal methods to infectious diseases with public health significance, including respiratory spread, gastrointestinal-related (GI) transmission, vector-borne transmission, zoonotic and emerging infectious diseases.

3.1 Respiratory Spread

Epidemics of tuberculosis (TB) have reemerged in recent years due to the increasing cases of acquired immunodeficiency syndrome (AIDS) and multi-drug resistant tubercle bacilli. The incidence rate of TB in the Fukuoka Prefecture urban area of Japan (Figure 10-3a) in 2001 was higher than that of the nationwide data. Using local cluster tests for pointed data by spatial scan statistics and spatial-temporal scan statistics, the spatial analysis alone identified TB clusters in different geographical areas of Japan that occurred in different years (Figure 10-3b), including: (1) Chikuho coal mining area in 1999, 2002, 2003 and 2004, (2) Kita-Kyushu industrial area in 2000, and (3) Fukuoka urban area in 2001 [30]. However, using the space-time analysis, the most likely clusters were found in the Kita-Kyushu industrial area in

2000. In other words, clusters of cases had already appeared in the Kita-Kyushu industrial area in 2000 before the occurrence of other spatial clusters from 2002 to 2004. Further analysis found that the occurrence of TB in the clusters located in northern Fukuoka Prefecture in 2000 were also significantly higher than those clusters identified in other years. In conclusion, spatial method alone can be used to evaluate the cluster cases in each year whereas spatial-temporal methods can be applied to find out where cluster cases are within a specific time period and their dynamic changes over different time periods and places as well.

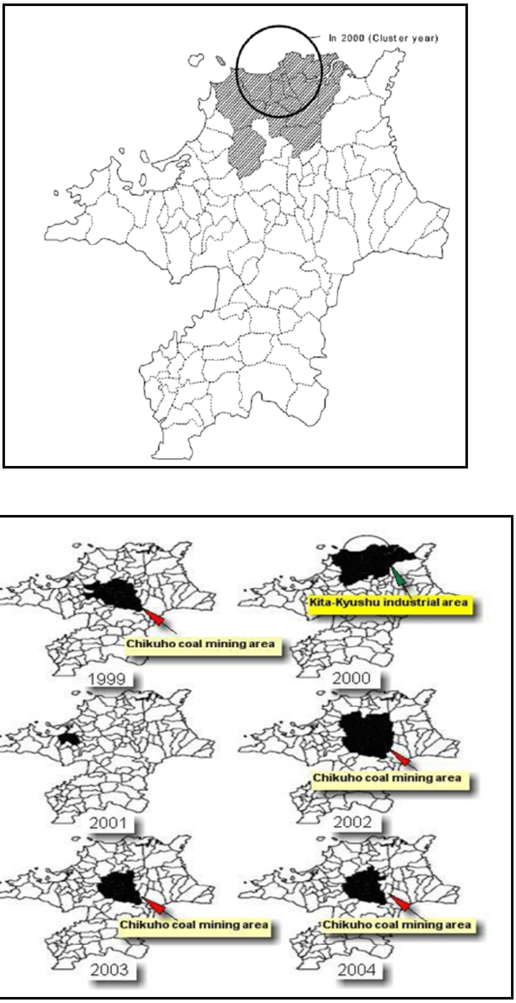


Figure 10-3. (**a**)The Space-time Analysis detected clusters of TB cases in Kita-Kyushu industrial area located in the northern Fukuoka Prefecture during 1999~2004, based on the historical data from 1999 to 2004 [30]. (**b**) Locations of the clusters of TB cases detected in Fukuoka Prefecture from 1999 to 2004, based on a purely spatial analysis [30].

3.2 GI-Related Transmission

Giardia lamblia is the most frequently identified human intestinal protozoa in Canada with an estimated prevalence of 4–10%. The spatial scan statistic was used to identify local spatial clusters of those cases with pointed data, to measure the possible "rural" effect from the distribution of the giardiasis and to explore the associations between the area-specific giardiasis rates and the manure application on agricultural land and livestock density [31]. Finally, giardiasis clusters in southern Ontario were identified (Figure 10-4a). However, neither livestock density (Figure 10-4b) nor manure application on agricultural land plays an important role in the epidemiology of giardiasis there.

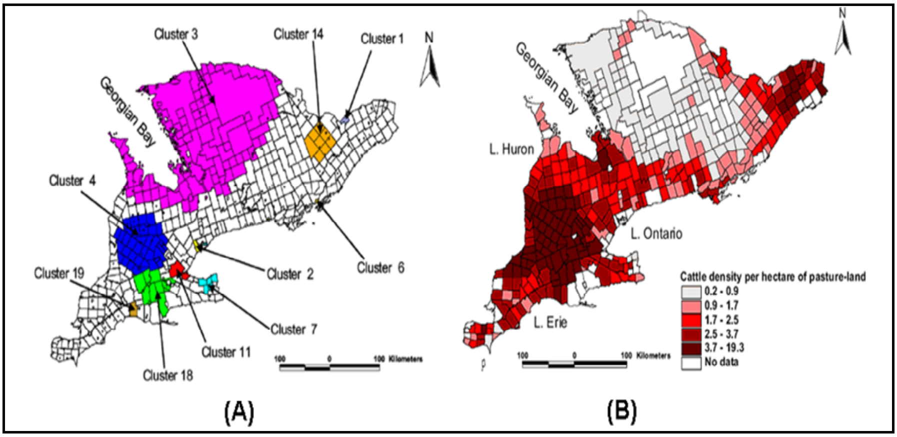


Figure 10-4. (**a**) Spatial distribution of giardiasis with significant high rate of giardiasis clusters located in southern Ontario during 1990–1998, (**b**) Spatial distribution of cattle density in southern Ontario[31].

3.3 Vector-Borne Transmission: Dengue as an Example

To retrospectively detect spatio-temporal dengue clusters at patients' homes (point-type data) in Iracoubo, French Guiana and the disease onset dates during 2001 [32], GIS integrated with the Knox method was employed. Heterogeneity in the variations of relative risk (RR) in space and time was found to be associated with mosquito factors, including mosquito feeding cycle, host-seeking behavior, and life span of mosquitoes. Particularly, higher RR values were more likely to be identified in the time periods and areas with shorter temporal and spatial distances (Figure 10-5a) and more clear suspected/confirmed dengue clusters were detected in shorter time distances (Figure 10-5b). In addition, confirmed dengue cases showed more clear higher risk (in red color) than suspected dengue cases, illustrating the importance of laboratory diagnosis. The cluster analysis also proved that the probability of observing a dengue case outside of 100 m around the dengue foci, a distance measured to correspond to a statistical threshold, was low. However, this threshold could vary if the case numbers increased with the improved surveillance system. By contrast, Taiwan's GIS analysis of confirmed dengue cases showed that the relocation diffusion occurred more frequently as the duration of the epidemic wave in that epidemic site became longer [33]. In other words, spatial limit of transmission, expanding distribution of mosquito vectors even after control efforts, and dynamic changes in populations at risk (e.g., susceptible) can be obtained more precisely once integrated temporal and spatial data are simultaneously analyzed. The situation might be even more complicated for malaria, which involves

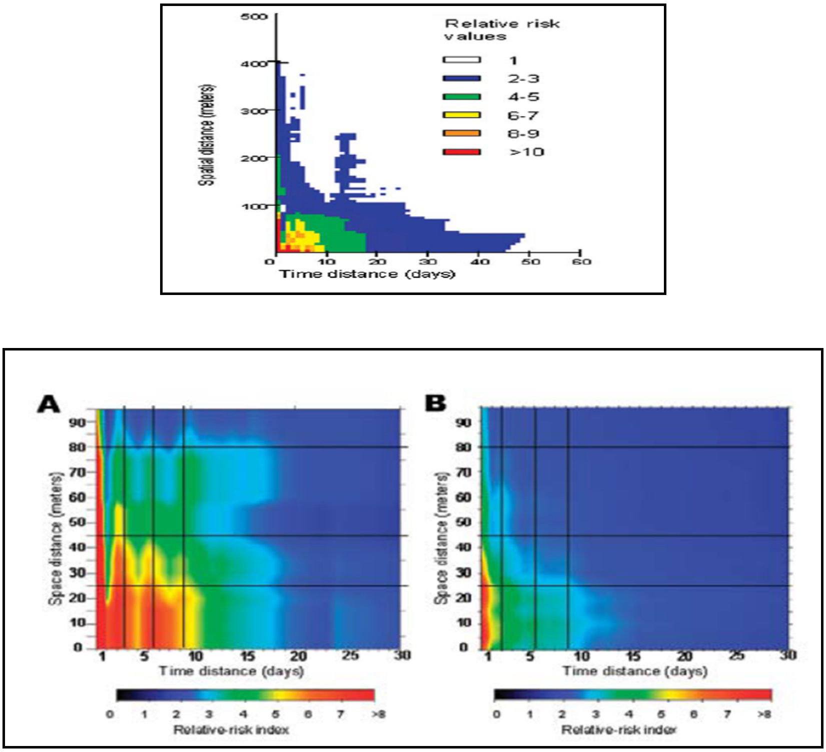
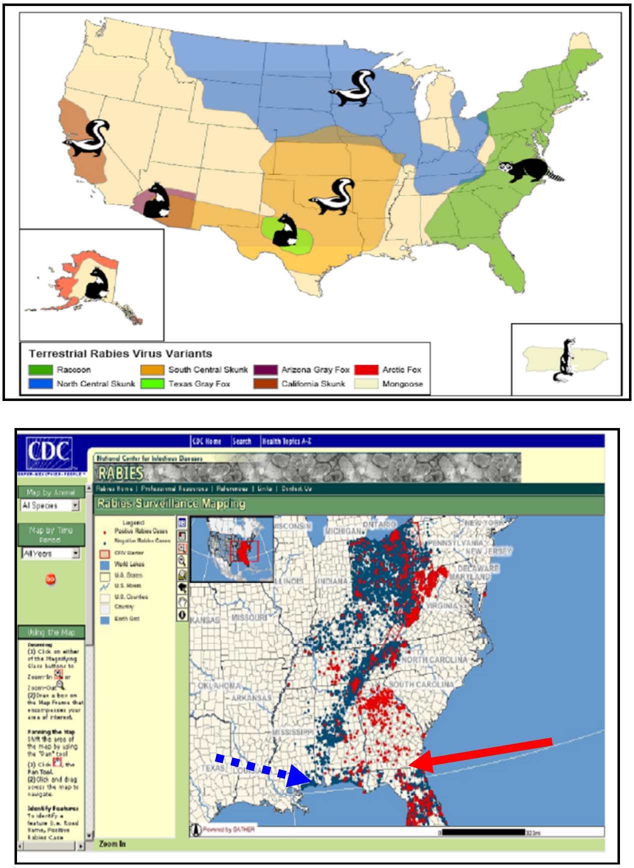


Figure 10-5. (**a**) The relative risk (RR) calculated from the confirmed dengue cases in Iracoubo, French Guiana during 2001, when space-distance and time-distance from the first index suspected dengue case varied from 0 to 500 m and from 0 to 60 days, respectively. Color areas indicated their RR values significantly greater than 1 ($p < 0.001$) [32]. (**b**) Main risk area for dengue fever (within 100 m and 30 days boundaries), derived from both the laboratory-positive dengue cases (A) and all suspected cases (B) in Iracoubo, French Guiana. *Vertical dark lines* indicate an apparent temporal periodicity, and *horizontal dark lines* correspond to apparent spatial breaks [32].

different species of mosquitoes and their variations in ecology [34], and for yellow fever, in which the immunity following vaccination means public health officials need to consider the "population at risk" as well as naturally acquired infection [35].

3.4 Zoonosis: Rabies as an Example

Zoonotic diseases involve the ecology of infectious diseases while animals in nature are sick. Therefore, the surveillance of zoonosis must start from the targeted animal population, its ecological niche and possible associated



Using the Map

Zooming
(1) Click on either of the Magnifying Glass buttons to Zoom-In or Zoom-Out.
(2) Draw a box on the Map Frame that encompasses your area of interest.

Panning the Map
Shift the area of the map by using the "Pan" tool
(1) Click , the Pan Tool.
(2) Click and drag across the map to navigate.

Identify Features
To identify a feature (i.e. Road Name, Positive Rabies Case

Figure 10-6. (**a**) Reservoirs for Rabies Virus in the United States. Geographic distribution of terrestrial rabies virus variants – defined by monoclonal antibody typing [36]. (**b**) The distribution of rabies in the USA, using complete web-based easily updated rabies RabID data of both animals and humans. Solid arrow (Right) indicates the geographical distribution of "positive" rabies identified by laboratory diagnosis. Dashed arrow (Left) indicates reported cases with the samples tested as "negative" for rabies [36].

environmental factors [4]. Particularly, a disease such as rabies would lead to higher case fatality rates if proper treatment and control were not implemented in a timely manner. Therefore, a GIS web-based platform for the surveillance of rabies named "RabID" was set up to rapidly map the animal rabies cases, to track the rabies reservoir and then to disseminate this information for public education at the US-CDC [36]. Several geographically discrete terrestrial wildlife reservoirs were identified (Figure 10-6a). In addition, real-time and web information on the type of animals infected and associated genotypes and strains of rabies virus identified by monoclonal antibodies have been shared among local animal and public health personnel across various geographical areas. This information will certainly facilitate the timely management of rabies control (Figure 10-6a and b). In other words, GIS information with integrated spatial epidemiologic characteristics is very useful for prevention and control on zoonotic or vector-borne infectious diseases, from public health planning to implementation and evaluation of the effectiveness of control.

3.5 EID: Avian Influenza as an Example

Avian influenza has been an increasing public health threat since the cross-country spread during 2003–2004. Between October 2005 and June 2006, 161 outbreaks of highly pathogenic avian influenza (HPAI) H5N1 occurred in poultry villages of Romania [37]. Using two combined temporal and geostatistical methods, Anselin's local indicator of spatial autocorrelation statistics (LISA) for area-type data and space-time permutation scan statistic for point-type data, the clusters of H5N1 were identified. The former method focuses only on spatial clusters and the latter method simultaneously considers temporal and spatial clusters. The space-time permutation scan statistic method is particularly useful in infectious diseases with shorter incubation period but closely associated with large-scope ecology and also in those situations where the numbers of populations at risk is unknown in syndromic surveillance [18]. The results found that the locations of the clusters were different by using the two different cluster algorithms (Figure 10-7a and b). The origin, evolution and increasing spread of the epidemic can be grasped more clearly. The outbreak first appeared in the region of the Danube River Delta by the introduction of the virus, implying the importance of landscape epidemiology. Then, the movement of poultry might facilitate its further spread to central Romania the next year. Using the spatio-temporal methods, the progression of the outbreak from a confined, local epidemic extended to a large, nationwide epidemic can be fully understood. Such efforts are very helpful to minimize the spread of the next H5N1 epidemic in other countries and the future global spread of HPAI H5N1 viruses.

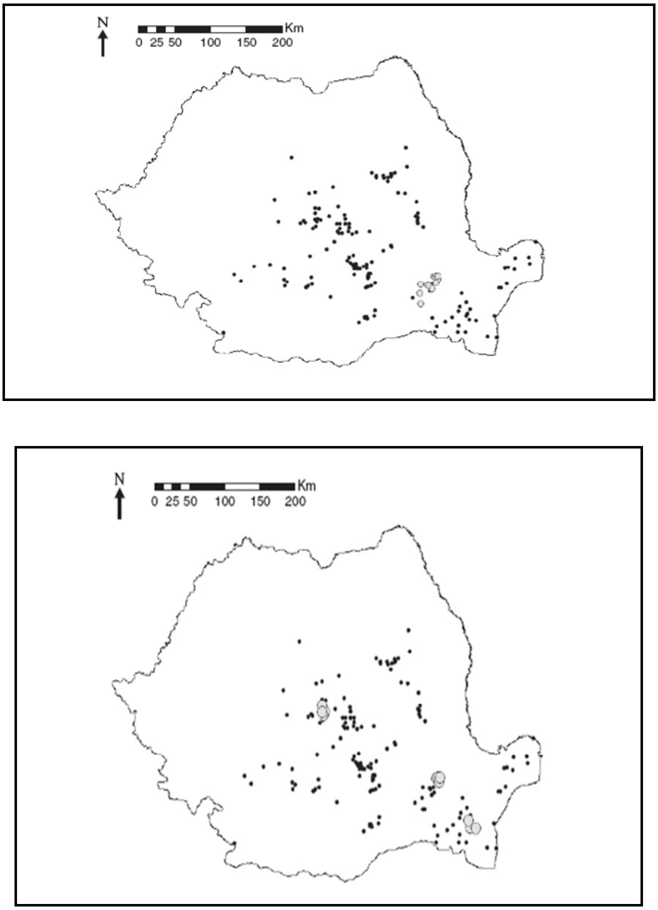


Figure 10-7. (**a**) One cluster (*open circle*) of HPAI subtype H5N1 and the villages with outbreaks (*filled circle*) in Romania, October 2005–June 2006, were identified by the local indicator of spatial autocorrelation statistic [37]. (**b**) Three clusters (*open circle*) of HPAI subtype H5N1 and the villages with outbreaks (*filled circle*) in Romania, October 2005–June 2006, were identified by the space-time permutation scan statistic [37].

To summarize the analysis of temporal and spatial clusters for infectious diseases with different modes of transmission, both the transmissibility and pathogenicity of the microbial agents are the key factors to determine the best method to be selected. For the diseases with high transmissibility and high pathogenicity, the rapidly cross-geographical spread shown by a cross-sectional map with the appearance of cases might indicate an emerging infectious disease which needs to use integrated spatio-temporal clustering methods for further data analysis. For the diseases with low transmissibility and high pathogenicity, the spatial clustering method can capture the distribution of the cases. For the diseases with high transmissibility and low pathogenicity, temporal clustering methods would need to be used to obtain warning signals as early as possible.

Until now, we have not had the perfect method to detect all kinds of clusters. Due to the unknown clustering pattern, it is better to use more than two methods to cross-validate the clusters before drawing a final conclusion.

4. CONCLUSIONS, LIMITATIONS AND FUTURE DIRECTIONS

4.1 What We Have Learned in the Past

Spatial and temporal clustering methods have been applied to prevention and control measures of infectious diseases, from improving surveillance systems, real-time integrating of clinical, microbiologic, environmental and epidemiologic data, to understanding the epidemiologic characteristics of infectious diseases and evaluating the effectiveness of control measures.

In routine surveillance systems, the algorithms such as CUSUM, ARIMA and Satscan have been widely used in different surveillance systems to detect early abnormal signals. The sensitivity and specificity of the algorithms for aberration detection need to be evaluated by each algorithm using different datasets. In general, the incorporation and integration of several different algorithms to complement each other can help to verify the occurrence of an outbreak.

For those infectious diseases with high communicability such as measles, smallpox or a disease with a high case fatality rate such as rabies, Ebola hemorrhagic fever and highly pathogenic avian influenza H5N1, or a disease with fast transmissibility such as the 2009 swine-origin H1N1 in human populations, the real-time integration of clinical, microbiologic, environmental and epidemiologic data is crucially important to increase the efficiency and accuracy of surveillance. Spatio-temporal analysis of the updated confirmed cases is frequently compared with that of the reported suspected cases for investigating how the epidemic expands rapidly and where analysis can be further improved. These analytic results can then provide positive feedback to improve the surveillance system and can also point out those high risk areas in need of more attention.

In the analysis of epidemiologic data of an infectious disease, spatial and temporal clustering algorithms can be applied after collecting the spatial information using GPSs of geo-coded addresses of the studied cases and their exposure sites. Through fully understanding the epidemiologic characteristics of the outbreak disease, the specific prevention and control strategies can be formulated based on scientific data. This is most important for emerging infectious diseases when the etiologic agent is not known, such as the cross-country outbreak of severe acute respiratory syndrome (SARS) in 2003. For example, the modes of transmission, the time period that is most communicable, and the spatial patterns of the cases with different social contacts are unclear at the initial stage of disease outbreaks [38]. The subsequent cases after the introduction of prevention and/or control measures can also be carefully evaluated to verify the most effective strategy, using time-based integrated surveillance data. The visualized dynamic distributions of cases in various time periods and places at different levels of the public health system, from local, state/provincial to national and international, can be presented to generate hypotheses and to verify the success of containing the outbreak for decision-makers. Most importantly, evidence of spatial clustering along with other epidemiological findings and laboratory tests may indicate a possible infectious etiology for emerging infectious disease, similar to Epstein Barr virus for Hodgkin's disease [39].

4.2 Limitations of GIS Studies and Unsolved Problems

Several limitations of GIS studies need to be improved including data collection, quality of data to statistical methods and interpretation of the data.

4.2.1 Data Collection and Quality of GIS Data

In data collection, timely data and "modifiable area unit problem" (MAUP) – similar to ecological fallacies in epidemiology – are the two major barriers. Available timely data are important to fast-spreading infectious diseases such as most respiratory infections. In addition, the high quality of GIS data is another limitation in many developing countries. By contrast, those pointed address data of cases related to privacy are generally inaccessible in developed countries. Most importantly, for infectious diseases involving higher social stigma or patients' private life such as tuberculosis, sexual transmitted disease (STD) or AIDS, the pointed data for spatial cluster analysis will be very difficult to obtain. Then, the problem of spatial precision or polygon data will make it very hard to investigate the evolution of the outbreak by time and place simultaneously or to search for interesting hypotheses. Since most public health systems are governed by local departments, it is very likely those surveillance data are frequently aggregated into administrative units. Unfortunately, different densities and distribution patterns of disease, such as cholera in Figure 10-8, exhibited from different aggregated administration

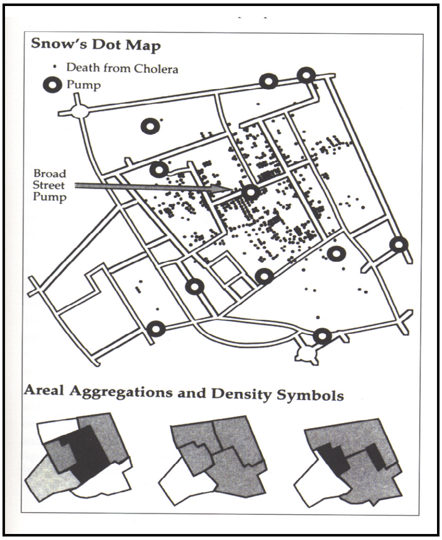


Figure 10-8. Different density maps and distribution patterns of cholera are shown by using different aggregative levels [40].

boundaries [40]. Therefore, researchers need to think about the most appropriate spatial units related to possible hypotheses at the initial stage of data collection.

4.2.2 Limitations in Statistical Methods and Interpretation of Data

Several unsolved statistical methods include too small value of relative risk to be detected, multiple covariate adjustment in spatial analysis, and better prediction during the occurrence of fast dynamic changes of cases in time and place. According to a simulated study [41], when the relative risk among study areas was lower than 1.5, the sensitivity of the detecting clusters dropped dramatically. However, the specificity can still keep a high performance level (above 95%). For an infectious disease with low pathogenicity, the relative risk is almost close to l and the sensitivity of the clustering algorithm should be low. Then, the false negative might be high. Under this circumstance, it is better to take specimens for laboratory diagnosis to increase the specificity. In addition, the demographic variables such as age structure and gender ratio are the frequently encountered confounding variables and other basic covariates should be adjusted for the risk. In dealing with fast-spreading infectious diseases, higher precision of the temporal and spatial units, real-time data gathering, and better statistical power all must be considered.

4.3 Future Directions

Many challenges of infectious diseases are common in different countries, including the impact of global warming on infectious diseases, emergency responses to EID, timely collection, and interchanges of high quality data to develop better control strategies. All these related issues need international collaboration. From our experiences, future global needs will involve flexible cluster methods to analyze irregular clusters, adjustment for personal risk factors, and application of Bayesian approaches to disease mapping and better prediction.

4.3.1 Flexibility of the Cluster Method in Detecting Irregular Clusters

Due to the natural barriers and the movement of humans, hosts and the vectors, the realistic shapes of clusters in most situations are irregular. If the algorithm intends to enhance the performance of detecting true clusters, flexibility of the shape will be needed. Risk-adjusted Nearest Neighbor Hierarchical clustering (RNNH), Support Vector Machine (SVM) [42], and ellipse shape Satscan [20] are all used to solve the problem of detecting clusters with irregular shape.

4.3.2 Adjustment for Personal Risk Factors

All ecologic data may involve the possible risk of "ecologic fallacy," and particularly the aggregated data might involve too many risk factors together. Detailed information is always difficult to collect through routine surveillance. In general, the demographic information such as age and gender, as total of ten covariates, would need to be adjusted using Satscan 8.0. In clustering algorithms, incorporating the data of important risk factors such as age, gender, occupation into analyses will help figure out the epidemiological conditions to form clusters.

4.3.3 Bayesian Method for Better Prediction [43]

Bayesian hierarchical spatial models have become widespread in disease mapping and ecologic studies of health-environment associations. In order to use posterior distribution of space–time interactions for predictions, information over space and time must be applied to estimate typical patterns for each area. Based on the extension of the Bayesian hierarchical models, the problems in detecting small numbers of events, particularly a small incidence of cases in the early wave of an outbreak, may soon be overcome in the future.

ACKNOWLEDGEMENTS

We greatly appreciate the National Health Research Institute (NHRI), Centers for Disease Control in Taiwan (Taiwan-CDC) for their financial support of research grants on GIS studies of dengue and establishing a syndromic surveillance system, respectively. In addition, we sincerely thank Dr. Michael Zilliox and Dr. Jaikumar Durisawamy at the Department of Microbiology and Immunology at Emory University in Atlanta, U.S.A. and Mr. Jim Steed in Kaohsiung, Taiwan for their efforts in English editing. We would like to express our sincere gratitude for the critical comments on surveillance by Dr. James Chin from the School of Pubic Health, University of California at Berkeley and on GIS and spatial statistics by Dr. Tzai-Hung Wen at the Institute of Epidemiology of DOH-NTU Infectious Diseases Research and Education Center, Taipei, Taiwan.

QUESTIONS FOR DISCUSSION

1. How can different clustering methods be applied to infectious diseases with various modes of transmission?
2. Are there any differences in using spatio-temporal analysis methods to analyze the data of an acute infectious disease versus a chronic disease?
3. Do you agree that the irregular clustering shapes and Bayesian model may enhance the capability to detect the true clusters?
4. Real-time syndromic surveillance is important for the early detection of abnormal events. Which clustering methods would you use to detect an early outbreak in a real-time manner?

REFERENCES

1. Elliott P, Wartenberg D: Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives* 2004, **112**(9):998–1006.
2. Spatial epidemiology [http://en.wikipedia.org/wiki/Spatial_epidemiology].
3. Gesler W: The uses of spatial analysis in medical geography: a review. *Social science & medicine (1982)* 1986, **23**(10):963–973.
4. Peterson AT: Ecologic niche modeling and spatial patterns of disease transmission. *Emerging infectious diseases* 2006, **12**(12):1822–1826.
5. Ali M, Emch M, Donnay JP, Yunus M, Sack RB: Identifying environmental risk factors for endemic cholera: a raster GIS approach. *Health & place* 2002, **8**(3):201–210.
6. Teutsch SM, Churchill RE: Principles and Practice of Public Health Surveillance, 2nd Edn. New York, NY: Oxford University Press; 2000.

7. Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, Elliott P: Statistical issues in the analysis of disease mapping data. *Statistics in medicine* 2000, **19**(17–18):2493–2519.
8. Wu TS, Shih FY, Yen MY, Wu JS, Lu SW, Chang KC, Hsiung C, Chou JH, Chu YT, Chang H *et al*: Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan. *BMC public health [electronic resource]* 2008, **8**:18.
9. Jackson ML, Baer A, Painter I, Duchin J: A simulation study comparing aberration detection algorithms for syndromic surveillance. *BMC medical informatics and decision making [electronic resource]* 2007, **7**:6.
10. Hutwagner L, Browne T, Seeman GM, Fleischauer AT: Comparing aberration detection methods with simulated data. *Emerging infectious diseases* 2005, **11**(2):314–316.
11. Stroup DF, Williamson GD, Herndon JL, Karon JM: Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in medicine* 1989, **8**(3):323–329; discussion 331–322.
12. Stroup DF, Wharton M, Kafadar K, Dean AG: Evaluation of a method for detecting aberrations in public health surveillance data. *American journal of epidemiology* 1993, **137**(3):373–380.
13. Williams SM, Parry BR, Schlup MM: Quality control: an application of the cusum. *BMJ (Clinical research ed.)* 1992, **304**(6838):1359–1361.
14. Longley PA, Goodchild MF, Maguire DJ, Rhind DW: Geographic Information System and Science. England: John Wiley & Sons, Ltd; 2001.
15. Kulldorff M: Statistical methods for spatial epidemiology: Tests for randomness. *GIS and Health* 1998.
16. Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR: Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American journal of public health* 1998, **88**(9):1377–1380.
17. Kulldorff M, Nagarwalla N: Spatial disease clusters: detection and inference. *Statistics in medicine* 1995, **14**(8):799–810.
18. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F: A space-time permutation scan statistic for disease outbreak detection. *PLoS medicine* 2005, **2**(3):e59.
19. Kleinman KP, Abrams AM, Kulldorff M, Platt R: A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and infection* 2005, **133**(3):409–419.
20. Kulldorff M, Huang L, Pickle L, Duczmal L: An elliptic spatial scan statistic. *Statistics in medicine* 2006, **25**(22):3929–3943.
21. Tango T, Takahashi K: A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics [electronic resource]* 2005, **4**:11.
22. Anselin L: Local indicators of spatial association – LISA. *Geographical Analysis* 1995, **27**:93–116.
23. Kelsall JE, Diggle PJ: Non-parametric estimation of spatial variation in relative risk. *Statistics in medicine* 1995, **14**(21–22):2335–2342.
24. Wen TH, Lin NH, Lin CH, King CC, Su MD: Spatial mapping of temporal risk characteristics to improve environmental health risk identification: a case study of a dengue epidemic in Taiwan. *The Science of the total environment* 2006, **367**(2–3):631–640.
25. Tango T: Score tests for detecting excess risks around putative sources. *Statistics in medicine* 2002, **21**(4):497–514.

26. Stone R: Investigation of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in medicine* 1988, **7**:649–660.
27. Knox G: The detection of space-time interactions. *Applied Statistics* 1964, **13**:25–29.
28. Pike MC, Smith PG: Disease clustering: a generalization of Knox's approach to the detection of space-time interactions. *Biometrics* 1968, **24**(3):541–556.
29. Gilman EA, McNally RJ, Cartwright RA: Space-time clustering of acute lymphoblastic leukaemia in parts of the U.K. (1984–1993). *European Journal of Cancer* 1999, **35**(1):91–96.
30. Onozuka D, Hagihara A: Geographic prediction of tuberculosis clusters in Fukuoka, Japan, using the space-time scan statistic. *BMC infectious diseases [electronic resource]* 2007, **7**:26.
31. Odoi A, Martin SW, Michel P, Middleton D, Holt J, Wilson J: Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *International journal of health geographics [electronic resource]* 2004, **3**(1):11.
32. Tran A, Deparis X, Dussart P, Morvan J, Rabarison P, Remy F, Polidori L, Gardon J: Dengue spatial and temporal patterns, French Guiana, 2001. *Emerging infectious diseases* 2004, **10**(4):615–621.
33. Kan CC, Lee PF, Wen TH, Chao DY, Wu MN, Lin NH, Huang SY, Shang CS, Fan IC, Shu PY, Huang JH, Pai L, King CC: Two clustering diffusion patterns identified from the 2001–2003 dengue epidemics, Kaohsiung, Taiwan. *The American journal of tropical medicine and hygiene* 2008, **79**(3):344–352.
34. Guerra CA, Snow RW, Hay SI: Defining the Global Spatial Limits of Malaria Transmission in 2005. In: *Global Mapping of Infectious Diseases – Methods, Examples, and Emerging Applications.* Edited by Hay SI, Graham A, Rogers DJ. Oxford, United Kingdom: Academic Press; 2007.
35. Rogers DJ, Wilson AJ, Hay SI, Graham AJ: The Global Distribution of Yellow Fever and Dengue. In: *Global Mapping of Infectious Diseases – Methods, Examples, and Emerging Applications.* Edited by Hay SI, Graham A, Rogers DJ. Oxford, United Kingdom: Academic Press; 2007.
36. Blanton JD, Manangan A, Manangan J, Hanlon CA, Slate D, Rupprecht CE: Development of a GIS-based, real-time internet mapping tool for rabies surveillance. *International journal of health geographics* 2006, **5**:47.
37. Ward MP, Maftei D, Apostu C, Suru A: Geostatistical visualisation and spatial statistics for evaluation of the dispersion of epidemic highly pathogenic avian influenza subtype H5N1. *Veterinary research* 2008, **39**(3):22.
38. Chen Y-D, Tseng C, King CC, Wu TSJ, Chen H: Incorporating Geographical Contacts into Social Network Analysis for Contact Tracing in Epidemiology: A Study on Taiwan SARS Data. In: *Intelligence and Security Informatics: Biosurveillance.* Edited by Zeng D, Gotham I, Komatsu K, Lynch C, Thurmond M, Madigan D, Lober B, Kvach J, Chen H. Heiderberg, Germany: Springer-Verlag; 2007.
39. Alexander FE, Williams J, McKinney PA, Ricketts TJ, Cartwright RA: A specialist leukaemia/lymphoma registry in the UK. Part 2: Clustering of Hodgkin's disease. *British journal of cancer* 1989, **60**(6):948–952.
40. Koch T: Cartographies of Disease: Maps, Mapping, and Medicine. Redlands, CA: ESRI Press; 2005.
41. Aamodt G, Samuelsen SO, Skrondal A: A simulation study of three methods for detecting disease clusters. *International journal of health geographics [electronic resource]* 2006, **5**:15.

42. Zeng D, Chen H, Lynch C, Eidson M, Gotham I: Infectious Disease Informatics and Outbreak Detection. In: *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. Edited by Chen H, Fuller SS. New York: Springer; 2005.
43. Abellan JJ, Richardson S, Best N: Use of space-time models to investigate the stability of patterns of disease. *Environmental health perspectives* 2008, **116**(8):1111–1119.

SUGGESTED READING

1. International Journal of Health Geographics, http://www.ij-healthgeographics.com
2. Waller LA, Gotway CA: Applied Spatial Statistics for Public Health Data. Hoboken, NJ: John Wiley & Sons, 2004.
3. Lawson AB: Statistical Methods in Spatial Epidemiology, 2nd Edn. Hoboken, NJ: Wiley, 2006.
4. Lawson AB: Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Boca Raton: CRC Press, 2009.

ONLINE RESOURCES

Free Software:

1. Epi Info, http://www.cdc.gov/epiinfo/downloads.htm
2. Quantum GIS 0.9, http://download.qgis.org/downloads.rhtml
3. R, http://www.r-project.org/
4. Satscan, http://www.satscan.org/download.html
5. Geosurveillance, http://www.acsu.buffalo.edu/~rogerson/geosurv.htm
6. Online Periodic Regression Models, http://www.u707.jussieu.fr/periodic_regression/.
7. Geoda, http://geodacenter.asu.edu/

Free Maps:

1. World Shapefile, http://www.cdc.gov/epiinfo/shape.htm
2. Geography Network Explorer: http://www.geographynetwork.com/

Chapter 11

AGE-ADJUSTMENT IN NATIONAL BIOSURVEILLANCE SYSTEMS

A Survey of Issues and Analytical Tools for Age-Adjustment in Biosurveillance

STEVEN A. COHEN[1,2] and ELENA N. NAUMOVA[1,2]

CHAPTER OVERVIEW

The practice of biosurveillance primarily involves measuring disease cases accurately and precisely in a given population. However, measuring the size and composition of the actual population at risk for the diseases under surveillance is just as important, particularly when the objective of surveillance is to measure rates of disease. The purpose of this chapter is to explore the issues pertaining to selection of the population denominator in population surveillance, with a particular focus on age and age-adjustment. This chapter presents an overview of several data sources commonly available to surveillance and epidemiological professionals, along with a synopsis of graphical and statistical tools to help assess and adjust for age effects in disease patterns.

Keywords: Population dynamics; Surveillance; Age-adjustment; Age-period-cohort analysis; Standardization

[1] *Tufts University School of Medicine, 136 Harrison Avenue, Boston, MA 02111, USA*

[2] *Tufts University Initiative for the Forecasting and Modeling of Infectious Diseases (InForMID), Tufts University, 136 Harrison Avenue, Boston, MA 02111, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_11,

1. INTRODUCTION

1.1 Disease Surveillance

Disease surveillance is one of the critical functions of local, state, regional, and national public health departments. The development and maintenance of health information systems that accurately and precisely collect surveillance data and disseminate this information to relevant parties serves as the foundation for nearly every aspect of public health programs and policies. Improved prediction and prevention of disease can result if epidemiologists, public health organizations, local, state, and federal government, and medical institutions work together to ensure that all diseases are properly reported and documented. Much of the resources, energy, and scope of disease surveillance are focused on accurately capturing all diseases that occur in the population being assessed, and rightfully so. However, the population of the United States is changing rapidly in terms of both size and composition. In order to correctly characterize disease in the US and its constituent parts, attention must be paid to properly estimate the population at risk, the denominator of the basic measures of epidemiology and surveillance: disease prevalence and incidence. There are numerous considerations to take into account when selecting the appropriate population denominator in surveillance. The remainder of this chapter outlines several of the important considerations of selecting a population denominator, with particular attention focused on the issue of age-adjustment in disease surveillance.

1.2 Case Studies of Influenza: Age-Specificity Within Population Subgroups

Few diseases affect the entire population with the same intensity. Many diseases disproportionately affect one or more population subgroups more than the rest of the population. For example, the age-specific mortality rate from pneumonia and influenza (P&I) deaths is nearly 100 times higher in people aged 65 and over (22.1/100,000 person-years) than for children under 1 year old, the group having the second-highest influenza mortality rate (0.3/100,000 person-years). From 1990 through 1998, 90% of influenza-associated deaths occurred among the population age 65 and older (Thompson et al., 2004). The following describes two specific case studies illustrating age differences in influenza patterns in two U.S. population subgroups.

1.2.1 Influenza and Respiratory Infection Hospitalizations in Milwaukee, Wisconsin (1996–2006)

On a smaller scale, age-specific disease patterns are discernible even within population subgroups such as children. In a study of respiratory disease surveillance in children, data were abstracted from a database of medical billing claims from Children's Hospital of Wisconsin, located in Milwaukee, from 1996 to 2006. The objective of the study was to determine how age patterns of several respiratory diseases differed from that of influenza. The distribution of respiratory infections in the population 0–18 years of age in Wisconsin is shown in Figure 11-1. The risk of upper respiratory infection peaks at age one, and then, as is the case for many other diseases in young children, declines rapidly with age with the rate of decline decreasing as age increases. Age patterns of influenza are similar in shape to both total and upper respiratory infection patterns, except much lower in overall magnitude. To that end, there appears to be more uncertainty or variability in the average annual counts of influenza in this population, as the trend for influenza is less smooth than for respiratory infections.

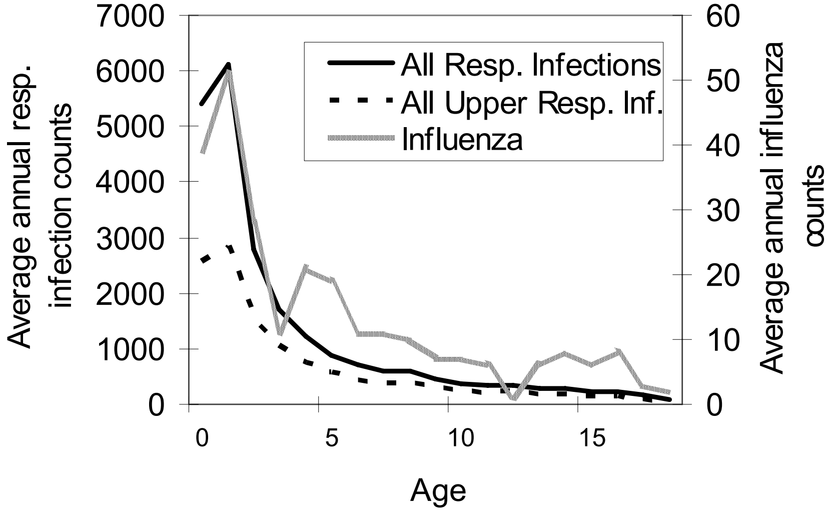


Figure 11-1. Average annual age-specific rates of respiratory infections and influenza in Wisconsin children (1996–2006) in a population of children from Children's Hospital of Wisconsin.

1.2.2 Pneumonia and Influenza in the US Elderly (1991–2004)

Although many diseases of early childhood peak in infancy and decrease with age in children, the opposite is true for many diseases in the elderly population. Disease incidence tends to increase with age in the older population, nearly exponentially, for some infections. To illustrate, approximately 14 million P&I cases from hospitalization claim records were abstracted from the Centers for Medicare and Medicaid Services' (CMS) surveillance data containing all Medicare-eligible hospitalizations in the United States from 1991 through 2004. As of July 1, 2005, there were approximately 42.4 million Medicare beneficiaries in the United States and its territories (The Henry Kaiser Family Foundation). To qualify for Medicare, a person must be a permanent resident of the United States. The person must be either age 65 or above and eligible to receive Social Security benefits, or under age 65 if already receiving Social Security benefits or if diagnosed with end-stage renal disease (Centers for Medicare and Medicaid Services). This case study focuses on the former: those age 65 and above.

Pneumonia and influenza cases peak in the elderly population in the early 1980s, yet because the population size decreases with increasing age, incidence rates increase nearly exponentially as age increases (Figure 11-2). These and

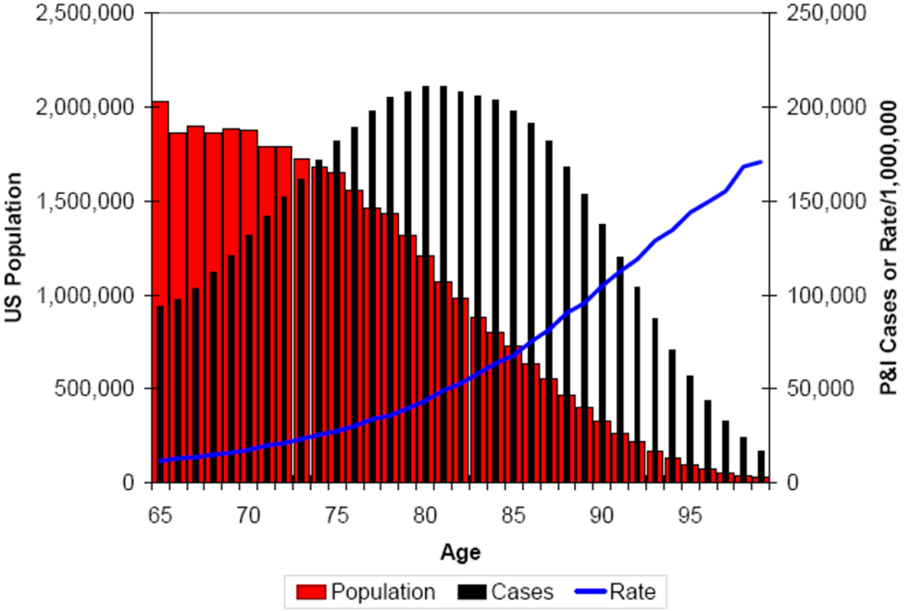


Figure 11-2. Average annual pneumonia and influenza hospitalization counts and rates (1998–2002) and population counts (2000) for the United States elderly (65+) (Source: Centers for Medicare and Medicaid Services).

other age trends in disease dynamics underscore the need to address age in population-based disease surveillance, particularly in the young and the elderly.

1.3 Population Dynamics

Population aging in the US results in both the numbers and proportions of people in the older age groups increasing. A few key forces are responsible for the aging of the population. First, in almost every developed nation, mortality has decreased markedly over the last century. In the U.S., this has resulted in life expectancies of 76.7 years for men and women combined in 2000. Life expectancy is projected to increase to between 80.5 and 82.9 years by 2050, with female life expectancy higher than that of males (Tuljapurkar et al., 2000). Decreasing mortality in older ages, coupled with a general decline in fertility (Bongaarts, 1998), has contributed to the "rectangularization" of the population (Cohen, 2003). This means that the population contains a relatively equal distribution of members in each age group, compared to other populations with higher fertility and mortality, where population size tends to decrease more sharply with age.

The second major demographic change affecting the US population with substantial implications for disease patterns in the elderly is the population size entering the older age groups. The Baby Boomers, born in the two decades following World War II, will begin to turn 65 in 2011, and by 2030 the number of people age 65 and over in the US is projected to be 71.5 million, compared to just 35 million in 2000, which represents just over 20% of the projected population (US Census Bureau Population Projections Website).

Population size and shape are of particular importance when describing infectious disease patterns in the population. The older population, as discussed above, is one of the population subgroups most detrimentally affected by morbidity and mortality from a variety of enteric diseases (Federal Interagency Forum on Aging-Related Statistics). Moreover, the oldest old are at particularly high risk of many infectious diseases. By 2050, persons age 85 and above are projected to represent 5% of the population, compared to just 1.5% in 2000, and the number of people in the US aged 85 and over is projected to be 21 million, compared to 4.1 million in 2000. For comparison, the youth dependency ratio (the ratio of persons less than 18 years of age to persons age 18–64), declined since the mid-century Baby Boom, and is projected to stabilize at approximately 40 youths per 100 working-age adults between now and 2040 (US Census Bureau Population Projections Website). Changes to the US population structure are of particular importance when considering that the two population subgroups that change in size and composition the most, the elderly and children, tend to be the same populations that experience the highest levels of disease.

2. DENOMINATOR DATA SOURCES

In population-based disease surveillance, special attention should be paid to the proper selection of the population denominator to provide an accurate picture of the population at risk of the disease. The US Census Bureau is possibly the most cited source for population data used for studies of disease in the US. Caution should be exercised when using Census data as population data for disease surveillance, however. There are several issues the researcher should consider when doing so, including:

1. Do the Census-based population counts cover the geographic population at risk?
2. Does the time interval the population counts represent match that of the disease surveillance data?
3. Do the Census population counts have the age breakdown (e.g., single year, 5 year age groups, etc…) to coincide with the disease counts?

For example, in the literature related to influenza research, several longitudinal population-based studies have employed US Census estimates for individual intercensal years (Simonsen et al., 2005; Rizzo et al., 2006). However, in some longitudinal influenza studies, researchers examined unadjusted numbers of influenza cases over time, which does not account for population change that may have played a role in causing numbers of cases to increase or decrease over time (Simonsen et al., 2000). This is analogous to using data from one census and applying that population longitudinally, as is done in some studies in other disciplines. Other influenza studies performed involving population estimates mentioned the use of Census data, but did not specify which data was used, for example, decennial census data, intercensal estimates, linearly interpolated data, or other related data (Thompson et al., 2004; Fry et al., 2005). Two of the main types of Census data publicly available to public health researchers are described below.

2.1 Decennial Census

The US Census Bureau conducts a national census every 10 years. This analysis uses data from Census 2000. In the decennial censuses, US households were sent a Census form to complete. Most households were sent a "short form," which asks basic demographic information about the household, including number of residents and ages of household members. One-sixth of households were sent a "long form," which is basically population-based, consisting of many questions related to socioeconomic status as well as demographic information not asked in the standard short form. The short form questions, mostly pertaining to population and related statistics, are summarized on various geographical levels in Summary File 1. Long form questions are generally tabulated in Summary File 3. Both files are publicly available on the Census website.

The data files from the decennial US Census are extremely rich and contain extensive and detailed information on the population structure of the United States, states, regions, and local areas. Most of these datasets are easily accessible and the online template found for data extraction is user-friendly. The main drawback of this data is the fact that these data exist only at 10-year intervals. In 1999, for example, the latest available data available from the decennial census represented the population 9 years prior. As discussed above, the population composition of the United States has changed and continues to change rapidly over time, and using outdated data might not reflect the true population composition over the time period of disease surveillance.

If the objective of disease surveillance is to examine temporal disease patterns for one or more decades of time, researchers should consider using some form of interpolation based on the decennial censuses. Although the population structure does not change exactly linearly from one census to the next, linear interpolation has been shown to closely approximate the population of study over time, particularly for large areas such as states and regions (Cohen et al., 2008).

2.2 Intercensal Population Estimates

The US Census Bureau provides population estimates for the years between and including the decennial census years. The Population Estimates Program, a division of the Census Bureau, produces these data, which can be found on the main Census Bureau website. The data are mostly publicly available and downloadable, but require some formatting manipulations to use in commercial spreadsheets and statistical software.

There are several limitations to the use of intercensal data in surveillance studies. The first is data accessibility. While the Census Bureau maintains a website containing detailed intercensal data, these data may not be available for all desired years and geographical levels that coincide with continuous surveillance data. Likewise, data is not available for all desired racial and ethnic groups, age categories, and other related factors at all geographic levels. As described above, the interface for extracting intercensal data is not as user-friendly as that of American FactFinder, which is used to extract decennial census data. Also, the sheer size of the data can be problematic not only for downloading, but also for manipulating within disease databases. For example, the size of county-level data files containing intercensal counts for each year by single year of age and gender is 229 MB.

Despite these drawbacks, intercensal data or interpolated data should be considered, whenever possible, to assign appropriate and corresponding population data to surveillance data to obtain population incidence or prevalence, particularly when assessing historical surveillance data over multiple years. These, together with specialized databases (e.g., Medicare enrollment files), can be utilized to draw a more complete and accurate picture of the population at risk for disease. The Census Bureau itself actually encourages the use of these data for disease surveillance. According to the US Census Bureau, some of the applications of intercensal estimates include federal funding allocations, denominators for per capita time series, as survey controls, in monitoring recent demographic changes, and, as in the case of this study, as denominators for vital rates (US Census Bureau Population Estimates Website).

3. GRAPHICAL TOOLS TO ASSESS AGE PATTERNS OF DISEASE

3.1 Population Pyramids

To illustrate the age and sex composition of a population, the population pyramid is a practical graphical tool that consists of two histograms aligned on the vertical axis, one typically representing males, the other representing females. The vertical axis itself represents age or age groupings. As with any histogram, the length of the bar represents the size of the population, either actual numbers or proportion of the total population in that age group.

An example from the US Census Bureau for the years 1970 and 2000 is shown in Figure 11-3. In these pyramids, age is broken down into 5-year age groups; males are shown on the left bars and females on the right bars. The length of the bar shows the total population counts for each gender and age group. The shape of the pyramid shows several properties of the population. In 1970, the US population could be described as young, in that the largest age-specific populations were those between ages 0–4 and 25–29. The population of older adults was comparatively small. The population pyramid in 1970 is roughly symmetric, except at the oldest ages, suggesting that there were no substantial discrepancies in population size between males and females. In 2005, the population shape became more rectangular, meaning that there was a more even distribution of population by age in the US. Because the bars represent population size, not proportion, it can also be observed that the overall population size increased between 1970 and 2000, because of the larger bar sizes in 2000. At the oldest ages, the female population is larger than the male population. This was true in 1970, but the magnitude of the absolute difference between males and females grew between 1970 and 2000.

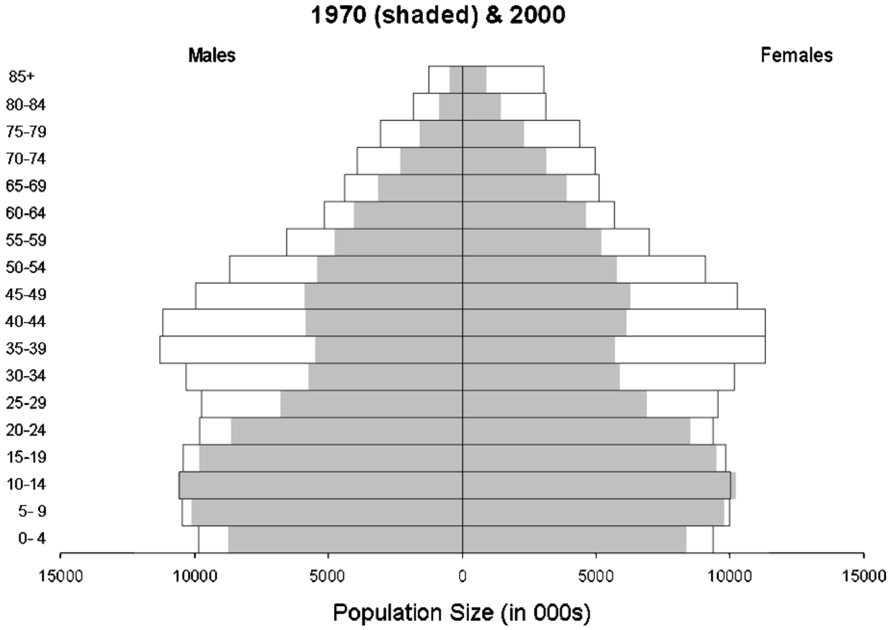


Figure 11-3. Population pyramids for US, 1970 and 2000.

3.1.1 Disease Pyramids

The utility of population pyramids extends to more than just the visualization of population alone. Population pyramids can also be adapted to describe age distributions of diseases, including counts and rates. Disease pyramids for upper respiratory diseases from the Children's Hospital of Wisconsin are shown in Figure 11-4. The disease pyramid illustrates that the majority of childhood upper respiratory hospitalizations occur in the population ages 0–2, with a peak at age 1. One of the important features of population or disease pyramids is that they allow the visual comparison of two groups by age; typically those two groups are males and females. In the case of the Children's Hospital data, we observe that males had slightly higher total counts of hospitalizations than females.

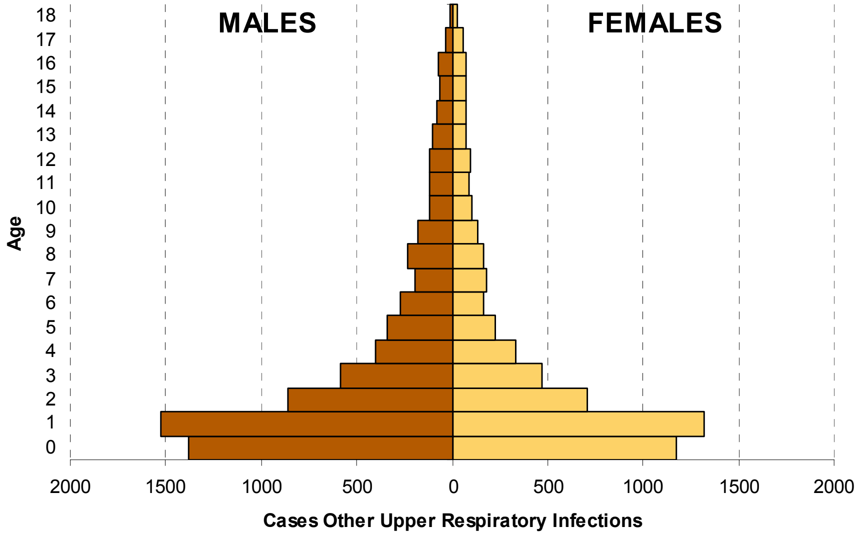


Figure 11-4. Disease pyramid showing average number of total annual upper respiratory infection cases extracted from Children's Hospital of Wisconsin (1996–2006).

A closer inspection of individual diseases reported by Children's Hospital suggests age and gender patterns of certain diseases (Figure 11-5). Specifically, the *sex ratio*, the ratio of asthma cases comparing females to males, differs by age in this population. Between ages 0 and 13, there were between 40 and 60 new cases of asthma in females for every 100 cases in males, but this ratio increased to nearly 100 female cases to every 100 male cases for children between age 14 and 18. Whether these differences are due to true patterns of disease in this population, or due to reporting practices or some other related factor is a matter of further study. Thus, disease pyramids can be employed to visually assess age patterns of disease and how those age patterns differ by gender or some other binary or categorical variable. Caution must be used when interpreting data from disease pyramids, however, and should generally be used only in conjunction with population pyramids for the area under surveillance. In the case of the Children's Hospital of Wisconsin, it could be speculated, for example, that as children enter the teenage years, they may be less likely to receive their healthcare from a children's hospital than younger children, which may influence, in part, the overall decline of asthma cases with age in this population of children.

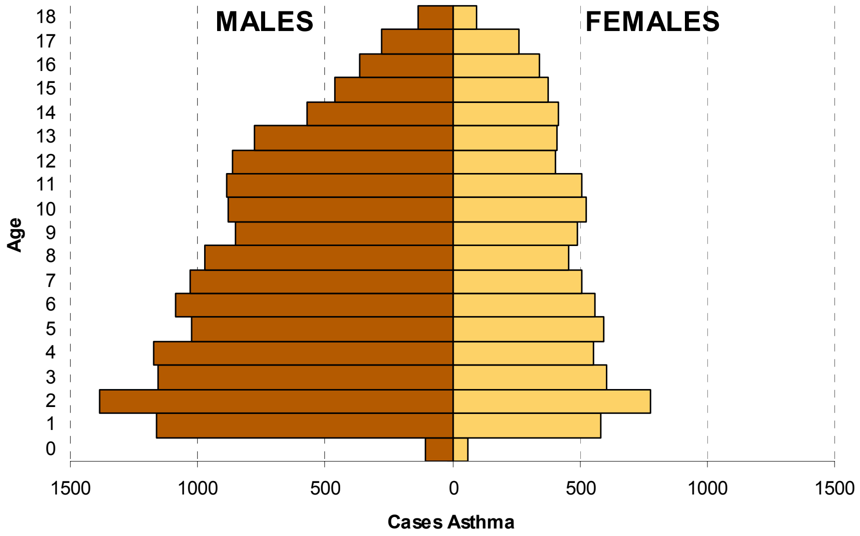


Figure 11-5. Disease pyramid showing average number of total asthma cases extracted from Children's Hospital of Wisconsin (1996–2006).

3.2 Lexis Surfaces

One of the major limitations of population and disease pyramids to show age effects is that they represent only one point in time, or, in the case of the above data, an average over time. A series of population or disease pyramids can be used to show temporal changes to the distribution of disease or population by age through presenting multiple pyramids at once, or through animation. However, presentation of multiple pyramids on a screen or page becomes cumbersome, and animation requires advance technical capabilities and still can only show population or disease pyramids one-at-a-time.

To overcome this limitation, users of historical surveillance data can employ the use of a graphical tool originally developed for use in demography, which can easily be adapted for use in public health surveillance. This tool, the Lexis diagram, is basically a three-dimensional chart with age on the vertical axis plotted against time on the horizontal axis (Vandeschrick, 2001). The z-axis for a Lexis surface can be age- and time-specific disease rates or counts. In traditional Lexis surfaces, the horizontal distance represented on the x-axis between two time points must be the same as the vertical distance between the same age differences on the y-axis (Figure 11-6). However, this rule can be relaxed, depending upon the data available and the level of desired detail.

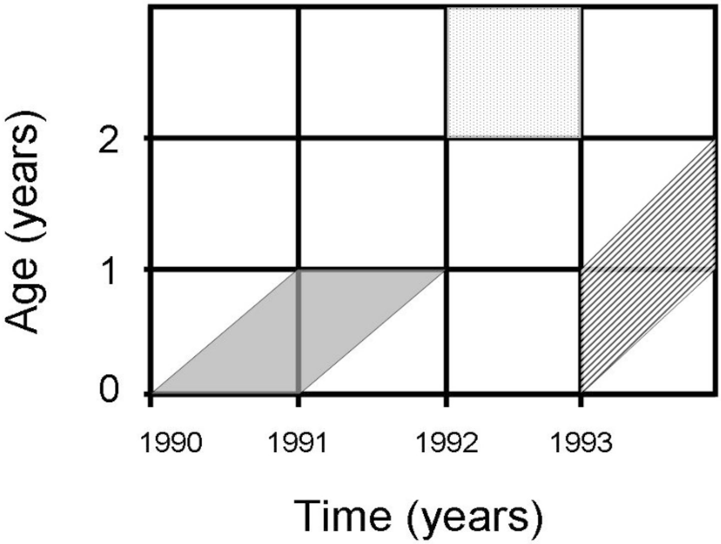


Figure 11-6. Schematic of Lexis diagram with illustration of cohort and period rates.

In the Lexis diagram, diagonal lines represent a single individual's life line. A collection of life lines can be viewed in a cohort or period perspective. In the Lexis diagram illustrated in Figure 11-6, the light gray square represents the age-specific disease rates for all 2-year-olds that occurred in 1992. The solid parallelogram in the lower left represents one type of cohort disease rate: the 1-year disease rate for all those born in 1990. The hashed parallelogram in the lower right is another type of cohort disease rate. This parallelogram represents the annual cohort disease rate for all those who were age 0 on January 1, 1993.

By incorporating the third dimension expressed by a color or texture scheme, Lexis surfaces can be adapted for a variety of surveillance settings. Modified Lexis surfaces were used in a 2007 study of influenza in New York City. This study explored the relationships between peak timing and severity of respiratory illness and influenza outbreaks by aggregated age groups for five influenza seasons (Olson et al., 2007). The counts and rates were aggregated into broad, uneven age groups, and thus did not focus specifically on the relationship between age and respiratory illness within population subgroups.

Using data from the Centers for Medicare and Medicaid Services, national pneumonia and influenza cases were tallied by age and influenza season, defined as July 1 through June 30 of the following year for the 65+ population. The Lexis surface illustrates the intensity of disease in the population as well as age patterns of disease in each influenza season (Figure 11-7). For each season, the maximum number of cases occurred in those between age 75 and 85. Clear differences among influenza seasons can also be observed.

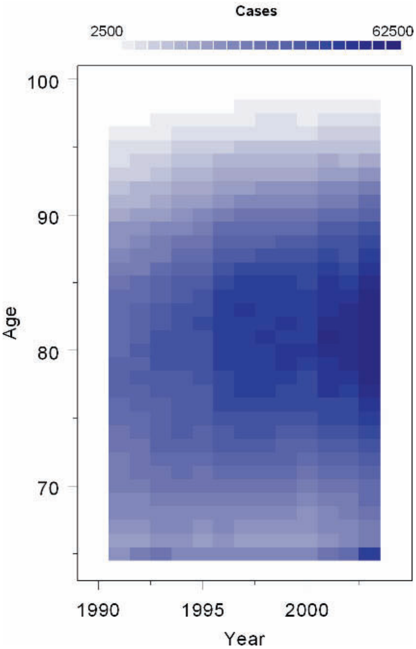


Figure 11-7. Pneumonia and influenza cases by year and age, July 1991–June 2004 (Source: Centers for Medicare and Medicaid Services).

4. ANALYTICAL TOOLS FOR AGE-ADJUSTMENT

4.1 Age-Period-Cohort Analysis

4.1.1 Background

Age-period-cohort models allow for the possibility that three related, but separate parameters – age, period, and cohort – are associated with patterns of morbidity and mortality (Sacher, 1956) and can be mathematically assessed using linear models accounting for each of these factors separately (Collins, 1982). This approach has particular utility in modeling historical surveillance data. The age-period-cohort approach has been applied extensively for the study of several diseases and mortality. Wilmoth and colleagues demonstrated that after age and period effects were taken into account, cohort effects on mortality were consistent and strong, providing a more comprehensive picture of mortality than the more traditional age and period models (Wilmoth et al., 1990). In a study of breast cancer in Spanish women, after age and period effects were taken into account, women born in the 1950s were three times more likely to die of breast cancer than those born in the 1890s (Cayuela et al., 2004). A modified version of the age-period-cohort model was employed to demonstrate key cohort differences in melanoma mortality using population-based surveillance data in Canada (MacNeill et al., 1995). The findings of the aforementioned research emphasize the importance of assessing these three related, but distinct, demographic dimensions of disease and mortality patterns.

Findings from these and other age-period-cohort modems is suggestive that certain exogenous factors exist for any given birth cohort, and that those factors can be directly or indirectly related with the overall health and viability of the individuals in that birth cohort (Derrick, 1927). Hobcraft and colleagues argue further that the idea of a cohort effect is valid only in the sense that cohorts are representative proxies for a set of underlying conditions that are common to people who are born in or live through contemporaneously and at similar ages. In other words, cohorts themselves are not the cause of any observed cohort effects, and thus must be treated accordingly (Hobcraft et al., 1982). Although the ascertainment of the specific set of environmental or socioeconomic conditions that account for any observed patterns of disease cannot be immediately determined using this approach, the age-period-cohort approach allows the researcher to assess general age, temporal, and cohort patterns of disease simultaneously.

4.1.2 Model Specification

The basic model equation is shown in Equation 11-1.

$$\log\left(\mathrm{P\&I\,Rate}_{\mathrm{ap}}\right) = \alpha_{\mathrm{a}} + \beta_{\mathrm{p}} + \gamma_{\mathrm{c}} \tag{11-1}$$

The parameters α_a, β_p, and γ_c, represent age, period, and cohort effects, respectively. Since, by definition, the sum of age and cohort effects equals the period effect using all possible age, period, and cohort effects, this model can be modified to account for the perfect collinearity that would result from such a model. By including indicator variables to represent cohort, one or more of these variables can be omitted from the model so as to reduce the possibility of collinearity from the identity properties of age, period, and cohort.

4.1.3 Graphical Tools

The Lexis surface is a logical and effective graphical tool for age-period-cohort models of disease rates. In the Lexis surface, vertical peaks and valleys are the period effects, horizontal peaks and valleys represent age effects, and diagonal patterns are cohort effects. Figure 11-8 illustrates how the Lexis surface reflects age, period, and cohort effects for influenza surveillance, where the horizontal axis is influenza season, defined as 1 year, July 1 through June 30, and the y-axis is age.

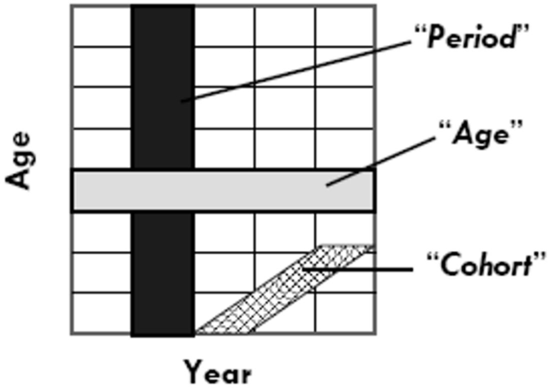


Figure 11-8. Schematic Lexis surface for age-period-cohort models.

4.2 Standardization and Decomposition

A crude disease rate is a simple and straightforward measure of disease burden in a population. It is defined simply as the total number of disease cases, incident or prevalent, divided by the total population count. As discussed above, population composition may differ among populations. For example, consider two hypothetical populations, one from State A and the other from State B. Let us assume that the population of State A is proportionately older than the population of State B. If there is a disease under surveillance that is more prevalent in older ages, the crude disease rate in State A will be greater than the rate in State B, even if the age-specific disease rates are identical. While this may accurately reflect that there is overall more disease in State A than in State B, the calculated crude rates mask the fact that the differences in disease burden between State A and State B are actually due to differences in population age composition, not differences in the magnitude of disease occurrences.

Standardization is a common method of accounting for such differences in age composition among populations when estimating disease rates. In practice, standardizing rates typically involves applying age-specific rates or counts comparing two or more populations using a single, common population structure, called a population standard. There are two types of standardization available for biosurveillance systems: direct and indirect standardization. Both can be utilized for comparing one region over time or comparing multiple regions, and both can be applied for the entire population age spectrum, or just a subset of the age spectrum.

4.2.1 Direct Standardization

Direct standardization is used when age-specific rates from two or more populations are known and applied to a common population standard. Table 11-1 shows counts and rates for a hypothetical disease, and actual population counts by broad age categories for Arizona and Utah.

Table 11-1. Disease rates for Arizona and Utah.

	Disease Counts		Population Counts[a]		Disease Rates (Per 1,000)	
Age	Arizona	Utah	Arizona	Utah	Arizona	Utah
0–19	909	554	1,518,188	810,977	0.599	0.683
20–39	3,067	2,054	1,498,212	702,911	2.047	2.922
40–59	4,212	1,944	1,242,696	466,604	3.389	4.166
60+	7,034	2,073	871,536	252,677	8.071	8.204

[a]Source US Census Bureau, 2000 decennial census

In this case, the crude disease rate for Arizona and Utah is identical, 2.97 per 1,000, which can be obtained by dividing the number of cases by the population at risk, which is 15,222 cases/5,130,632 people for Arizona and 6,625 cases/2,233,169 people for Utah. However, the age-specific rates are consistently higher in Utah than in Arizona for all four age groups. Utah also has a younger population, where 11.3% of the total population is age 60 and above, compared to 17.0% in Arizona. Likewise, the percent of the population under age 20 is 36.3 and 29.6% for Utah and Arizona, respectively.

To make the populations more comparable, direct standardization provides better estimates of disease burden in Arizona and Utah, since the age-specific disease rates are known in both states. In direct standardization, age-specific disease rates are applied to a common population. Here are the steps for direct standardization.

1. The first step in standardization is selecting the population standard. The choice of population standard is arbitrary; common choices include one of the populations being compared, the average population of the populations being compared, or some national or regional standard population.
2. Calculate the proportion of the entire population represented by each age group.

3. Multiply the age-specific disease rates for each population by the population proportions calculated in Step 2.
4. Sum the values calculated in Step 3. The results are age-adjusted disease rates.

The age-standardization for the Arizona and Utah example is shown in Table 11-2. The standard of choice was the US national population from Census 2000.

Table 11-2. Age standardization example with steps.

Age	Step 1	Step 2	Step 3		
	Standard Population (US)	Proportion of Total Population	Rate × Proportion (Arizona)	Rate × Proportion (Utah)	
0–19	80,473,265	0.285952	0.171285	0.195305	
20–39	81,562,389	0.289822	0.197985	0.846861	
40–59	73,589,052	0.261490	0.764109	1.089368	
60+	45,797,200	0.162735	0.677999	1.335078	
Total	281,421,906	1	1.81 per 1,000	3.47 per 1,000	← Step 4

Thus, the results obtained from direct standardization show age-adjusted disease rates of 1.81 and 3.47 cases per 1,000 for Arizona and Utah, respectively. The difference between the crude and the adjusted rates emphasizes the need to account for age in calculating population-based characteristics of disease occurrence. In this case, two states appear to have identical disease rates, when in reality, the disease rates in Arizona are much lower than those of Utah, with the age composition of the population responsible for masking the difference in rates.

4.2.2 Indirect Standardization

When age-specific disease rates are not available or if the data quality of these rates is poor due to small number of events or sample size, but the population structure and the crude disease rate are available, indirect standardization is a logical choice. One additional requirement for indirect standardization is that there must be some standard or comparison age-distribution of rates available. In the case of indirect standardization, this set of standardized rates is applied to the population composition of the study population to estimate a standardized prevalence or incidence rate ratio. Using the example of Arizona and Utah described in Sect. 4.1.1, Table 11-3 illustrates the steps involved in indirect standardization. For this example, it can be assumed that age-specific disease rates are unknown for both states, and only the crude rates are known. The standard disease rates will be provided by the national population.

1. Obtain a standard distribution of disease rates.
2. Obtain population distribution(s) of the population(s) under study.
3. Multiply the standard disease rates by each of the populations of study to estimate expected number of disease cases in each of the populations.
4. Sum the total number of expected cases.
5. To obtain the standardized disease rate, divide the expected number of cases in each population by the respective actual number of cases, and multiply that figure by the crude rate of the standard population:

$$\text{Standardized rate} = \frac{\text{Expected No. of cases}}{\text{Observed No. of cases}} \times \text{Crude rate in standard}$$

Table 11-3. First four steps of indirect standardization

Age	Step 1	Step 2		Step 3		
	Disease Rates in US/1,000	Population AZ (UT)	Composition	Expected Number of Cases AZ (UT)		
0–19	0.700	1,518,188	810,977	1,062.7	567.7	
20–39	2.323	1,498,212	702,911	3,480.3	1,632.9	
40–59	3.563	1,242,696	466,604	4,427.7	1,662.5	
60+	7.711	871,536	252,677	6,720.4	1,948.4	
Total	3.060	5,130,632	2,233,169	15,691	5,811	← Step 4

In this example, the standardized rate for Arizona is 15,691/15,222 × 3.060, which is equal to 3.154 cases per 1,000. The standardized rate for Utah is 6,625/5,811 × 3.060, which equals 3.489 cases per 1,000. These results, though starkly different than the results obtained through direct standardization, still indicate that disease rates are higher in Utah than in Arizona, which is consistent with the findings obtained through direct standardization. The choice of standardization method depends upon the surveillance data available to the researcher. Whenever possible, however, direct standardization should be used. Despite the choice of method, these findings emphasize the need to address the issue of confounding by age and provide two potentially useful tools for statistically controlling for this important characteristic in the calculation of prevalence or incidence rates from population-based disease surveillance.

4.2.3 Decomposition

To quantify exactly how much the difference between two disease rates – A and B – is due to differences in age composition between two populations, decomposition methods can be used. Consider an example of reported cryptosporidiosis in the elderly for two regions, the Northeast and the South. Let R_i and P_i represent the rate and proportion of the population represented in age group i, respectively. Superscripts will denote the region for each population in this example, but can denote different time points, states, etc… in other situations. The equation for decomposition analysis is found in Equation 11-2.

$$\Delta R = CDR^N - CDR^S = \sum_i 0.5\times[P_i^N - P_i^S][R_i^N + R_i^S] + \sum_i 0.5\times[R_i^N - R_i^S][P_i^N + P_i^S] \quad (11\text{-}2)$$

The first term represents the effect of age, and the second term represents the effect of rate itself. The crude rates of cryptosporidiosis are 5.588 per 10,000 for the Northeastern states and 5.241 per 10,000 for the Southern states, a difference of 0.347 per 10,000 or 6.5% (Table 11-4).

Table 11-4. Illustration of decomposition of differences between rates.

Age	*Pop.*^{N}	*Pop.*^{S}	P_i^N	P_i^S	R_i^N	R_i^S	Age Effect	Rate Effect
65–74	2,096,915	3,687,911	0.474	0.505	3.12	3.08	−0.0947	0.0196
75–84	1,644,519	2,603,338	0.372	0.356	5.98	5.55	0.0913	0.1566
85+	678,461	1,018,007	0.154	0.139	12.23	12.27	0.1804	−0.0059
Total	4,419,895	7,309,256	1	1			0.1769	0.1703

The results indicate that 0.177 or 51.0% of the difference in crude rates is attributable to differences in age composition, and 0.170 or 49.0% of the difference in crude rates is attributable to differences in age-specific rates of cryptosporidiosis.

This section discussed two-factor decomposition, which allows for the possibility of two factors, the difference in rate and the difference in age, to comprise the difference in disease rates completely. There are additional methods, such as three-factor decomposition, and two-factor decomposition with an interaction term, that are available for use if such a factor is of interest to the researcher, but these topics will not be discussed in this chapter. Please see the Suggested Readings for further details.

Standardization and decomposition can be applied in a variety of settings and for almost any disease for which there is variation in rates by age to account for compositional effects of age. The strength of standardization lies more in adjusting for age effects rather than quantifying age effects, whereas decomposition allows the researcher to quantify the effect of age, but cannot provide adjustments for age effects. It should be noted that there are no unique solutions to decomposition analyses; there are several possible ways to decompose a difference in disease rates between populations. However, the decomposition of a difference in rates by age using a two-factor decomposition provides an efficient and easily interpretable measure of the age effect in population-based disease rates.

4.3 Summary Disease Measures

In some situations, knowing age-specific disease rates may be useful with respect to understanding disease dynamics in the population, but may be cumbersome to deal with when analyzing surveillance data. One of the major shortcomings of standardization and decomposition is that the adjusted rates are still just one rate, and do not express the true underlying distribution of disease by age in the population. There are other, more case-specific measures available and others that can be developed to fulfill this need.

For example, using the Medicare data described above, it was observed that pneumonia and influenza hospitalization rates increase approximately exponentially between ages 65 and 99. Since P&I rates increase approximately exponentially with age in the elderly population, the age-acceleration coefficient can be defined as the log of the rate of increase in disease rates with age. The age increase of P&I is a relative measure of disease burden in the elderly. This is based on a new approach that uses and takes advantage of having disease rates by single-year of age (Figure 11-9).

The burden of pneumonia and influenza is highest in the oldest old, and this measure captures the relationship between age and disease rates. The standard approach of using overall age-adjusted rates gives priority to the age groups with the highest populations, which is generally the youngest elderly, given that population size decreases with increasing age. This outcome variable, however, effectively weights each age equally, and since P&I rates increase substantially with age in the oldest old, this approach reflects P&I patterns across all ages simultaneously in the elderly population, including those at the highest risk of disease (Cohen et al., 2007). The age-acceleration coefficient is just one of numerous potential measures of disease burden that summarize disease patterns with respect to age into one meaningful measurement that can be compared over time and across geographic regions.

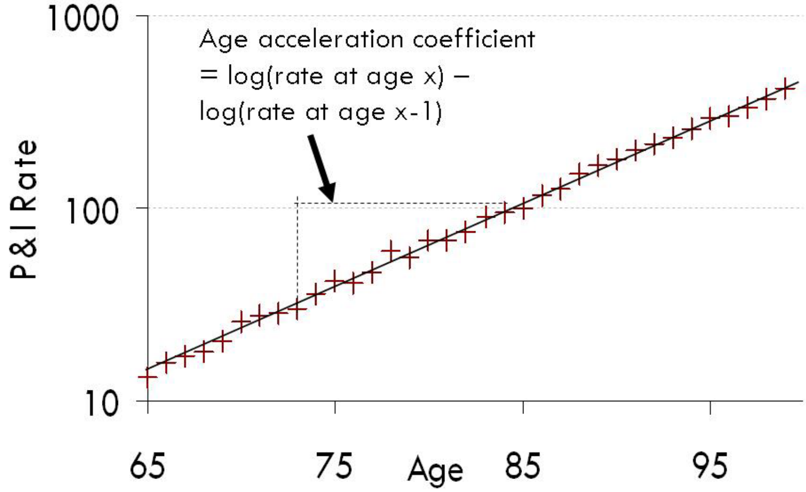


Figure 11-9. Illustration of age acceleration coefficient methodology.

5. CONCLUSIONS

Disease processes are inherently linked to population dynamics that play an important role in the magnitude and scope of disease distribution. Though often overlooked, age patterns of disease can serve as a potential confounder in the estimation of disease rates, but at the same time, can reveal important aspects of the disease and disease transmission within the population at risk. There are a variety of methods to account for age, some graphical, some statistical. Recognizing these effects and making appropriate adjustments should be a key component in disease surveillance. Seemingly small absolute differences in population counts can make a substantial difference in the estimation of age-specific rates, and that proper estimation of the denominator is critical in public health surveillance of disease.

ACKNOWLEDGEMENTS

We wish to thank Julia Wenger of the Tufts University School of Medicine for her assistance with organizing and preparing several of the databases used in this chapter. We also wish to thank the team of the Tufts Initiative for Forecasting and Modeling Infectious Diseases (InForMID) for valuable contributions. This research is being supported by the National Institutes of Health grants NIH-NIAID U19 AI62627, "Robust T-cell Immunity to Influenza in Human Populations" (PI J. Gorski, Project Leader - E.N. Naumova), and N01 AI50032, "Generation and Decay of Memory T cells in Young, Old, and Immunocompromised Populations" (PI J. Gorski, Project Leader – E.N. Naumova). We also wish to acknowledge our data providers: Centers for Medicare and Medicaid Services and Children's Hospital of Wisconsin.

QUESTIONS FOR DISCUSSION

1. What are the inherent problems with comparing raw disease counts over time in the US, besides the issue of age? What are the issues around taking an average annual disease rate for multiple years?
2. What kinds of circumstances could lead to a birth cohort experiencing higher or lower rates of disease than surrounding birth cohorts?
3. What does the approximate shape of the distribution of deaths by age look like for a developed society, such as the US? What diseases follow a similar pattern?
4. Assume that a disease peaks at age 0, declines rapidly above age 0, but the rate of decline decreases with age. What type(s) of summary measure(s) of disease can be estimated from this observation, akin to the age-acceleration coefficient described in Sect. 4.3?
5. What advantages and disadvantages does the estimation of cohort disease rates have over the estimation of period rates?

REFERENCES

Bongaarts J. Fertility and reproductive preferences in post-transitional societies. Popul Dev Rev 1998;27:260–81.

Cayuela A, Rodríguez-Domínguez S, Ruis-Borrego M, Gili M. Age-period-cohort analysis of breast cancer mortality rates in Andalucia (Spain). Ann Oncol 2004;15:686–8.

Centers for Medicare and Medicaid Services. Medicare eligibility tool. Website: http://www.medicare.gov/MedicareEligibility/Home.asp?dest=NAV |Home|GeneralEnrollment#TabTop. Accessed March 20, 2007.

Cohen JE. Human population: the next half century. Science 2003;302:1172–5.

Cohen SA, Naumova EN. Population dynamics in the elderly: the need for age-adjustment in national biosurveillance systems. Lect Notes Comput Sci 2007;4506:47–58.

Cohen SA, Oomer IA, Naumova EN. Regional differences in the estimation of influenza burden in the elderly: does choice of population denominator matter? Presented at the Population Association of America Annual Meeting, New Orleans, LA, April 16–19, 2008.

Collins JJ. The contribution of medical measures to the decline of mortality from respiratory tuberculosis: an age-period-cohort model. Demography 1982;19:409–27.

Derrick VPA. Observations of (1) errors of age in the population statistics of England and Wales and (2) the changes of mortality indicated by national records. J Inst Actuar 1927;58:117–59.

Federal Interagency Forum on Aging-Related Statistics. Older Americans 2004. *Key Indicators of Well-Being. Federal Interagency Forum on Aging-Related Statistics*. Washington, DC: U.S. Government Printing Office.

Fry AM, Shay DK, Holman RC, Curns AT, Anderson LJ. Trends in hospitalizations for pneumonia among persons aged 65 years or older in the United States, 1988–2002. JAMA 2005;294:2712–9.

Hobcraft J, Menken J, Preston S. Age, period, and cohort effects in demography: A review. Pop Index 1982;48:4–43.

MacNeill IB, Elwood JM, Miller D, Mao Y. Trends in mortality from melanoma in Canada and prediction of future rates. Stat Med 1995;14:821–39.

Olson DR, Heffernan RT, Paladini M, Konty K, Weiss D, Mostashari F. Monitoring the impact of influenza by age: emergency department fever and respiratory complaint surveillance in New York City. PLoS Med 2007;4:1349–61.

Rizzo C, Viboud C, Montomoli E, Simonsen L, Miller MA. Influenza-related mortality in the Italian elderly: no decline associated with increasing vaccination coverage. Vaccine 2006;24:6468–75.

Sacher GA. On the statistical nature of mortality with special reference to chronic radiation mortality. Radiology 1956;67:250–7.

Simonsen L, Fukuda K, Schonberger LB, Cox NJ. The impact of influenza epidemics on hospitalizations. J Infect Dis 2000;181:831–7.

Simonsen L, Reichert TA, Viboud C, Blackwelder WC, Taylor RJ, Miller MA. Impact of influenza vaccination on seasonal mortality in the US elderly population. Arch Intern Med 2005;165:265–72.

The Henry Kaiser Family Foundation: Medicare. Website: http://www.statehealthfacts.org/cgi-bin/healthfacts.cgi?action=compare&welcome =1&category=Medicare. Accessed March 20, 2007.

Thompson WW, Shay DK, Weintraub E, Brammer L, Cox N, Anderson LJ, Fukuda K. Mortality associated with influenza and respiratory syncytial virus in the United States. JAMA 2004;289:179–86.

Tuljapurkar S, Li N, Boe C. A universal pattern of mortality decline in G7 countries. Nature 2000;405:789–92.

United States Bureau of the Census. Projections of the US population: 1999–2100. Available at: http://www.census.gov/population/www/projections/natproj.htm. Accessed February 8, 2007.

US Census Bureau. Population Estimates. Website: http://www.census.gov/popest/estimates.php, Accessed February 22, 2008.

Vandeschrick C, The Lexis diagram, a misnomer. Demogr Res 2001;4:97–124.

Wilmoth J, Vallin J, Caselli G. When does a cohort's mortality differ from what we might expect? Popul English Selection 1990;2:93–126.

SUGGESTED READING

Rowland DT. *Demographic Methods and Concepts*. Oxford: Oxford University Press.

This is an essential overview of a variety of basic demographic concepts and procedures and serves as an excellent resource for public health researchers and practitioners into this relevant field.

Preston SH, Heuveline P, Guillot M. *Demography: Measuring and Modeling Population Processes*. Oxford: Blackwell Publishing.

This book provides an overview of some of the demographic theory and intermediate methods described in this chapter, as well as other, related demographic tools available for public health research. It is more mathematical and statistical than Rowland's text.

Du P, Coles FB, O'Campo P, McNutt LA. Changes in population characteristics and their implication on public health research. Epidemiol Perspect Innov. 2007;4:6.

This is an excellent article outlining some of the major challenges facing public health research in regard to changing population distributions and socioeconomic factors.

Tapia Granados JA. Economics, demography, and epidemiology: an interdisciplinary glossary. J Epidemiol Community Health. 2003;57:929–35.

This article contains a glossary of terms and concepts essential to the fields of demography, epidemiology, public health, and economics.

ONLINE RESOURCES

The US Census Bureau maintains an excellent and user-friendly database of data from the most recent decennial censuses. Most decennial Census data are available for many levels of geography, including states, countries, cities and towns, ZIP codes, census tracts and blocks, and much more. These data can be found on the Census Bureau's website, or on a data clearinghouse webpage known as Data Ferret.

The websites for this database can be found at:

- http://www.factfinder.census.gov/servlet/DatasetMainPageServlet?_program=DEC&_submenuId=&_lang=en&_ts=
- http://www.dataferrett.census.gov/

The US Census Bureau's website also maintains a database containing intercensal population estimates from the Population Estimates Program. This one page contains links to all the publicly available datasets and is organized by geographic level and contains data dating back to 1990.

- http://www.census.gov/popest/datasets.html

The archived population estimates, dating back in some cases to 1900, can be found at:

- http://www.census.gov/popest/archives/

Chapter 12

MODELING IN IMMUNIZATION AND BIOSURVEILLANCE RESEARCH

C. RAINA MACINTYRE[1,*], JAMES G. WOOD[1],
ROCHELLE WATKINS[2], and ZHANHAI GAO[1]

CHAPTER OVERVIEW

This chapter introduces the key concepts in mathematical modeling of vaccine-preventable diseases, and special features of vaccination such as herd immunity, disease elimination and waning immunity. It also reviews the interface of biosurveillance with monitoring and control of vaccine-preventable diseases.

Keywords: Vaccines; Immunization; Mathematical Modeling; Biosurveillance

1. INTRODUCTION

Before immunization was widespread, infectious diseases were the leading cause of death in children. Vaccination is named from the Latin "of a cow" (vaccines) in honor of Edward Jenner, who discovered smallpox vaccination. Louis Pasteur, in 1881, generalized the term "vaccination" to include preventive inoculation with all kinds of infectious agents. The first vaccines included smallpox (1804), plague (1890), diphtheria antiserum (1895) and typhoid (1899). The public health impact of vaccination in the twentieth century was significant. Campaigns to vaccinate for/against diphtheria were

[1*] *School of Public Health and Community Medicine, Faculty of Medicine, UNSW, NSW 2052, Australia. r.macintyre@unsw.edu.au*

[2] *Faculty of Health Science, Curtin University of Technology, Perth, WA 6845, Australia*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_12,

introduced in schools between 1933 and 1936 and for infants from 1940 to 1945; for pertussis in the early 1940s; and the triple antigen DTP was introduced in the early 1950s. The greatest success story has been the smallpox vaccine, which led to the global eradication of smallpox (Hinman, 1999).

Yet infectious diseases are still with us, resulting in deaths from diseases preventable by vaccines. In addition, emergence of new infections and re-emergence of old infections continue to challenge us. There are many historical examples of epidemics of previously rare diseases when immunization programs have failed or ceased. For example, there were two major epidemics of polio in Holland in 1984 and in 1991 in a religious group who refused immunization (Oostvogel et al., 1994). In the United Kingdom, epidemics of whooping cough in the 1970s and 1980s and of measles in current times have followed decreased immunization rates (Miller et al., 1992; Ashmore et al., 2007; Dobson, 2008). In the former USSR, collapse of public health infrastructure resulted in epidemics of diphtheria, which was previously rare. More than 115,000 cases and 3,000 deaths from diphtheria were reported from 1990 to 1997 in the Russian Federation (Centers for Disease Control and Prevention, 1996).

In 2000 The Institute of Medicine produced a report, Vaccines for the twenty-first Century, which attempted to rank priorities for vaccine development in the USA (Institute of Medicine, 2000). The highest priority was for vaccines against CMV, influenza, diabetes, multiple sclerosis, rheumatoid arthritis, group B streptococcus and pneumococcus (Institute of Medicine, 2000). The trend of vaccine development with more immunogenic and less reactogenic vaccines as well as more complex combination vaccines has seen a steep increase in the number of childhood vaccines routinely used in developed countries.

1.1 Role of Modeling for Vaccination Programs

In most sciences, research questions are answered by planned repeated experiments. For infectious diseases (IDs), experimenting in communities is rarely ethical or possible. Instead, we rely on observational data that are subject to reporting delays, supplemented by periodic population samples such as cross-sectional serosurveys. The fluctuating nature of infectious disease risks make mathematical models of pathogen transmission an attractive tool to aid policy formulation from such data. Modeling aids policy decisions on how best to respond to emerging infections such as SARS, pandemic influenza or deliberately released smallpox. It can also assist in determining and measuring the effect of introducing a new vaccine and the best age to administer it. Correct predictions of epidemics of measles in the United

Kingdom and in New Zealand based on modeling have emphasized its relevance to immunization policy (Gay et al., 1995; Roberts and Tobias, 2000).

Population effects such as indirect protection of unvaccinated individuals due to reduced transmission (herd immunity) are difficult to capture in clinical trials. Although specialized trial designs can measure aspects of herd immunity (e.g., the protection conferred to staff in a childcare centre where children are vaccinated for influenza), it is relatively simple to use models to calculate herd immunity effects from standard clinical trial data.

However, continued surveillance of vaccination-related data is necessary both to determine the epidemiologic impact and to refine models used for predictive purposes. For example, in the UK, the Netherlands and parts of the USA there has been a resurgence of Haemophilus Influenzae disease 10 years after commencement of population vaccination (Galil et al., 1999; Rijkers et al., 2003; Trotter et al., 2003). This is attributed to the change to a less effective vaccine, an accelerated schedule and reliance on a catch-up campaign. Modeling is a useful tool for probing the relative impact of these kinds of effects and, therefore, has an important role in monitoring, evaluating and predicting the impact of vaccination programs.

In the first half of this chapter, we introduce key concepts in infectious disease transmission and describe the Susceptible-Infectious-Recovered model. We then go on to illustrate the relationship between vaccine coverage, the reproduction number and disease elimination. We also summarise a number of the refinements of this model that are commonly used in immunization modeling today.

1.2 The Interface with Biosurveillance

Even the simplest mathematical models of disease transmission rely on good surveillance data to make meaningful predictions. Traditional surveillance tools for immunization tend to involve specific studies such as testing for immunity in cross-sectional studies. However, data streams such as notifications of confirmed infections and vaccination coverage data are now available at relatively short delays and considerable theoretical work has recently been devoted to analysis of outbreak data.

In the second half of the chapter, we cover some of the basic biosurveillance techniques for infectious diseases and their application to estimation of basic infectious disease parameters. Recent work that incorporates more structured transmission models into biosurveillance analysis is also described, with discussion of the strengths and weaknesses.

2. MODELING OF VACCINATION PROGRAMS

Mathematical modeling of infectious diseases has a lengthy history (Heesterbeek, 2002). There are now several quality texts which describe modeling approaches to infectious diseases in detail (Bailey, 1975; Anderson and May, 1999; Diekmann and Heesterbeek, 2000; Daley and Gani, 2001; Keeling and Rohani, 2007). In this section we illustrate the basic concepts and outline some refinements to this approach.

2.1 Key Concepts

2.1.1 The Basic Reproductive Number (R_0)

R_0 is the average number (technically a ratio) of secondary cases generated by a typical infectious person at the beginning of an epidemic of a given infectious disease in a fully susceptible population. While R_0 reflects the intrinsic transmission potential of the organism, it is also influenced by environmental factors such as genetics, behavior and climate, so that different values can apply in different communities. The lower the value of R_0, the easier it is to eradicate the disease, provided an effective vaccine exists.

In general, however, we deal with populations that have some immunity to disease. In this situation, we can define a related quantity R_t, which is the average number of secondary cases generated by a typical infectious person at time t. The value of R_t varies with time and will decrease as population immunity rises, reflecting a reduced potential for disease transmission.

2.1.2 Force of Infection

There are measurable risks of acquiring non-communicable diseases, and the situation is the same for communicable diseases. However, the ability of communicable diseases to be transmitted from person to person and the potential for immunity to arise from infections or immunization leads to a more specialized form of risk. The instantaneous risk of a susceptible person acquiring a communicable disease is known as the force of infection for that disease. The force of infection (often denoted by the symbol λ) is specific to the infectious agent and the population and time period of interest.

Like risks for non-communicable diseases, the force of infection is dependent on social and behavioral factors (e.g., hygiene in the case of Hepatitis A infection (Jacobsen and Koopman, 2005)) but a key difference is that it can also depend on the prevalence of infection. Acquisition of a communicable disease often requires contact with an infectious person, and in many cases the risk is proportional to the prevalence of infection with that disease in the community.

2.1.3 Disease States

For many communicable diseases (e.g., measles, rubella, smallpox), a population can be divided into three broad categories: Susceptible people (*S*), Infectious people (*I*) and Immune people (*R*) (Anderson and May, 1999). Although there are many exceptions to this simple framework, it provides a basis for a mathematical description of infectious disease transmission in a large population.

2.2 The *SIR* Model

Imagine that we wish to describe the transmission of an infectious disease in a large population of *N* individuals. Assuming that infections are directly transmitted from person to person, we can express the force of infection as:

$$\lambda(\mathrm{I}) = \beta \times \frac{\mathrm{I}}{N} \qquad (12\text{-}1)$$

Here we have introduced the parameter β representing the instantaneous rate at which infections are caused by a single infectious person. We also require a parameter to describe the rate of recovery from infection (denoted by γ). This parameter is the inverse of the duration of infection (*D*), so that $\gamma = 1/D$.

In Figure 12-1, a schematic of the *SIR* model is shown, with disease states and rates of transfer between these states

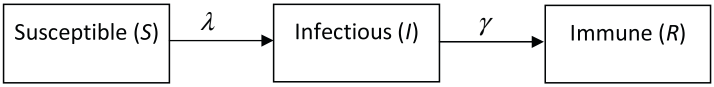


Figure 12-1. Schematic for *SIR* model, showing states and flow variables.

Although in reality infections lead to discrete changes in the numbers of individuals in each state, in large populations it is convenient to disregard this and allow a continuous flow between these states. This allows us to represent the *SIR* model as a system of differential equations:

$$\frac{dS}{dt} = -\frac{\beta IS}{N}$$
$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I \qquad (12\text{-}2)$$
$$\frac{dR}{dt} = \gamma I$$

This model leads to epidemic dynamics as illustrated in Figure 12-2 (β = 0.5, D = 6, N = 1 million).

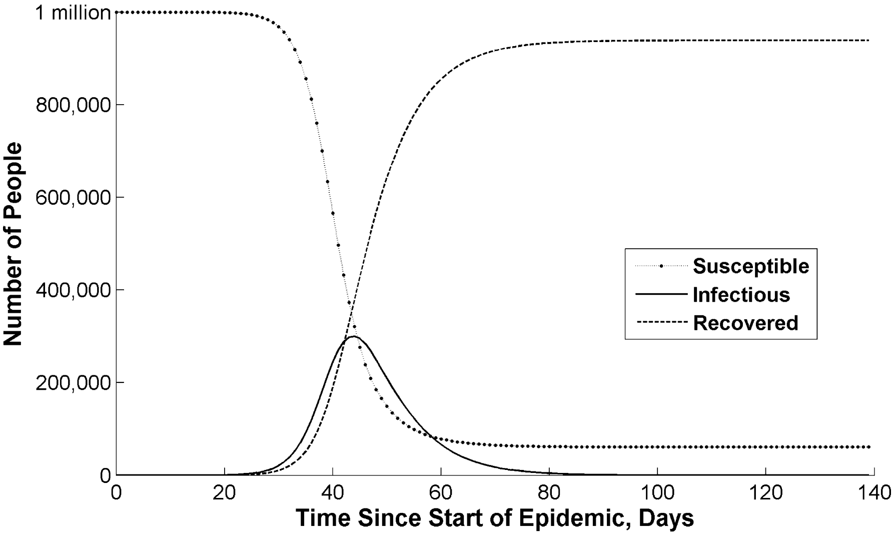


Figure 12-2. Change in susceptible, infectious and recovered populations as calculated with an *SIR* model with parameters $\beta = 0.5$, $D = 6$, $N = 1$ million.

2.2.1 The Basic Reproduction Number in the *SIR* Model

In this basic *SIR* model, there is a key parameter that determines if an epidemic occurs or not. Imagine that a single infectious case is introduced into an otherwise susceptible population at the beginning of their infectious period. At this time, $S \approx N$ so that $dI/dt \approx (\beta - \gamma)$I. Thus the number of infectious people will grow only when the ratio $\beta/\gamma > 1$. This ratio is called the basic reproduction number and is denoted by R_0, and for the *SIR* model defined in Equation 12-2 is given by:

$$R_0 = \frac{\beta}{\gamma} \qquad (12\text{-}3)$$

The dynamics of transmission in the *SIR* model change at $R_0 = 1$, with an epidemic occurring if $R_0 > 1$ but not if $R_0 \leq 1$.

2.3 Endemic Dynamics

How is infection sustained in a community? In the epidemic *SIR* model given in Equation 12-2, only one epidemic can occur, after which the majority of the community is left immune. The proportion left immune is equivalent to the attack rate (a_r) during the epidemic, and can be calculated by solving the non-linear equation:

$$(1-a\ )=e^{-a_rR_0} \tag{12-4}$$

What would happen if infection were reintroduced to this population? In this case, we know that $S \approx (1 - a_r)N$, so that $dI/dt \approx [\beta(1 - a_r) - \gamma]I$. An epidemic can only begin if $a_r < 1 - 1/R_0$. In fact, this is never the case following an epidemic in a completely susceptible population because the attack rate is always greater than $(1 - 1/R_0)$.

If we define the proportion of the population that is susceptible at time t as $x(t) = S(t)/N$, and the effective reproduction number as $R_t = R_0x$, then a threshold in a population with some immune people occurs at $R_t = 1$, with epidemics occurring when $R_t > 1$ but not for $R_t \leq 1$. What is needed to sustain long-term infectious dynamics is a source of new susceptible individuals. This occurs in real populations via births, or through non-immune migrants.

2.3.1 The Endemic *SIR* Model

Consider an idealized population, with a life-expectancy of L years (generated by a constant risk of death $\mu = 1/L$) and a birth rate of B in the whole population. When children are born, they have maternal antibodies to a number of infectious diseases in their body (aided by breast milk), which provide some protection against infection during their first few months of life. This is relevant to immunization, but for simplicity we will ignore this and assume that newborns are born susceptible if unvaccinated and immune if vaccinated. On the basis of these assumptions, we can write down the endemic *SIR* model incorporating vaccination:

$$\begin{cases} \dfrac{dS}{dt}=-\dfrac{\beta IS}{N}+B(1-p)-\mu_d S \\ \dfrac{dI}{dt}=\dfrac{\beta IS}{N}-(\gamma+\mu_d)I \\ \dfrac{dR}{dt}=\gamma I+Bp-\mu_d R \end{cases} \tag{12-5}$$

Here, p is the proportion of infants who are successfully vaccinated. Assuming that the population remains constant in size ($B = \mu N$), the model leads to repeated epidemics of decreasing size such as shown in Figure 12-2. Vaccination results in a decreased overall incidence, an increase in the inter-epidemic period and a decrease in the amplitude of the oscillation of incidence, with more marked effects at higher vaccine coverage levels.

2.3.2 Immunization

Equation 12-5 describes the infection dynamics when immunization is provided at birth. This is easy to incorporate into the endemic *SIR* model – instead of directing all newborns into the susceptible compartment, a proportion p are directed into the immune class instead. This proportion p is actually the product of two factors: the proportion of infants who are vaccinated (v) and the efficacy of the vaccine (VE).

2.3.3 Waning of Immunity

The duration of immunity also varies between infections, and may differ for vaccine-derived immunity as opposed to natural infection. For diseases such as measles, natural infection seems to produce lifelong immunity. For varicella (chicken pox), second attacks are observed with some regularity suggesting that not all individuals experience lifelong immunity. The immune response after vaccination can differ from that due to natural infection and in some cases is of limited duration.

This effect can easily be incorporated into the endemic *SIR* model: if we assume that immunity due to vaccination is equivalent to that due to natural infection, then the model takes the form:

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta IS}{N} + B(1-p) - \mu_d S + \omega R \\ \dfrac{dI}{dt} = \dfrac{\beta IS}{N} - (\gamma + \mu_d) I \\ \dfrac{dR}{dt} = \gamma I - \mu_d R + Bp - \omega R \end{cases} \tag{12-6}$$

2.3.4 Equilibrium

After a period of time, the model states tend toward constant values. If $R_0 < 1$, there will be no epidemic, and the infection will die out. If $R_0 > 1$, a series of epidemics with shrinking peaks will occur as the dynamics lead toward equilibrium levels of susceptible, infectious and immune individuals. Introducing a vaccination program changes the dynamics in several ways: it results in a reduced overall incidence, a longer inter-epidemic period and an increase in the amplitude of the oscillation of incidence, with more marked effects as vaccine coverage rises.

The equilibrium values can be calculated by setting the right hand side of Equation 12-6 equal to zero. The prevalence of infection is either zero, or it satisfies the equation:

$$I = \frac{B\left(1 - p + \dfrac{p\omega}{\omega + \mu_d}\right)}{\gamma + \mu_d - \dfrac{\omega\gamma}{\omega + \gamma}}\left(1 - \frac{1}{R_0\left(1 - p + \dfrac{p\omega}{\omega + \mu_d}\right)}\right) \tag{12-7}$$

The significance of this expression can be illuminated through some simple examples.

1. No vaccination ($p = 0$) and lifelong immunity ($\omega = 0$): in this case, we find that I is proportional to $1 - 1/R_0$, so that whenever $R_0 > 1$ infection will persist, while when $R_0 < 1$ the equilibrium prevalence of infection is zero.
2. Vaccination with lifelong immunity: in this case, I is proportional to $1 - 1/((1 - p)R_0)$. Thus if p is large enough so that $(1 - p)R_0 < 1$, then the vaccination program should eliminate the infection.
3. 100% coverage with a 100% effective vaccine ($p = 1$) with waning rate equal to the death rate: in this case, I is proportional to $1 - 2/R_0$. Thus, even when the average duration of immunity is equal to the average life-expectancy, infection can only be eliminated if $R_0 < 2$. Thus, the duration of immunity can severely limit the effectiveness of vaccination programs.

Figure 12-3 illustrates the effect of vaccination ($VE = 1$, no waning) on incidence. Note the reduced average incidence, and the longer time between epidemic peaks as coverage rises.

Figure 12-4 shows the effect of vaccination on the reproduction number over time (R_t). Near equilibrium, $R_t \approx 1$ but the start of a vaccination program upsets this equilibrium, pushing R_t below 1. However, after a period of time, susceptibles build up enough to start a new epidemic if vaccination coverage is only 50%. This cycle repeats itself as the system moves towards a new equilibrium value. However, if coverage is as high as 70%, this is sufficient to push R_t below 1 permanently, leading to elimination of infection. In this case, the parameters $\beta = 0.5$, $D = 6$ lead to an R_0 of 3, which means that if more than 67% of the population are immune, the disease will be eliminated.

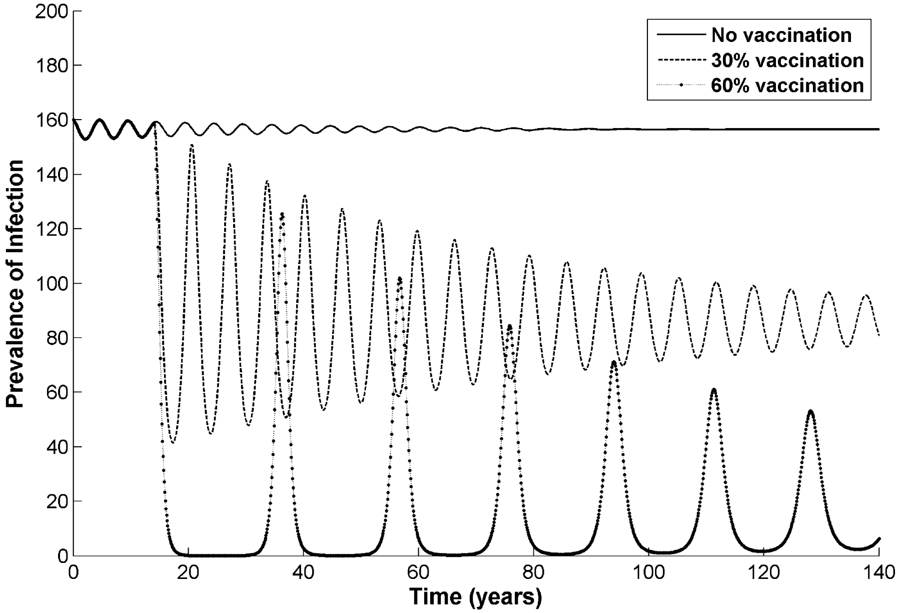


Figure 12-3. Effect of increasing vaccination coverage starting from a nearly constant prevalence of infection, as calculated using the endemic *SIR* model with parameters $\beta = 0.5$, $D = 6$, $N = 1$ million.

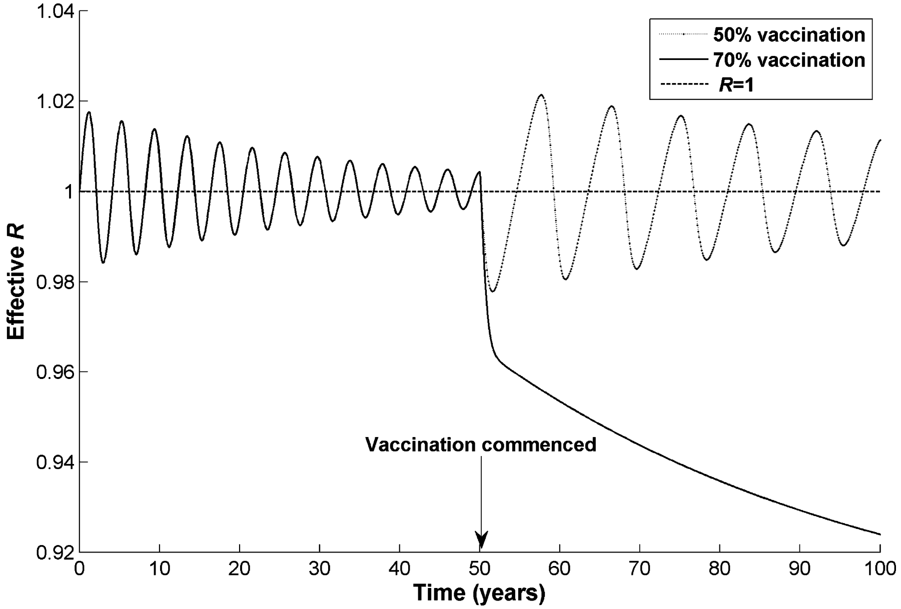


Figure 12-4. Effect of vaccination on the effective reproduction number (R_t). Note that for the parameters $\beta = 0.5$, $D = 6$, $N = 1$ million, 70% vaccination coverage is enough to reduce R_t below 1 permanently.

2.4 More Realistic Models

The above models are valuable because they reproduce observed aspects of infectious diseases, including epidemic cycles and changes in epidemiology as a result of vaccination, yet rely on only a few parameters. However, they provide only a rather crude representation of reality and often more complicated models are necessary. Here we detail a few common refinements.

2.4.1 Age-Related Risks

In the above model the population implicitly has an age-distribution that declines exponentially with age. This is a reasonable model for some developing countries but does not describe the demography of developed countries, which have an almost constant age-distribution with the exception of the very old. While arbitrary age-distributions can be accommodated in refinements of the endemic model, a good approximation to the dynamics is gained by using a constant age-distribution with a maximum age equal to the life-expectancy in the population (Anderson and May, 1999).

The risk of infection can also vary with age. It is self-evident that personal contacts of individuals tend to be clustered in certain age groups (typically near the same age as the individual). This suggests that the force of infection should also vary by age. This is usually incorporated into models by using a Who-Acquires-Infection-From-Whom (WAIFW) matrix, which can be partially estimated from incidence and immunity data (Anderson and May, 1999) and augmented with data from studies specifically tracking age-related contacts (Mossong et al., 2008).

2.4.2 Vaccine Efficacy

Immunization programs also change age-related risks. In the above model, vaccines are provided to a proportion of each birth cohort as soon as they are born. However, in practice, maternally-derived immunity of infants can require delaying vaccination by up to 12 months, and multiple doses of vaccine, spaced at intervals of a month are often required. In addition, children often receive vaccine doses later than the time recommended in the schedule, and the immune response may be delayed by up to 1 month following vaccination. These kinds of features can be incorporated by allowing explicit age-dependency in the model equations (they become partial differential equations).

Modeling immunization by routine vaccination of birth cohorts is also only applicable in some cases. Vaccination programs for influenza tend to target a proportion of older individuals each year, while "catch-up" vaccination of a number of birth cohorts at once is common in developing countries and is often part of new immunization programs in developed countries.

2.4.3 Stochasticity

The models described above are deterministic and do not include random variation. This is justified on the basis of applying them to large populations, in which prevalence of the infection is considerable. When the population is small, or prevalence is very low, chance events can greatly influence the dynamics. This element of chance can be incorporated by assigning probabilities instead of rates to transitions between states. Stochastic methods are commonly used in individual-based models, in which characteristics of individuals are seen as an important component of dynamics. A detailed introduction to stochastic methods in infectious disease modeling can be found in (Daley and Gani, 2001).

2.5 Special Issues for Vaccination

Modeling of vaccination programs must account for the complexities of infectious diseases and the impact of vaccination programs. For example, vaccination programs change disease epidemiology, usually resulting in a right-shift of age-specific incidence. Universal vaccination programs can also confer herd immunity, the strength of which depends on the coverage and efficacy of the vaccine as well as the R_0 for the pathogen.

For some pathogens, such as influenza, cross-protection against related strains must be taken into consideration, as must the potential for strain replacement with non-vaccine strains. This may also be an issue for pneumococcal vaccination programs, where there is evidence of changing colonisation patterns in highly-vaccinated populations (Veenhoven et al., 2004). Vaccination can also have secondary benefits in cases such as super-infection, where, for example, invasive pneumococcal disease often follows influenza infection (Brundage, 2006). Preventing influenza infection is likely to also prevent the invasive disease, so that the benefit of vaccination is enhanced.

Finally, vaccination is unique because it affords primary prevention to healthy people, with the possibility of adverse events or side effects. For many individuals, the benefits of vaccination may not occur or be delayed into the distant future. On the other hand herd immunity can provide protection even to those who choose not to become vaccinated. This provides the prospect of eliminating or eradicating infection when a vaccine is effective enough to make this possible. However, since vaccination programs across the world vary greatly, elimination may need to be sustained in individual countries or regions for many years before eradication is achieved. This means that even though the current risk of infection in the elimination phase is very small, vaccination programs for diseases such as polio must be maintained until eradication is achieved in order to avoid a return to endemic infection.

2.5.1 Data Requirements and Surveillance

Surveillance systems are a core contributor to modeling data and comprise the ongoing acquisition of information for use in public health action. Surveillance should be practical, timely and uniform, and the data are not necessarily complete. Surveillance systems can be passive, active or sentinel in nature. Of these, passive surveillance is the most commonly used. It is initiated by the data provider, is cheaper and easier to establish than active surveillance, but often underestimates disease burden. In active surveillance, the investigator actively solicits reports from providers. This type of surveillance is used commonly in outbreaks, is usually more complete than passive surveillance, but can be resource-intensive and expensive. A compromise between active and passive systems can be achieved with sentinel surveillance, which utilizes a sample of sites, can provide relatively detailed information, is cheaper than active surveillance, but may be less complete. Other data that contribute to mathematical models include seroepidemiologic data from population-based serosurveys, enhanced surveillance data, vaccine coverage data and vaccine efficacy estimates from clinical trials.

In terms of pandemics and emerging infections, early detection systems such as signalling of elevated emergency department presentations of atypical pneumonia to trigger active containment are important. These should aim to have the capacity to estimate both the effective reproduction number R_t, and to be sensitive enough to detect the impact of interventions such as isolation and quarantine, pharmaceutical treatment and prophylaxis, genetic changes in the pathogen such as development of resistance to treatments and vaccine effectiveness.

3. MATHEMATICAL MODELS FOR BIOSURVEILLANCE OF VACCINE-PREVENTABLE DISEASE

The field of biosurveillance has advanced rapidly over the last decade, making use of varied forms of data from diverse sources. The most noticeable advances in biosurveillance practice have occurred among real-time syndromic surveillance systems which aim to allow the early detection of outbreaks of both known and unknown pathogens. Although diagnostic data sources are by definition slower to register an event than syndromic (or pre-diagnostic) sources, the surveillance of diagnostic data provides valuable specific information for monitoring population health outcomes and informing health policy.

While routine analyses of real-time biosurveillance systems have been developed to facilitate early detection of bioterrorist or emerging disease threats, analysis of vaccine-preventable disease data is generally retrospective in nature, and not performed in real-time. In addition to monitoring laboratory confirmations of notifiable infectious diseases, biosurveillance for immunization programs is largely aimed at capturing data that informs influential assumptions in transmission models, such as age-specific immunity. Information on immunity is derived from large-scale cross-sectional serosurveys that are resource-intensive and are generally performed at intervals of several years. More routine surveillance of population immunity data would be particularly beneficial for diseases where vaccine-derived immunity is lost over time and would substantially improve the accuracy of related models. The routine application of modeling approaches to vaccine-preventable disease data, including both diagnostic and immunization data, offers the means to ensure that the impact of population health interventions are closely monitored.

Biosurveillance methods conventionally applied to both syndromic and diagnostic disease data, including aberration and cluster detection, have the potential to improve the analysis of immunity and morbidity data for vaccine-preventable diseases, and are currently underutilised in this field. Approaches to spatio-temporal analysis such as scan statistics could support assessment of the need for targeted interventions to decrease susceptibility or control disease transmission, and compliment more detailed modeling approaches. Cluster detection methods provide a different means of identifying local indicators of increased disease risk that could be used to inform targeted interventions to improve disease control.

More structured modeling approaches can also be incorporated in biosurveillance to evaluate disease risk and disease transmission within populations. The routine use of simple modeling approaches can greatly improve our understanding of the impact of disease prevention and control programs. Emerging disease threats such as pandemic influenza highlight the need for modeling methods to provide early, accurate and timely epidemiological information to inform and evaluate disease control efforts. Traditional deterministic models of infectious diseases, such as those outlined above, are useful for describing transmission of established diseases in large populations. However, the need to identify the key characteristics of transmission from small numbers of cases has stimulated development of models that are based on surveillance and contact tracing data and reflect the closer integration of surveillance, modeling and disease control interventions (Matthews and Woolhouse, 2005).

Stochastic effects have been shown to be important in the early stages of epidemics. Studies of disease transmission traditionally divide populations into subgroups as a means to include heterogeneity. However, these methods are unable to capture the role of individual variation in outbreak dynamics, and estimates of reproductive numbers at the population level can mask significant individual-level variation in transmission. Individual variation in infectiousness can have important implications for evaluating transmission and predicting the epidemic course in emerging disease outbreaks (Lloyd-Smith et al., 2005).

The increasing availability of detailed outbreak surveillance data has provided more evidence of the stochastic nature of epidemics and improved the ability of modeling methods to produce meaningful epidemiological information from small case numbers (Matthews and Woolhouse, 2005). Rapid collection and modeling of surveillance data can provide valuable information for public health decision-making.

3.1 Models of the Reproductive Number in Immunization Biosurveillance

Recently, statistical techniques have been developed for estimating and monitoring the effective reproductive number of an outbreak in its early stages. These methods could contribute to the rapid evaluation of disease control measures during an epidemic and facilitate improved disease control. The effective reproduction number is a time-dependent parameter that quantifies the average number of secondary infections caused by a typical infectious case. The reproductive number provides an indicator of the future course of the epidemic and critical information for planning for epidemic control (Wallinga and Lipsitch, 2007).

Real-time estimation of reproductive numbers would have practical value for communicable disease control, enabling epidemiologists to evaluate the impact of the control measures implemented. Several models have been designed to allow estimation of the reproductive number from routine surveillance data, and three prominent approaches are used here to illustrate the variation in practical applications and methods. Together, these approaches allow routine surveillance of the reproduction number under the conditions of disease elimination as well as during epidemics, providing sufficient case-based surveillance data are available.

3.1.1 Surveillance of Disease Elimination

Farrington and co-workers (2003) describe the use of a branching process model for the surveillance of vaccine-preventable diseases which has practical application for monitoring the elimination of diseases that are controlled by mass vaccination. This model uses an approximation to the epidemic process which requires only surveillance data on the occurrence of cases and epidemiological information that permits the linkage of cases to specific outbreaks. This approach contrasts with earlier modeling methods which, although being more accurate, require information on the number of susceptible individuals which may not be available in a surveillance setting. A high level of case ascertainment is required, which limits the applicability of the model in some settings and for diseases which are not well-controlled. A stochastic random variable is used to model the number of secondary cases produced by each case in each generation of the epidemic. External causes of stochasticity and the importation and exportation of cases are ignored. The model assumes a homogenous pattern of disease spread, and has been used to estimate the reproduction number for measles in the United States of America (Farrington et al., 2003) and Australia (Becker et al., 2005).

3.1.2 Surveillance of Epidemics

In an epidemic context where transmission rates are high, and where partial tracing information is available, Cauchemez et al. (2006) propose a generic method that enables the real-time estimation of changes in the reproductive number of an outbreak. The method requires surveillance count data, data on the onset of symptoms, and contact tracing data for a subset of cases. As tracing of transmission of disease between all individuals in an outbreak is time consuming and generally infeasible, models estimate the epidemic parameters based on the assumption that the traced cases represent a random sample of all cases.

Inferences about the temporal pattern of the reproduction number are based on the output of a Bayesian hierarchical model which uses a Markov Chain Monte Carlo (MCMC) algorithm to identify the generation interval based on the case tracing data. The output from the MCMC is used to allocate untraced cases to an observed primary case using a Monte Carlo algorithm, and correction for censorship is performed. The method does not require knowledge of the generation interval and is suitable for use with daily real-time surveillance data. The model is associated with six main assumptions about the epidemic and data, including that all cases are detected, and that secondary cases are always reported after their index case. (Cauchemez et al., 2006) This method allows real-time evaluation of the efficacy of outbreak control measures based on variation in disease transmission.

3.1.3 Parameter-Free Epidemic Surveillance

Unlike the preceding methods, Haydon et al. (2003) developed a straightforward parameter-free method to estimate epidemic parameters based on reconstructing the epidemic tree using available contact tracing data. A flexible set of assumptions allow allocation of untraced cases to the epidemic tree based on selection from cases that were known to be infectious at the time of infection using both spatial and temporal data. Reconstructing the outbreak allows the reproduction number to be estimated without having to fit equations to the data and with fewer assumptions than required for deterministic modeling approaches. This approach allows the estimates to be obtained more directly from the data, assuming that the methods of reconstructing the disease transmission pathways are not biased. Reconstruction of the outbreak also allows the model to encompass the spatial and temporal variability inherent in the data, estimation of the variability in the reproduction number over space and time, the exploration of the effect of alternative control measures, and the effect of long-range transmission events (Haydon et al., 2003).

3.2 Summary

In summary, modeling approaches for the analysis of infectious disease surveillance data have been developed which are suitable for use in both the environments of disease elimination and epidemics. The routine use of these types of models in surveillance systems will allow further evaluation of the benefits of mathematical modeling in the applied surveillance context.

Practical application of the modeling methods reviewed in the routine surveillance of vaccine-preventable diseases is dependent on the availability of sufficiently detailed case-based surveillance data and high case detection rates. These data are not always routinely available, even in developed countries. Advancements in disease control leading to a decrease in the burden of disease may improve the ability of health jurisdictions to collect the required data (Leung et al., 2008).

The parameter-free method described by Haydon et al. (2003) provides a straightforward means to analyse spatial data associated with epidemics, and the usefulness of this approach in practice requires further investigation. Spatio-temporal analysis of disease transmission provides a method which could be integrated with other information sources such as population and environmental data, to improve models of disease spread.

ACKNOWLEDGEMENTS

James Wood and Rochelle Watkins are supported by a National Health and Medical Research Council Capacity Building Grant in Population Health #358424.

QUESTIONS FOR DISCUSSION

1. Eradication of smallpox is the greatest success of immunization to date. Currently the WHO is conducting immunization programs aimed at eradicating polio and measles. The R_0 for measles has been estimated at 15 or higher, whereas for smallpox it was between 3 and 6. Discuss how the value of R_0 can influence the success of vaccination programs. (Hint: read point 2) under the section Equilibrium.)
2. By reducing the prevalence of infection, population-based immunization programs reduce exposure to infection. This tends to raise the average age at which people become infected. Can you think of situations in which this might have negative consequences? (Hint: infections with Rubella.)
3. When an epidemic occurs, one of the easiest things to record is the timing of cases. However, it is not enough to tell us the R_0 of the pathogen. What other data would need to be collected to help us determine the R_0? (Hint: think about the timing involved in infection.)
4. As an epidemiologist in charge of the overall public health management of a recent outbreak of a vaccine-preventable disease in the local community, how could the results of a mathematical model of the early progress of the outbreak facilitate your planning for outbreak control?

REFERENCES

Anderson, R. M. and M. M. May (1999). *Infectious Diseases of Humans, Dynamics and Control*. London, Oxford University Press.

Ashmore, J., S. Addiman, et al. (2007). "Measles in North East and North Central London, England: a situation report." *Euro Surveill* **12**(9): E070920.2.

Bailey, N. (1975). *The Mathematical Theory of Infectious Diseases*. London, Charles Griffin and Company.

Becker, N. G., Z. Li, et al. (2005). "Monitoring measles elimination in Victoria." *Aust N Z J Public Health* **29**(1): 58–63.

Brundage, J. F. (2006). "Interactions between influenza and bacterial respiratory pathogens: implications for pandemic preparedness." *Lancet Infect Dis* **6**(5): 303–12.

Cauchemez, S., P. Y. Boelle, et al. (2006). "Estimating in real time the efficacy of measures to control emerging communicable diseases." *Am J Epidemiol* **164**(6): 591–7.

Centers for Disease Control and Prevention (1996). "Update: diphtheria epidemic – New Independent States of the Former Soviet Union, January 1995–March 1996." *MMWR Morb Mortal Wkly Rep* **45**(32): 693–7.

Daley, D. and J. Gani (2001). *Epidemic Modelling: An Introduction*. Cambridge, UK, Cambridge University Press.

Diekmann, O. and J. Heesterbeek (2000). *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. New York, John Wiley & Sons.

Dobson, R. (2008). "England and Wales are among European countries at highest risk of measles epidemic." *BMJ* **336**(7635): 66.

Farrington, C. P., M. N. Kanaan, et al. (2003). "Branching process models for surveillance of infectious diseases controlled by mass vaccination." *Biostatistics* **4**(2): 279–95.

Galil, K., R. Singleton, et al. (1999). "Reemergence of invasive Haemophilus influenzae type b disease in a well-vaccinated population in remote Alaska." *J Infect Dis* **179**(1): 101–6.

Gay, N. J., L. M. Hesketh, et al. (1995). "Interpretation of serological surveillance data for measles using mathematical models: implications for vaccine strategy." *Epidemiol Infect* **115**(1): 139–56.

Halloran, M. E., I. M. Longini, Jr., et al. (1999). "Design and interpretation of vaccine field studies." *Epidemiol Rev* **21**(1): 73–88.

Haydon, D. T., M. Chase-Topping, et al. (2003). "The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak." *Proc Biol Sci* **270**(1511): 121–7.

Heesterbeek, J. A. (2002). "A brief history of R_0 and a recipe for its calculation." *Acta Biotheor* **50**(3): 189–204.

Hinman, A. (1999). "Eradication of vaccine-preventable diseases." *Annu Rev Public Health* **20**: 211–29.

Institute of Medicine (2000). *Vaccines for the 21st Century: Tools for Decision Making*. Washington, DC, National Academy Press.

Jacobsen, K. H. and J. S. Koopman (2005). "The effects of socioeconomic development on worldwide hepatitis A virus seroprevalence patterns." *Int J Epidemiol* **34**(3): 600–9.

Keeling, M. and P. Rohani (2007). *Modeling Infectious Diseases in Humans and Animals*. New Jersey, Princeton University Press.

Leung, J., A. Rue, et al. (2008). "Varicella outbreak reporting, response, management, and national surveillance." *J Infect Dis* **197 Suppl 2**: S108–13.

Lloyd-Smith, J. O., S. J. Schreiber, et al. (2005). "Superspreading and the effect of individual variation on disease emergence." *Nature* **438**(7066): 355–9.

Matthews, L. and M. Woolhouse (2005). "New approaches to quantifying the spread of infection." *Nat Rev Microbiol* **3**(7): 529–36.

Miller, E., J. E. Vurdien, et al. (1992). "The epidemiology of pertussis in England and Wales." *Commun Dis Rep CDR Rev* **2**(13): R152–4.

Mossong, J., N. Hens, et al. (2008). "Social contacts and mixing patterns relevant to the spread of infectious diseases." *PLoS Med* **5**(3): e74.

Oostvogel, P. M., J. K. van Wijngaarden, et al. (1994). "Poliomyelitis outbreak in an unvaccinated community in The Netherlands, 1992–93." *Lancet* **344**(8923): 665–70.

Rijkers, G. T., P. E. Vermeer-de Bondt, et al. (2003). "Return of Haemophilus influenzae type b infections." *Lancet* **361**(9368): 1563–4.

Roberts, M. G. and M. I. Tobias (2000). "Predicting and preventing measles epidemics in New Zealand: application of a mathematical model." *Epidemiol Infect* **124**(2): 279–87.

Trotter, C. L., M. E. Ramsay, et al. (2003). "Rising incidence of Haemophilus influenzae type b disease in England and Wales indicates a need for a second catch-up vaccination campaign." *Commun Dis Public Health* **6**(1): 55–8.

Veenhoven, R. H., D. Bogaert, et al. (2004). "Nasopharyngeal pneumococcal carriage after combined pneumococcal conjugate and polysaccharide vaccination in children with a history of recurrent acute otitis media." *Clin Infect Dis* **39**(7): 911–9.

Wallinga, J. and M. Lipsitch (2007). "How generation intervals shape the relationship between growth rates and reproductive numbers." *Proc Biol Sci* **274**(1609): 599–604.

SUGGESTED READING

Anderson, R. M. and M. M. May (1999). *Infectious Diseases of Humans, Dynamics and Control*. London, Oxford University Press.

Heesterbeek, J. A. (2002). "A brief history of R_0 and a recipe for its calculation." *Acta Biotheor* **50**(3): 189–204.

Hinman, A. (1999). "Eradication of vaccine-preventable diseases." *Annu Rev Public Health* **20**: 211–29.

Matthews, L. and M. Woolhouse (2005). "New approaches to quantifying the spread of infection." *Nat Rev Microbiol* **3**(7): 529–36.

Trotter, C. L., M. E. Ramsay, et al. (2003). "Rising incidence of Haemophilus influenzae type b disease in England and Wales indicates a need for a second catch-up vaccination campaign." *Commun Dis Public Health* **6**(1): 55–8.

Chapter 13

NATURAL LANGUAGE PROCESSING FOR BIOSURVEILLANCE

Detection and Characterization from Textual Clinical Reports

WENDY W. CHAPMAN[1,*], ADI V. GUNDLAPALLI[2],
BRETT R. SOUTH[2], and JOHN N. DOWLING[1]

CHAPTER OVERVIEW

Information described in electronic clinical reports can be useful for both detection and characterization of outbreaks. However, the information is in unstructured, free-text format and is not available to computerized applications. Natural language processing methods structure free-text information by classifying, extracting, and encoding details from the text. We provide a brief description of the types of natural language processing techniques that have been applied to the domain of outbreak detection and characterization. We group textual data generated by a healthcare visit into four classes: chief complaints, emergency care notes, hospitalization notes, and discharge reports. For each class of data, we illustrate uses of the data for outbreak detection and characterization with examples from real applications that extract information from text. We conclude that a modest but solid foundation has been laid for natural language processing of clinical text for the purpose of biosurveillance, with the main focus being on chief complaints. To provide more accurate detection and to assist in investigating and characterizing outbreaks that have already been detected, future research should focus on tools for extracting detailed clinical and epidemiological variables from clinical

[1*] *Department of Biomedical Informatics, University of Pittsburgh, 200 Meyran Avenue, Pittsburgh, PA 15260, USA, wendy.w.chapman@gmail.com*

[2] *Division of Epidemiology, School of Medicine, University of Utah, 30 North 1900 East, AC230A, Salt Lake City, UT 84132, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_13,

reports. Areas of challenge include identifying contextual modifiers of clinical conditions, identifying useful temporal information, and integrating information across reports for a more complete view of a patient's clinical state.

Keywords: Natural language processing; Biosurveillance; Syndromic surveillance; Text processing; Information extraction; Infectious disease

1. INTRODUCTION

> It would be so nice if something made sense for a change.
> Alice in *Alice in Wonderland* by Lewis Carroll

This quote from another century sums up the challenges faced by those working in text processing of medical data. Although the "sense" – context and details of a medical note describing a patient's visit and illness – are readily apparent to a trained human reader, training computers to understand the same information is a daunting task.

Consider a situation in which a novel strain of avian influenza (H5N1) with pandemic potential is causing outbreaks of disease among domestic and commercial poultry in many parts of the world with sporadic transmission to humans. Public health agencies and medical personnel everywhere including the U.S. are concerned about the first imported case of avian influenza that would most likely go undetected. Identifying subsequent cases too would pose a challenge if by coincidence the index case occurred during the annual winter respiratory season.

Whereas structured data such as chief complaints and ICD-9 diagnoses are often the first pointers to a pool of patients that fit a particular case definition, there are epidemiologic and clinical clues in the patient's encounter that are accessible only by reviewing the entire clinical record. These clues include a history of recent travel to a disease-affected region, contact with others with similar symptoms, and unusual severity of disease in otherwise healthy individuals. A manual reading of records, such as provider notes and microbiology, radiology, and pathology reports, offers the most reliable, though resource-intensive, method of extracting relevant information. As an alternative, mining the electronic records using natural language processing (NLP) offers information extraction and discovery in a reproducible, resource-sparing, and efficient manner.

In this chapter, we provide an overview of the data sources used for biosurveillance (Sect. 2), focusing on clinical data sources that are stored in textual format. We briefly describe the natural language processing techniques currently in use to process free-text clinical reports into a form that can be used by surveillance systems and illustrate uses of textual clinical data for outbreak detection and characterization with examples from real applications (Sect. 3).

2. OVERVIEW OF DATA SOURCES FOR BIOSURVEILLANCE

The earlier the detection, the more time stakeholders have to mitigate an infectious disease outbreak. To this end, a variety of data sources have been utilized for predicting an outbreak before detailed and confirmatory medical information is available. Several surveillance systems have utilized population- and individual-level data of human behavior prior to seeking medical attention. Others have employed traditional healthcare data to detect an outbreak when patients seek medical care in ambulatory, emergency, and inpatient settings. Yet others have used various combinations of data sources to further refine biosurveillance and increase specificity. In all these systems, the data used for biosurveillance can be broadly classified as either structured data – such as ICD-9 codes – or unstructured data – such as free text.

Structured data are relatively easy to parse, process, manipulate, and analyze mainly by matching codes and terms. Though the overall benefit or utility of different data sources for biosurveillance is not fully established, it is generally accepted that structured data provide greater sensitivity over specificity, thus identifying events that warrant further investigation by public health agencies and other stakeholders. Unstructured free text in the patient's medical record allows access to information and details that are generally not available in structured data sources. Mining and encoding this text may allow for increased specificity in detecting an infectious disease outbreak and identifying epidemiologic details.

2.1 Surveillance from Non-clinical Data Sources

Although the electronic medical record (EMR) contains a wealth of data potentially useful for biosurveillance, many non-clinical data sources may provide earlier evidence of an outbreak or may provide information complementary to that contained in an EMR. Below, we provide a short overview of structured and unstructured data sources outside of the EMR that are being used for surveillance.

2.1.1 Structured Non-clinical Data

Structured data sources have been used extensively for biosurveillance, mainly because of their availability and relative ease of accessing and processing. It is also postulated that these data sources may reveal patterns of behavior among the population *before* they seek formal medical attention, thus providing a very early indicator of an infectious disease outbreak. Some sources are self-evident with regard to the population seeking symptomatic relief of, or expressing concern for, symptoms, including over-the-counter medication and health-related sales of items used for common syndromes, such as respiratory, febrile and gastrointestinal illnesses, calls to local 911 centers, poison control center calls (for gastrointestinal illnesses), health-related web sites access data, medical provider advice hotline data or symptoms reported to a telephone health advice service, healthcare provider database searches, and Internet-based illness reporting. Others have attempted to link changes in patterns of everyday life to disease outbreaks, and these include airline travel volume (influenza spread and mortality), school/work absentee volume, and commuter road, mass transit and entertainment venue usage. Other data sources that may be important for predicting outbreaks, including zoonotic diseases, take account of weather data, disease vector data and illnesses and death in animals.

2.1.2 Textual Non-clinical Data

The importance of unstructured or textual data, especially Internet-based information for monitoring emerging infectious diseases and reporting public health outbreaks, is growing. A number of public and private global surveillance systems aggregate and post unstructured data concerning public health from a variety of Internet-based sources in an attempt to create an automated real-time Internet surveillance for epidemic intelligence. For example, ProMed-mail (the Program for Monitoring Emerging Diseases) is an Internet-based reporting system that rapidly disseminates information on outbreaks of infectious diseases and acute exposures to toxins that affect human health. Sources of moderated information include local and national media reports, official reports, online summaries, local observers and others.

In contrast, HealthMap [1] and Global Health Monitor [2] automatically collect news from the Internet about human and animal health and plot the data on a World map, using an established dictionary to map textual data and extract geographical location and mention of the disease [3]. Global Health Monitor is an online system for detecting and mapping infectious disease outbreaks that appear in news stories. Global Health Monitor takes an ontology-centered approach to knowledge understanding and linkage to external resources. Indicators and warnings gathered from public domain reports of infectious disease outbreaks have been mined using keyword and text searching to create a heuristic staging model for the detection and assessment of outbreaks for both front-line personnel and decision-makers [4]. The Public Health Agency of Canada sponsors a similar early warning system that mines available free-text Internet reports in seven languages for disease outbreaks and provides the feed for subscribers as the Global Public Health Intelligence Network [5].

2.2 Surveillance from Clinical Data Sources

Once a patient presents to a healthcare facility, structured and unstructured data are generated and stored in the EMR, and that data may be repurposed from the goal of documentation to assist surveillance systems in detecting and characterizing outbreaks. Below, we summarize some structured and unstructured data sources in the EMR.

2.2.1 Structured Clinical Data

As an individual seeks medical attention, structured data are generated as part of the encounter. In general, these vary in terms of availability in electronic format, accessibility, and timeliness. Early indicators in this context are referred to as pre-diagnostic indicators and include such data as ambulance call chief complaints, emergency department or ambulatory clinic total patient volume or structured chief complaint data, volume of unscheduled hospital or intensive care unit admissions, influenza-like illness (ILI) monitoring by sentinel physicians, laboratory order data, test ordering patterns (e.g., influenza, lumbar puncture), and radiology imaging ordering volume. As the patient advances through the healthcare system, the data sources available include outcomes of emergency department and ambulatory clinic visits by diagnosis, usually in the form of ICD-9 codes, laboratory test results, including microbiology, pharmacy prescriptions filled, acute diagnoses in nursing home populations, unexplained deaths and medical examiner case volume, and insurance claims or billing data, including Medicare or Medicaid claim forms filed.

2.2.2 Textual Clinical Data

Unstructured data in the form of free text comprises most of the medical record and is manifest mainly as notes generated by healthcare providers and laboratory and imaging reports. Although more difficult to access, the majority of the data regarding the health status of the patient, including details of laboratory and radiology exams, along with past medical, family, social, and travel history, are only found in these free-text records. For biosurveillance purposes, highly relevant epidemiologic clues to infectious diseases, such as exposure to someone with a similar illness, travel to an endemic region, or specific symptoms, are available only in free-text medical records. Thus, extracting patient-level data from unstructured free-text data is an important first step in understanding the clinical state of a patient. Identifying features from text in turn should translate to improved identification of patients of interest and ultimately to improved care at both the patient and population level.

Most commonly available textual clinical data include emergency department (ED) triage chief complaint data in free text, clinical narrative reports (often in the EMR), ED patient summaries, radiology imaging dictated reports, and admission, progress, procedure, operative, and discharge notes.

In this chapter, we focus on these data sources – clinical textual data generated by a patient's visit to an inpatient or outpatient healthcare facility. In the next section, we describe the different types of text and natural language processing methodologies currently in use to analyze these data sources and provide an overview of the clinical textual data sources being used for biosurveillance.

3. SURVEILLANCE FROM TEXTUAL CLINICAL DATA SOURCES

3.1 Methodologies for Processing Clinical Textual Data

Natural language processing applications strive to classify, extract, and encode information from text. Several strategies exist for automated processing of electronic clinical text. Applications developed for biosurveillance can be described using three main categories: (a) keyword-based, (b) statistical, and (c) symbolic. Many applications use one or more of these techniques to extract and classify information from biomedical texts. Simpler text processing techniques may not be appropriately labeled as natural language processing methodologies, per se, but for simplicity, we use the term NLP to describe all of the text processing techniques described throughout this chapter.

Below we give a broad description of the three methodologies with examples from the biosurveillance domain. For a more detailed description of the challenges inherent in processing textual clinical documents, see [6].

3.1.1 Keyword-Based NLP Techniques

Keyword-based methods involve text processing techniques that utilize a specific list of terms, keywords, or phrases to identify variables of interest. For instance, a keyword search for identifying fever in a report may include the terms "fever" and "febrile." This list of words or phrases can be expanded by finding synonyms and term variants stored in standard vocabularies such as the UMLS Metathesaurus [7]. Because many of the conditions in clinical reports are described as being absent in the patient [8], term lists are often coupled with a negation algorithm [8–11] to differentiate between conditions the patient has (e.g., "complains of fever") and conditions that are negated (e.g., "denies fever").

The main advantage of keyword-based methods is their simplicity – keyword searches are easy to implement, and they are flexible in their ability to include new keywords as needed. Keyword searches perform quite well in a few circumstances. First, keyword searches designed for a small number of clinical concepts can be successful, because it may be possible to come up with the majority of textual variants used for a few concepts. The example of fever supports this claim – one study showed that a keyword-based algorithm could successfully identify fever in ED reports with a few synonyms for fever, a simple negation algorithm, and a regular expression pattern for the patient's temperature [12]. Second, if the input text is linguistically simple, keyword searches may be sufficient. Many of the chief complaint classifiers described in Sect. 3.3.1 use keyword searches to effectively classify patients into syndrome categories and identify specific clinical conditions from chief complaints. For example, the New York City Syndromic Macros [13] classify patients with chief complaint strings including the words "cough," "coughing," "sob," and "wheezing" as Respiratory patients and can classify respiratory chief complaints with fairly high sensitivity and specificity [14].

There are several disadvantages to keyword-based methods. First, keyword searches cannot recognize words that are not in their keyword list. There are too many synonyms, abbreviations, and textual variants for expressing medical concepts to maintain an accurate keyword list for a large set of clinical concepts. Second, the more a textual document resembles natural language, the more linguistically complex the text becomes and the less effective a keyword search will be. Since keyword-based approaches primarily focus on identifying concepts in isolation they ignore important contextual information crucial for understanding the description of a patient's clinical state, such as uncertainty, temporality, and relationships among concepts. Statistical and symbolic techniques attempt to address the more challenging aspects of processing natural language.

3.1.2 Statistical NLP Techniques

Statistical text classification techniques employ the frequency distribution of the words to automatically classify text into one of a distinct set of predefined categories. Various statistical models have been applied to the problem of text classification, including Bayesian belief networks, decision trees, regression models, nearest neighbor algorithms, neural networks, and support vector machines. The basic component in all statistical techniques is the frequency distribution of words or characters in the text. In the domain of biosurveillance, statistical text classification techniques have been applied to triage chief complaints [15–18] and chest radiograph reports [19]. Several chief complaint classifiers use statistical techniques to classify free-text triage chief complaints into syndrome categories. Some statistical classifiers use the frequency distribution of the words in the chief complaints to predict the classification. For example, “vomiting” would be more likely to be classified as Gastrointestinal than Respiratory or Neurological, because chief complaints in the training set that contained the word “vomiting” were classified most frequently as Gastrointestinal. Other statistical chief complaint classifiers model the frequency distribution of contiguous characters, such as “asth” and “heez,” rather than entire words [20].

Beyond frequency distribution of words, statistical techniques can be used for higher level reasoning. MPLUS uses Bayesian networks to make inferences about the underlying meaning of words. For instance, the phrase “hazy opacity in the left lower lobe” implies the concept *localized infiltrate* which implies the disease *pneumonia* [21]. One chief complaint classifier [22] uses a weighted semantic similarity score grouping method that is capable of automatically assigning previously un-encountered symptoms to appropriate syndrome groups using the UMLS Metathesaurus [7].

An advantage of statistical techniques is the ability to reason under uncertainty. Disadvantages include the need for a labeled training set for learning distributions and relationships in the data and less ability to explain the outcome to users. For some of the same reasons keyword-based methods are ineffective in understanding clinical text, unless a statistical model includes linguistic and domain knowledge, statistical techniques can also fall short in extracting meaning from text.

3.1.3 Symbolic NLP Techniques

Symbolic NLP techniques utilize linguistic information in attempting to interpret free text. NLP methodologies applied to biosurveillance have leveraged knowledge of syntax (the way words are arranged together in a sentence), semantics (the meaning of linguistic expressions), and discourse (relationships between groups of sentences) to interpret text. Syntax can be important in understanding the context surrounding a relevant condition in a document. For example, syntax informs us of the scope of the negation term “denies” in the sentence “Patient denies cough but has experienced shortness of breath” so that we understand *cough* is absent but *shortness of breath* is present. Having a semantic model of a domain can be critical in understanding the meaning represented in a clinical document. Ontologies that explicitly model relationships among concepts in a domain can be helpful in making inferences that humans make, such as the fact that a *stomach cramp* is a type of *abdominal pain* or that a *localized infiltrate* occurs in patients with *pneumonia*. Discourse knowledge is also important in understanding complex narratives like ED reports or discharge summaries, which describe a patient's course over time. Making sense of the narrative involves understanding relationships among concepts in the text, such as temporal relations (e.g., "the vomiting began after she ate at the fast food restaurant") and coreferential relationships (e.g., the fact that "chest pain" and "it" refer to the same condition in "Patient complains of chest pain. It radiates down her left arm."). Most NLP applications that address clinical text more complicated than chief complaints incorporate some type of symbolic techniques in their processing.

3.1.4 Evaluation of Text Processing Methods in Biosurveillance

Three types of evaluations have been applied to NLP applications for biosurveillance: feature detection, case detection, and outbreak detection. In this chapter, we focus on the first two.

Feature detection studies validate the ability of the methodology to classify, extract, or encode features from text. A feature may be among other things a syndrome category, a specific clinical concept, or a concept modifier. Feature detection studies evaluate the technical accuracy of the application based on the input text without considering whether or not the input text represents the true clinical state of the patient. For instance, a feature detection study for chief complaint classification may measure whether the chief complaint classifier assigns the correct category to the chief complaint, using the meaning of the chief complaint as the reference standard. Feature detection studies ignore the possibility that the input text may present an incomplete or even inaccurate representation of the patient's clinical state. Still, feature detection studies perform a useful function in determining whether the NLP application is performing the task it was designed to perform. The majority of evaluations of NLP applications in the biosurveillance domain are feature detection studies that evaluate the ability of the applications to classify chief complaints into syndrome categories or identify pneumonia-related concepts from chest X-ray reports, for example.

Once the ability to detect features from text is verified, the ability of encoded features to diagnose individual cases of interest (case detection) can be addressed. The reference standard for case detection studies is not necessarily the text being processed, because the concern is no longer how well the application can identify information from the textual source but rather how well the identified feature can identify cases of interest. Therefore, the reference standard depends on the finding, syndrome, or disease being diagnosed. A reference standard for the syndrome a patient exhibits when presenting to the ED may include the dictated ED note; a reference standard for a patient's final diagnosis may comprise expert review of the complete medical record.

An example of how feature identification and case detection studies differ is a study on fever detection [12]. The application for identifying fever in chief complaints performed with perfect sensitivity and specificity in the feature identification evaluation. However, in the case detection study that measured how well the automatically extracted variable of fever from chief complaints identified patients who had a true fever based on the reference standard of ED record review, the chief complaint detector performed with sensitivity of only 61%. Despite the fact that the NLP application made no mistakes in determining if fever was described in a chief complaint, the chief complaints did not always mention fever when the patient was febrile in the ED, resulting in imperfect case detection ability.

3.2 Textual Documentation Generated from a Visit to a Healthcare Facility

When a patient comes in contact with the healthcare system, multiple types of data are electronically generated to document the clinical state of the patient and the types of procedures and tests performed on the patient. A patient with acute onset of a disease of concern in biosurveillance often makes first contact with an ambulatory care clinic, which may be an outpatient primary care clinic or emergency room. After an ambulatory visit, a patient is released to return home or may be admitted to the hospital. During the ambulatory and the hospitalization phase of patient care, vast amounts of clinical data are generated and often stored in an electronic medical record system. Figure 13-1 demonstrates the flow of clinical data over time from an ambulatory visit to hospitalization.

3.2.1 Making Use of Textual Documentation for Detection and Characterization

Automated surveillance systems have been successfully developed and implemented in a relatively short period of time in large part because they leverage pre-existing electronic data, including clinical data generated from a healthcare visit. Surveillance systems initially focused their efforts on early detection of outbreaks. More recently, surveillance systems are attempting not only to detect outbreaks but to characterize the nature of the outbreaks. The best data source for surveillance depends on the question being asked of the system. Early detection requires early data that may not be as complete, whereas characterization may call for more detailed data that are not available until later in a patient's course of care.

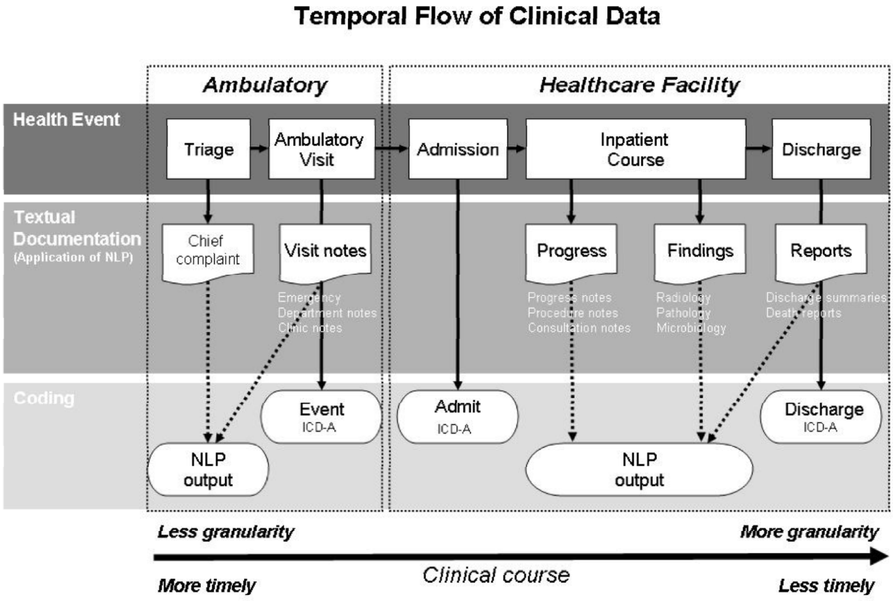


Figure 13-1. Temporal flow of clinical data when patient visits a healthcare facility. ICD-A refers to versions of ICD coding.

Most automated biosurveillance systems currently monitor clinical data generated from the ambulatory care process, which is one of the earliest points in which a patient makes contact with the healthcare system. Systems monitoring ambulatory data from military hospitals have access to encoded admit diagnoses like ICD-9 codes, which are rarely available outside of military clinical care. Systems that monitor ICD admit codes typically group patients into syndrome categories by comparing a patient's admit code to a list of codes that represent a particular constellation of symptoms of interest, such as a list of respiratory or gastrointestinal codes. If the admit code assigned to the patient is in the list of respiratory codes, for example, the patient is classified as a respiratory case. The system monitors the temporal frequency and the spatial distribution of patients with the same classification to look for aberrations that may indicate an outbreak.

After an ambulatory care visit, most hospitals generate a discharge diagnosis for the patient in the form of an ICD code. However, in most hospitals, that code is not available for hours or even days after the patient leaves the urgent care facility. Therefore, many surveillance systems monitor triage free-text chief complaints, which are nearly ubiquitously available from emergency departments and acute care clinics throughout the United States as soon as a patient is admitted to the ambulatory clinic. Like classifying patients from coded admit diagnoses, surveillance systems monitoring chief complaints typically classify patients into syndrome categories based on the chief complaint. However, because chief complaints comprise free-text descriptions, the chief complaints must first be processed to determine what category the text represents. In Sect. 3.3.1.2 we describe the state-of-the-art in chief complaint classification.

Earlier identification of suspicious symptoms or constellations of symptoms promises earlier detection of outbreaks, and chief complaints are available early in the patient care process. However, there is a tradeoff between the timeliness and the completeness of clinical data. Most current surveillance systems monitor admit codes or chief complaints; however, some systems are beginning to monitor more detailed clinical data that are available later in a patient visit, such as ambulatory notes, radiology reports, and discharge summaries. Because electronically available clinical reports are not as timely as chief complaints, they may not be as useful for early detection of small outbreaks. But clinical reports contain a wealth of information regarding a patient's clinical condition that can aid an investigation of a suspected outbreak, along with epidemiological factors useful for characterizing or responding to an outbreak.

The CDC has noted the need for an expansion of surveillance data sources to include laboratory, radiology, and even outpatient records [23] so that we can track more specific case definitions with less baseline noise [24]. Moreover, characterization of an outbreak requires more detailed clinical information to provide health situational awareness during a community outbreak or disaster to track and manage available hospital resources. To do this, we need real-time clinical connections with hospitals and health systems to obtain de-identified but detailed clinical data for patients in acute care settings.

In the section below we describe in detail the different types of clinical data used for surveillance, focusing on current research and applications of natural language processing technologies for detecting cases of concern and for characterization of outbreaks.

3.3 Overview of Clinical Textual Data Sources and Their Application in Biosurveillance

Table 13-1 summarizes much of the research on NLP applications for detecting and characterizing disease outbreaks from textual clinical data sources. For each different data source, such as chief complaints and chest X-ray reports, the table lists whether the evaluation addressed feature identification or case detection, whether the goal was detection or characterization, the target output from the NLP application, such as syndrome classification or coded clinical condition, the reference standard used in the study, and the

Table 13-1. Overview of biosurveillance feature identification and case detection studies of NLP methods.

Data Source	Evaluation Type	Goal of System: Detection or Characterization	Target	Reference Standard	Technique Used
Chief complaints	Feature identification	Detection: syndrome categories	Seven syndromes [1]	Physician classification based on chief complaints	Naïve Bayesian classifier (CoCo)
Chief complaints	Feature identification	Detection: syndrome categories	Seven syndromes [2]	Physician classification based on chief complaints	MPLUS
Chief complaints	Feature identification	Detection: syndrome categories	Respiratory syndrome [3]	ICD-9 discharge diagnoses	Ngram classifier trained on chief complaints whose classifications came from ICD-9 diagnoses
Chief complaints	Feature identification	Detection: syndrome categories	Seven syndromes [4]	Physician and nurse classification based on chief complaints	BioPortal chief complaint classifier
Chief complaints	Feature identification	Detection: syndrome categories	Seven syndromes [5]	Physician classification based on chief complaints	Naïve Bayesian classifier CoCo and NYC DOHMH keyword search with and without preprocessing by CCP and EMT-P
Chief complaints	Feature identification	Detection: syndrome categories	Respiratory syndrome [6]	Physician classification based on chief complaints and triage notes	NCDetect classifier with and without preprocessing by EMT-P
Chief complaints	Feature identification	Detection: syndrome categories	Eight syndromes [7]	Physician classification based on chief complaints	Chinese-English translation followed by classification by BioPortal
Chief complaints	Feature identification	Characterization: clinical conditions comprising syndromic peaks	Clinical conditions associated with respiratory and gastrointestinal syndrome [8]	Physician classification based on chart review	Manual annotation of conditions for cases identified by NYSDOH chief complaint classifier

(*Continued*)

Data Source	Evaluation Type	Goal of System: Detection or Characterization	Target	Reference Standard	Technique Used
Chief complaints	Feature identification	Detection and characterization: clinical conditions	Clinical conditions associated with seven syndromes [9]	Physician classification based on chief complaints	Naïve Bayesian Classifier (SyCo)
Chief complaints	Feature identification	Detection and characterization: clinical conditions	Clinical conditions [10]	No reference standard – case mix prevalence	Coded chief complaints (CCC-EDS)
Chief complaints and emergency department notes	Feature identification and Case detection	Detection: physical finding	Fever [11]	Physician classification based on review of chief complaints and ED notes	Keyword search, NegEx, and identification of hypothetical statements
Emergency department notes	Feature identification	Detection: syndrome category	Acute lower respiratory syndrome [12]	Physician classification based on review of ED notes	Manual annotation of features subsequently classified by machine learning algorithms
Emergency department notes	Feature identification	Detection and characterization: clinical conditions	Clinical conditions related to lower respiratory illness [13]	Physician classification based on review of ED notes	MetaMap
Emergency department reports	Feature identification	Detection and characterization: clinical conditions	Clinical conditions related to lower respiratory illness [14]	Physician classification based on review of ED reports	Topaz
Emergency department triage notes	Feature identification	Detection and characterization: clinical conditions	Negated clinical conditions [15]	Physician review of system output	Keyword search and NegEx
Chest X-ray reports	Feature identification	Detect and characterization: radiological findings and impressions	Mediastinal findings consistent with inhalational anthrax [16]	Physician classification based on review of chest X-ray reports	Identify Patient Sets (IPS) system to create a probabilistic keyword search that uses NegEx for negation identification
Chest X-ray reports	Feature identification	Detection and characterization: radiological findings and impressions	Findings related to acute bacterial pneumonia [17]	Physician classification based on review of chest X-ray reports	SymText, keyword search

Chest X-ray reports	Case detection	Detection: disease cases	Nosocomial pneumonia in infants in the NICU setting [18]	Physician classification based on chart review and NNIS definition	MedLEE
Chest X-ray reports	Case detection	Detection: hospital infections	Bloodstream infections [19]	Physician classification based on chart review	SymText
Chest X-ray reports	Case detection	Detection: radiological findings and impressions	Findings consistent with tuberculosis [20]	Local health department's tuberculosis registry	Electronic medical record and a clinical event monitor with MedLEE
Chest X-ray reports	Feature identification and Case detection	Detection and characterization: Radiological findings and impressions	Findings related to pneumonia [21]	Physician classification based on review of chest X-ray reports	Multi-threaded Clinical Vocabulary Server (MCVS)
Chest X-ray reports	Case detection	Detection and characterization: disease cases	Severity classes for patients with community-acquired pneumonia [22]	Physician classification based on review of chest X-ray reports and discharge summaries	MedLEE
Discharge summaries	Case detection	Detection and characterization: disease cases	Severity classes for patients with community-acquired pneumonia [22]	Physician classification based on review of chest X-ray reports and discharge summaries	MedLEE
Full electronic patient record	Feature identification	Detection and characterization: clinical conditions and epidemiological variables	Conditions related to Influenza-like illness and epidemiologic variables [23]	No reference standard – prevalence comparison against structured patient record	Keyword search and NegEx, MedLEE
Full electronic patient record	Feature identification and Case detection	Detection and characterization: clinical conditions and epidemiological variables	Clinical conditions and epidemiologi c variables [24]	Physician classification based on chart review	Multi-threaded Clnical Vocabulary Server (MCVS)
Full electronic patient record	Case detection	Detection: syndrome cases	Influenza-like illness [25]	Physician classification based on chart review	Keyword search and NegEx

NLP technique that was evaluated. In this section, we describe the data sources people have used in these studies and highlight some of the results from evaluation of NLP methodologies applied to clinical text for biosurveillance.

Because timeliness is critical in detecting outbreaks, automated surveillance systems have focused on the earliest clinical data sources: ICD admit codes and triage chief complaints. Therefore, the bulk of the research on outbreak detection from textual clinical data is aimed at chief complaint classification, which we describe below.

3.3.1 Triage Chief Complaints

Chief complaints are generally the first electronically available clinical description of a patient – a short description of what brings the patient to medical attention. The chief complaint is usually recorded by a healthcare professional, such as a triage nurse or physician, but may also be entered by a clerk. Most chief complaints are recorded as unstructured free text. However, some institutions maintain a pick list from which the user must select a particular chief complaint for the patient. In rare cases, the chief complaint is immediately converted to an ICD code. The chief complaint may be in the patient's own words (or those of a relative or friend). But more often, some interpretation is added by the recorder. For example, if the patient complains of "pain in the chest going into the left arm," the chief complaint may be recorded as "angina." In many cases, only a limited number of characters are allowed to be entered. In an attempt to be brief, the nurses or clerks sometimes utilize creative abbreviations, summarize the complaints in their own words, and make decisions about the most important or relevant complaint in order to reduce the size of the textual entry. Therefore, some complaints conveyed by the patient may not be included in the entry.

The fundamental objective of syndromic surveillance is to identify illness clusters early, before diagnoses are confirmed by laboratory testing and reported, and to mobilize a rapid response. Most syndromic surveillance systems classify patients into syndrome categories based either on the ICD admit diagnosis, which is only available in limited settings such as military hospitals, or the triage chief complaint. To classify a patient based on a chief complaint, some systems first preprocess the textual string to remove punctuation and replace abbreviations and acronyms with standardized terms. Next, a chief complaint classifier assigns a syndrome category to the chief complaint. Once a patient is classified into a syndrome category, a surveillance system can monitor patients for spatial and temporal aberrations that may indicate an outbreak.

Various chief complaint classifiers have been developed and evaluated for their ability to classify patients into accurate syndrome categories. Next, we describe both the characteristics of existing chief complaint classifiers and the performance with which the classifiers can detect and characterize outbreaks.

3.3.1.1 Characteristics of Chief Complaint Classifiers

In order to examine how chief complaint classifiers work and how they differ from each other, we distributed a survey to the developers of 12 classifiers:

ESSENCE CCP classifier [25]
Ontology-enhanced BioPortal CC classifier [22]
NC DETECT Syndrome Case Report [26]
Coded Chief Complaints (CCC-EDS) [27]
CC-MCSVM (Chief Complaints Multiclass SVM)
N-gram CC Classifier [16]
CoCo [15]
SyCo [17]
NYC Syndromic Macros [13]
BioSense Sub-syndromes [28]
EARS (TSS) [29]
MPLUS [18]

We characterized the individual classifiers based on the syndrome categories they map to and the methods they use for mapping text to syndromes.

The 12 classifiers in the survey map chief complaints to 20 unique syndrome categories. The only syndrome all classifiers map to is Respiratory syndrome. Other syndrome categories and the number of classifiers that use them include

Respiratory (12)
Gastrointestinal (7)
Hemorrhagic (5)
Severe illness or Death (3)
Localized cutaneous lesion (3)
Constitutional (2)
Diarrhea (2)
Vomiting (1)
Cold (1)
Injury (1)
Rash (7)
Fever (5)
Lymphadenitis (3)
Specific infection (2)
Influenza-like (2)
Meningoencephalitis (1)
Sepsis (1)
Shock/Coma (1)
Asthma (1)

The 12 classifiers also vary in whether they use keyword-based, statistical, or symbolic techniques, and in whether they map to concepts in a standardized vocabulary, as shown in Table 13-2. Statistical techniques include support vector machines (CC-MCSVM), n-gram classification using words and characters as n-grams (N-gram), naïve Bayes' (CoCo and SyCo), and Bayesian networks (MPLUS). BioPortal and NC Detect first map the text to UMLS concepts and then map the concepts to syndrome categories.

Table 13-2. Methods used to classify chief complaints.

Classifier	Keyword	Statistical	Symbolic	Map to Vocabulary
Essence	●			
BioPortal	●		●	●
NC Detect	●		●	●
CC-EDS	●			
CC-MCSVM		●		
N-gram		●		
CoCo		●		
SyCo		●		
NYC	●			
BioSense	●			
EARS (TSS)	●			
MPLUS		●	●	

As shown in Table 13-3, some classifiers classify the chief complaint directly to a syndrome category, e.g., "SOB" = Respiratory. Others first classify the chief complaint string to a clinical concept name, then map the concept name to a syndrome category using deterministic, probabilistic, or ontology-based mapping, e.g., "SOB" = *dyspnea* = Respiratory. The advantage of mapping directly to a syndrome category is simplicity. The advantage of mapping to concepts and then to syndromes is the ability for a user to create individualized syndrome definitions based on a set of limited reason-for-visit categories. This ability would be particularly useful for monitoring syndromes of interest in a particular location or if the user believed a new syndrome not ordinarily being monitored was developing (such as SARS in 2002). Techniques for mapping a concept to a syndrome are deterministic (e.g., *dyspnea* always gets mapped to Respiratory), probabilistic (e.g., P(Respiratory|*dyspnea*) = 0.92), or ontologic (e.g., *dyspnea* is a breathing disorder, which is a Respiratory illness).

Chief complaints may describe more than one reason for a patient's visit. Some chief complaint classifiers assign a single syndrome to a string, e.g., "short of breath/vomiting" = Respiratory. However, as shown in Table 13-4, most classifiers can assign multiple syndromes to a chief complaint string, e.g., "short of breath/vomiting" = Respiratory and Gastrointestinal.

Table 13-3. Mapping from chief complaints to syndromes.

Classifier	Map Directly to Syndromes	Map to Concepts Then to Syndromes	Technique for Mapping from Concept to Syndrome
Essence		●	Deterministic
BioPortal		●	Ontology
NC Detect	●		
CC-EDS		●	Deterministic
CC-MCSVM	●		
N-gram	●		
CoCo	●		
SyCo		●	Probabilistic
NYC	●		
BioSense		●	Deterministic
EARS (TSS)		●	Deterministic
MPLUS		●	Probabilistic

Table 13-4. Number of syndromes output for single chief complaint.

Classifier	Single Syndrome	Multiple Syndromes
ESSENCE		●
BioPortal		●
NC Detect		●
CC-EDS		●
CC-MCSVM		●
N-gram		●
CoCo	●	
SyCo		●
NYC	●	
BioSense		●
EARS (TSS)		●
MPLUS		●

3.3.1.2 Performance of Chief Complaint Classifiers

Because most surveillance systems rely on chief complaints to classify patients into syndromes, many of the studies applying NLP to biosurveillance have been aimed at evaluation of chief complaint processing.

Feature identification studies for chief complaint classification – Several studies have evaluated the ability to classify chief complaints into various syndrome categories in order to detect possible outbreaks. All but a few studies to date have addressed classification of chief complaints in the English language. One study used the BioPortal chief complaint classifier to classify Chinese chief complaints that were first translated with a Chinese-English translation module [30]. An n-gram classifier has been applied to chief complaints in Turkish [31] and Italian [32].

Keyword-based, statistical, and symbolic techniques have all successfully been applied to the problem of syndrome category classification for respiratory, gastrointestinal, neurological, constitutional, hemorrhagic, botulinic, rash, and other syndromes [13, 15–18, 22, 25–29], with sensitivities and specificities generally between 80 and 100%. A main challenge in classifying chief complaints into syndrome categories is the prevalence of acronyms, abbreviations, and misspellings in chief complaint text [33, 34]. For that reason, many of the chief complaint classifiers use a preprocessor for normalizing the text. However, one study showed that typical preprocessing techniques such as spell-checking and synonym replacement did not improve classification performance for a statistical or a keyword-based classifier [14]. Several classifiers [14, 22, 35] preprocess chief complaints with the emergency medical text processor (EMT-P), a system for cleaning chief complaint text data [33]. EMT-P not only cleans chief complaints but also splits chief complaints into multiple problems when relevant. For example, EMT-P would split "left sided numbness/pain" into one complaint for "left sided numbness" and one for "left sided pain." Splitting chief complaints significantly improved classification performance of the naïve Bayesian classifier CoCo [14] by allowing CoCo to assign multiple classifications to a single chief complaint. In addition, EMT-P maps each problem in a chief complaint to a UMLS concept, which can then be mapped to a syndrome category.

Case detection studies for chief complaint classification – Although applications that process chief complaints show high technical accuracy, when applied to the problem of case detection, performance is only moderate due to the fact that chief complaints often do not contain enough clinical detail to accurately represent the patient's clinical state. Case classification studies have compared the classification made from the chief complaint against a reference standard classification based on either ICD-9 discharge diagnoses [36, 37] or manual review of clinical textual reports [12, 38, 39].

The studies have shown sensitivities ranging from 10 to 77%, depending on the syndrome. Chief complaint classification typically performs with higher sensitivity when classifying more common syndromes like respiratory and gastrointestinal and with lower sensitivity on more rare syndromes like botulinic syndrome. Moderate case detection ability is a disadvantage of using chief complaints for surveillance, but the predictive performance must be weighed against the great advantages of nearly ubiquitous availability and timeliness.

When attempting to detect more specific syndromes, like febrile syndromes in which a patient has a fever and a symptom related to a relevant syndrome (e.g., "fever and cough"), sensitivity plunges to between 0 and 12% [40]. This study suggests that chief complaints, although useful for general syndrome classification, do not contain enough information to identify febrile syndromes.

Characterization of outbreaks from chief complaints – Because chief complaints do not present a granular view of a patient's clinical state, they may not provide an accurate method for characterizing outbreaks, but a few studies have shown that chief complaints can contribute to outbreak characterization. A study on fever detection from chief complaints showed 61% sensitivity but 100% specificity, indicating that all patients with chief complaints describing a fever actually had a fever. Although not all patients with fever were detected from chief complaints, investigation of an outbreak for which chief complaints showed a high rate of fever could maintain confidence in the positive febrile classifications. Another study investigated peaks indicated by chief complaint syndrome classifications to look for concerning signs and symptoms [41]. By allowing the chief complaint-generated peaks to direct the investigation, the investigators could quickly and easily target patients of concern for manual chart review.

In our introductory scenario of attempting to detect cases of avian influenza, classification of chief complaints to Respiratory syndrome would be expected to have a moderate to high sensitivity for identifying outbreaks of respiratory disease. However, the specificity would be low; the outbreak could be due to avian influenza or any other respiratory condition, including non-infectious causes. The cases in the Respiratory outbreak would have to be examined further to determine if the features of avian influenza were present. Ultimately, natural language processing could also be used to identify conditions of concern in textual reports for patients comprising those peaks, and several researchers are developing and evaluating tools for just that purpose, as we describe next.

3.3.2 Ambulatory Visit Notes

In some institutions, triage nurses not only generate a chief complaint but also a triage note that expands on the chief complaint. After triage, once a patient is admitted to an ambulatory facility and has been examined by a physician, the physician generates an ambulatory note that typically describes the patient's reason for coming to the facility, the history of the present illness, the patient's past medical history, relevant diseases experienced by family members, results of a physical examination and of tests performed while the patient was in the ambulatory facility, a diagnosis if there is one, and the plan for managing the patient's illness.

Ambulatory visit notes are required documents necessary for evaluation of quality of care delivery, for billing, and for legal purposes. The notes may be generated by physicians through dictation and transcription, may be directly entered into the EMR through typing or structured entry, or may be handwritten. Therefore, not all healthcare facilities have electronically available ambulatory visit notes, and the timeliness of electronic notes varies depending on the mode of generation.

Ambulatory visit notes contain symptoms, findings, test results, descriptions of chronic conditions, and diagnoses that can be very useful in detecting and understanding outbreaks. Since many people with sudden onset of a new illness may first present to an ambulatory care facility, the reports promise granular and timely information on the nature of an outbreak.

A few studies have verified the intuitive hypothesis that information in emergency department notes can more accurately classify patients into syndrome categories than can chief complaints [42–44]. Researchers have begun developing and evaluating information extraction applications for identifying clinical conditions relevant for detection and characterization of outbreaks from triage notes and emergency department reports. EMT-P has been used in conjunction with a negation algorithm called NegEx [8] to index UMLS concepts in triage reports [45]. Indexed triage notes can provide a timely but more descriptive characterization of a patient than a chief complaint can. A system called Topaz [46] identifies 55 respiratory-related clinical conditions such as shortness of breath, hypoxia, and pneumonia from emergency department reports. Identifying the conditions with Topaz involves indexing UMLS terms using MetaMap [47], applying regular expressions to conditions that require a numeric value (such as fever), and determining whether the condition is negated, occurred in the past history, was mentioned hypothetically, or was experienced by someone other than the patient [48].

Having access to coded representations of clinical conditions in ambulatory notes promises to improve the ability to detect relevant cases and therefore to detect outbreaks by providing a more granular description of a patient's illness in a relatively timely manner. For example, the detection of coded representations of influenza-like illness could lead to the timely detection of an outbreak of avian influenza with much greater specificity than monitoring for an increase in respiratory syndrome cases. In addition to improving detection capabilities, coded clinical conditions from ambulatory care reports could be helpful in guiding and informing investigation of a potential outbreak. In the case of avian influenza, ED reports could inform investigators of the clinical features and the severity of the disease, for example, whether pneumonia was a characteristic of the illness.

3.3.3 Inpatient Reports: Progress and Findings

Once a patient is admitted to a healthcare facility there are important sources of free-text clinical data generated during the normal clinical care process that provide useful information for biosurveillance. These inpatient free-text clinical documents can be classified into two categories: reports that document inpatient clinical progress or lack thereof and reports that characterize and document specific findings that occur as a result of tests and procedures. Progress reports include notes written by clinicians documenting clinical progress, procedures, and consultations requested or provided. Inpatient progress notes may contain only narrative text, may follow a traditional SOAP note format (i.e., subjective, objective, assessment, plan), or may include templated sections including subheadings that direct the content of each specific note section. Findings are documented in free text from specific diagnostic testing, including impressions and results from radiology procedures, pathology, and microbiology. Both of these inpatient report types may contain specific header information that lists the patient, date of patient encounter, diagnosis or service provided, and contextual information regarding history of illness.

Free-text clinical documents are used to convey the decision-making process, perspectives, and plans of the clinicians responsible for a patient's care. Therefore, inpatient free-text narratives link the various components of the medical record and describe the temporal clinical course as well as severity of a patient's hospitalization. Inpatient narratives are also used by coders and financial departments for medical billing purposes and to create the legal medical record for the patient. Inpatient records could theoretically be useful in a more complete characterization of avian influenza cases. They might be the only source of information such as patient travel to an area where the disease is already known to be occurring or the fact that a patient is homeless. Inpatient records could also be the source of information on testing done for viral diseases and the results, although, if available, direct examination of laboratory records would probably provide easier access to such information.

Though electronic inpatient reports are a central component of the electronic medical record, they are not readily extractable or widely available from most EMR systems. This is starting to change, for example, in the case of healthcare systems and hospitals like the Department of Veterans Affairs, Kaiser Permanente, the Mayo Clinic, the Cleveland Clinic, Columbia Presbyterian Medical Center, and the University of Pittsburgh Medical Center, but the majority of this development is largely driven by vendor-supported product solutions and large data warehousing efforts at the local level.

Compared with ED chief complaints and ambulatory care notes, inpatient reports provide a less timely source of information for biosurveillance purposes (see Figure 13-1). In some systems, through concurrent charting, progress reports are immediately available; in other systems, there is a substantial delay in availability of the documents. Even if the information is not timely enough for early detection of an outbreak, because of the granularity of the information described in these reports, information extracted from inpatient document sources can be used to describe the magnitude and severity of an outbreak.

Like ambulatory visit notes, progress and finding reports are complex narratives involving detailed descriptions of the diagnostic process over time and therefore pose substantial challenges to NLP applications. There may be a certain amount of redundancy found in inpatient clinical document sources, particularly for those documenting patient clinical progress – this redundancy can be a blessing and a curse to NLP applications. In some cases, these documents are also driven by templated document structures meant for ease of information entry and may also contain imbedded note sections that contain laboratory, medication management, and other results information. Inpatient documents, particularly progress reports, may also be very lengthy and may contain copying and pasting of previously entered information by other providers, thus introducing ambiguity and contradictory information that NLP applications or biosurveillance algorithms may have difficulty reconciling.

In the context of outbreak detection, many of the studies using inpatient reports have focused on chest radiograph reports for identification of patients with pneumonia-related findings [19, 49, 50]. These applications have been able to identify pneumonia-related concepts in radiology reports with accuracies similar to that of physicians reading the reports, with sensitivities and specificities above 90%. One study created a text processor to identify chest radiographs with mediastinal widening consistent with an anthrax infection [51]. And another identified tuberculosis from chest X-ray reports [52] using an NLP application called MedLEE [53]. MedLEE contributed to a larger decision support system for identifying patients with possible tuberculosis who were not yet isolated [52, 54], and the chest radiograph-based rule that used MedLEE's output was the most useful rule for improving tuberculosis respiratory isolation.

To our knowledge, no studies have compared accuracy and timeliness of inpatient reports for case detection. But one study did show increased sensitivity and specificity of case detection for influenza-like illness from the full patient record, which includes inpatient reports, when compared to chief complaints or ambulatory notes [44]. Inpatient reports may also contribute to outbreak characterization by providing descriptions of clinical conditions that are not contained in the structured record. Moreover, inpatient notes often record epidemiological factors that can help locate the source of an outbreak or characterize the extent of spread. Gundlapalli and colleagues [55] used MedLEE to identify factors such as homelessness, alcohol or drug abuse, and exposure to infected individuals among a cohort of patients seen at a VA healthcare system.

3.3.4 Discharge Reports

Discharge reports describe the clinical condition of the patient at the time of hospital discharge or transfer to another facility for additional care. These document sources also provide the reason for hospitalization, a summary of the care provided during the course of hospitalization, significant findings, procedures performed, discussion and planning for follow-up care, and sometimes clinician instructions given to other providers who may provide future care for that patient. Discharge reports may also contain brief instructions given to the patient and family for recovery and follow-up care. There are two general categories of discharge reports: those that are written for patients discharged from the hospital alive, called discharge summaries, and death reports written for patients who die during hospitalization.

Discharge reports are required documents necessary for evaluation of quality of care delivery, billing, and legal purposes. Thus every inpatient admission should have at least a short summary of pertinent information for documentation of the clinical care provided during hospitalization and the status of the patient at the time of discharge. Like electronic inpatient notes, the availability of electronic discharge reports depends on the institution providing care and the degree with which that institution has implemented EMR technology. One study compared electronic to handwritten discharge summaries and found that that electronic discharge summaries are in some ways lower quality than handwritten ones [56].

Because discharge reports contain a summary of the clinical course and status of the patient at time of discharge, the reports provide a level of granularity of information other document sources do not contain. Death reports usually have a short free-text section listing various conditions leading to the cause of death and are provided to public health entities for purposes of public record.

Compared with emergency department chief complaints, discharge reports are on the opposite side of the spectrum of information timeliness. Discharge reports may be available anywhere from a day to weeks after hospitalization. Because of their lack of timeliness, discharge notes would likely add little to the detection or characterization of an outbreak of an acute infectious disease like avian influenza, which may account for the dearth of biosurveillance NLP research directed at discharge summaries in spite of several applications developed for information extraction from discharge summaries, such as pneumonia severity, negated conditions, de-identification of patient health information, and smoking status.

The causes of death from death certificates are routinely coded to yield structured data and are used for determining the mortality associated with pneumonia and influenza. These are reported as weekly statistics by the CDC (http://www.cdc.gov/flu/weekly/) and the data are used to assess the severity of the annual influenza season in the U.S. Although there are references to mining cause of death from death certificates for various conditions, we know of no research on information extraction from death notes for the specific purpose of biosurveillance. Death certificates could be important in retrospectively searching for patients with diseases such as avian influenza who died before a confirmatory diagnosis was made as they may have been coded generically as influenza and/or pneumonia.

4. CONCLUSION AND DISCUSSION

A patient's encounter with the healthcare system generates a variety of clinical information, a large portion of which is in free-text format. For the purposes of biosurveillance, it is first important to clarify the question we wish to answer. Are we concerned with the detection of an infectious disease outbreak among a group of patients seen at a facility? Or are we concerned with looking deeper into the characterization of the outbreak in terms of epidemiological factors that are associated with that disease? The data sources and techniques required to answer these two distinct, though related, questions may be different. Another key issue to consider is timeliness – early detection relies on data sources that are available earlier and that are amenable to efficient processing. In contrast, clinical notes that offer epidemiological clues are available later and often require more detailed processing.

In spite of years of research applying NLP techniques to clinical reports, much work still remains before we reach the stage when all clinical notes are routinely processed, annotated, and catalogued by computers for use in biosurveillance or other application areas. Routine application of NLP for biosurveillance will require moving beyond the task of identifying clinical conditions in isolation (e.g., finding instances of cough in a report) toward understanding the context of the clinical condition within the reports and across different reports. To understand the content of a report the way a human reading the report would, our applications must identify information such as whether the condition is negated, how much certainty the physician has in the diagnosis, who experienced the condition, when the condition began and how long the condition has existed. Moreover, to automatically summarize a patient's clinical state over time for investigation of an outbreak, an NLP application will require the ability to integrate information across reports and model the flow of information over time.

At one level, we may need a paradigm shift in NLP methodologies for a machine to read and summarize a clinical note as intended by the original author. Understanding the situation as the author of the report does may require a fusion of patient-level data with population-level knowledge of circulating infectious diseases. Though challenging, the benefits of this work will extend beyond biosurveillance to other domains such as adverse events detection, injury surveillance, chronic disease management, and clinical/ translational research.

In the meantime, the techniques and applications we have described in this chapter can be useful tools in combination with human knowledge for detection and characterization of outbreaks. Natural language processing techniques may not be capable of providing intelligent substitutes for clinical record investigation, but extracting and encoding information from the unstructured portion of the EMR can assist humans in detecting, investigating, and responding to outbreaks.

ACKNOWLEDGMENTS

Authors funded by NLM R01LM009427-01 "NLP Foundational Studies and Ontologies for Syndromic Surveillance from ED Reports," CDC Grant 1 U38 HK000013, NIH-NIAID Grant U54 AI-065357, and CDC Utah Center of Excellence in Public Health Informatics 1 PO1 CD000284-01. We thank the University of Pittsburgh Medical Center, University of Utah, and VA Salt Lake City Health Care System for resources and access to data. We also thank developers of chief complaint-based surveillance systems for their input, responses to surveys, and willingness to collaborate on education and research in spite of scarce resources.

QUESTIONS FOR DISCUSSION

1. What types of epidemiological and clinical variables can be gleaned from free-text data?
2. Describe the benefits of structured vs. unstructured data sources in terms of sensitivity and specificity of data.
3. Describe the strengths and weaknesses of keyword, statistical, and symbolic NLP techniques.
4. Design a three-stage evaluation for surveillance from a particular free-text clinical data source. Describe how you would evaluate an NLP application extracting information from the text for feature detection, case detection, and outbreak detection. For each evaluation stage, describe what the reference standard would be.
5. Compare and contrast timeliness of data and granularity of information. Which is most important for outbreak detection?
6. Why would application of chief complaint classification produce different results across syndromic categories? Which syndrome categories do you think are easiest/hardest to classify based on chief complaints? Why?
7. Compare and contrast different sources of textual data for biosurveillance. List one potential biosurveillance application for each source.
8. Describe how textual data can be used to identify case severity.
9. What information is contained in discharge summaries that would be useful for biosurveillance?

REFERENCES

1. Brownstein J, Freifeld C. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill 2007;12(11):E071129.5. Epub 2007 Nov 29.
2. Doan S, Hung-Ngo Q, Kawazoe A, Collier N, editors. Global Health Monitor – a Web-based system for detecting and mapping infectious diseases. Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Companion Volume; 2008.
3. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008;15(2):150–7.
4. Wilson JMt, Polyak MG, Blake JW, Collmann J. A heuristic indication and warning staging model for detection and assessment of biological events. J Am Med Inform Assoc 2008;15(2):158–71.
5. Canada Global Public Health Intelligence Network (GPHIN). Public Health Agency of Canada 2008 [cited 2008 May 26, 2008]; Available from: http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk_e.html.

6. Chapman WW. Natural language processing for biosurveillance. In: Wagner MM, Moore AW, Aryel RM, editors. Handbook of Biosurveillance. Burlington: Elsevier Academic Press; 2006.
7. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34(1–2):193–201.
8. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
9. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001;8(6):598–609.
10. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005;5(1):13.
11. Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc 2007;14:304–11.
12. Chapman W, Dowling J, Wagner M. Fever detection from free-text clinical records for biosurveillance. J Biomed Inform 2004;37:120–7.
13. Heffernan R, Mostashari F, Das D, Karpati A, Kuldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerg Infect Dis 2004;10(5):858–64.
14. Dara J, Dowling JN, Travers D, Cooper GF, Chapman WW. Evaluation of preprocessing techniques for chief complaint classification. J Biomed Inform 2008;41(4):613–23.
15. Olszewski RT, editor. Bayesian classification of triage diagnoses for the early detection of epidemics. FLAIRS Conference; 2003; St. Augustine, FL.
16. Brown P, Halasz S, Cochrane DG, Allegra JR, Goodall CR, Tse S. Optimizing performance of an Ngram method for classifying emergency department visits into the respiratory syndrome. Adv Dis Surveill 2007;2:1.
17. Espino JU, Dowling J, Levander J, Sutovsky P, Wagner MM, Cooper GF. SyCo: a probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. Adv Dis Surveill 2007;2:5.
18. Chapman W, Christensen L, Wagner M, Haug P, Ivanov O, Dowling J, et al. Classifying free-text triage chief complaints into syndromic categories with natural language processing. Artif Intell Med 2005;33(1):31–40.
19. Fiszman M, Chapman W, Aronsky D, Evans R, Haug P. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.
20. Brown P, Halasz S, Cochrane DG, Allegra JR, Goodall C, Tse S. Optimizing performance of an *Ngram* method for classifying emergency department visits into the respiratory syndrome. Adv Dis Surveill 2007;2:1.
21. Christensen L, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proc Workshop on Natural Language Processing in the Biomedical Domain 2002:29–36.
22. Lu HM, Zeng D, Trujillo L, Komatsu K, Chen H. Ontology-enhanced automatic chief complaint classification for syndromic surveillance. J Biomed Inform 2008;41(2):340–56.
23. Moran GJ, Talan DA. Update on emerging infections: news from the Centers for Disease Control and Prevention. Syndromic surveillance for bioterrorism following the attacks on the World Trade Center – New York City, 2001. Ann Emerg Med 2003;41(3):414–8.

24. Broome CV, Pinner RW, Sosin DM, Treadwell TA. On the threshold. Am J Prev Med 2002;23(3):229–30.
25. Sniegoski CA. Automated syndromic classification of chief complaint records. Johns Hopkins APL Technical Digest 2004;25(1):68–75.
26. Scholer MJ, Ghneim GS, Wu SW, Westlake M, Travers DA, Waller AE, et al. Defining and applying a method for improving the sensitivity and specificity of an Emergency Department early detection system. Proc. 2007 AMIA Fall Symposium; 2007.
27. Thompson DA, Eitel D, Fernandes CM, Pines JM, Amsterdam J, Davidson SJ. Coded chief complaints – automated analysis of free-text complaints. Acad Emerg Med 2006;13(7):774–82.
28. Hales C, Coberly J, Tokars J. Defining clinical condition categories for biosurveillance. Adv Dis Surveill 2007;4:95.
29. Lawson BM, Fitzhugh EC, Hall SP, Franklin C, Hutwagner LC, Seeman GM, et al. Multifaceted syndromic surveillance in a public health department using the early aberration reporting system. J Public Health Manag Pract 2005;11(4):274–81.
30. Lu HM, King CC, Wu TS, Shih FY, Hsiao JY, Zeng D, et al. Chinese chief complaint classification for syndromic surveillance. Lect Notes Comput Sci 2007;4506:11–22.
31. Brown P, Oktay C, Cevik AA, Kilicaslan I, Goodall CR, Halasz S, et al. Sensitivity and specificity of an Ngram method for classifying emergency department visits into the respiratory syndrome in the Turkish language. Adv Dis Surveill 2007;4:44.
32. Brown P, Morabito G, Halasz S, Goodall CR, Cochrane DG, Tartaglino B, et al. The performance of a NGram classifier for patients' chief complaint based on a computerized pick list entry and free text in an Italian emergency department. Adv Dis Surveill 2007;4:45.
33. Travers DA, Haas SW. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. J Biomed Inform 2003;36(4–5):260–70.
34. Shapiro AR. Taming variability in free text: application to health surveillance. MMWR Morb Mortal Wkly Rep 2005;53:95–100.
35. Travers D, Shiying W, Scholer MJ, Westlake M, Waller A, McCalla AL. Evaluation of a chief complaint pre-processor for biosurveillance. AMIA 2007 Symposium Proceedings; 2007:736–40.
36. Chang HG, Cochrane DG, Tserenpuntsag B, Allegra JR, Smith PF. ICD9 as a surrogate for chart review in the validation of a chief complaint syndromic surveillance system. Adv Dis Surveill 2006;1:11.
37. Chapman W, Dowling J, Wagner M. Classification of emergency department chief complaints into seven syndromes: a retrospective analysis of 527,228 patients. Ann Emerg Med 2005;46:445–55.
38. Chang HG, Cochrane DG, Tserenpuntsag B, Allegra JR, Smith PF, editors. Validation of a syndromic system based on patients' chief complaints using chart review. National Syndromic Surveillance Conference; 2004; Boston, MA.
39. Beitel A, Olson K, Reis B, Mandl K. Use of emergency department chief complaint and diagnostic codes for identifying respiratory illness in a pediatric population. Pediatr Emerg Care 2004;20(6):355–60.
40. Chapman W, Dowling J. Can chief complaints identify patients with febrile syndromes? Adv Dis Surveill 2007;3:1–9.
41. Chang HG, Chen JH, Cochrane D, Allegra J, Smith P. The use of sub-syndromes to investigate peaks in a syndromic surveillance system. Acad Emerg Med 2007;14(5 Suppl 1):S179–80.

42. Elkin PL, Brown SH, Balas A, Temesgen Z, Wahner-Roedler D, Froehling D, et al. Biosurveillance evaluation of SNOMED CT's terminology (BEST trial): coverage of chief complaints. Int J Med Inform 2010;79(4):e71–5.
43. Chapman W, Dowling J, Cooper G, Hauskrecht M, Valko M. A comparison of chief complaints and emergency department reports for identifying patients with acute lower respiratory syndrome. Adv Dis Surveill 2007;2:195.
44. South BR, Chapman WW, Delisle S, Shen S, Kalp E, Perl T, et al. Optimizing syndromic surveillance text classifiers for influenza-like illness: does document source matter? Proc 2008 AMIA Fall Symposium (under review) 2008.
45. Ising A, Travers D, Crouch J, Waller AE. Improving negation processing in triage notes. Adv Dis Surveill 2007;4:50.
46. Chu D, Dowling JN, Chapman WW. Clinical feature extraction from emergency department reports for biosurveillance [Master's]. Pittsburgh: University of Pittsburgh; 2007.
47. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.
48. Chapman WW, Chu D, Dowling JN, editors. ConText: an algorithm for identifying contextual features from clinical text. BioNLP Workshop of the Association for Computational Linguistics. Czech Republic: Prague; June 29, 2007.
49. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122(9):681–8.
50. Elkin P, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma HB. NLP-based Identification of Pneumonia Cases from Free-Text Radiological Reports. AMIA Annu Symp Proc. 2008; 2008:172–176.
51. Chapman W, Cooper G, Hanbury P, Chapman B, Harrison L, Wagner M. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 2003;10(5):494–503.
52. Hripcsak G, Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Inform Assoc 1997;4(5):376–81.
53. Friedman C, Alderson P, Austin J, Cimino J, Johnson S. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.
54. Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. Infect Control Hosp Epidemiol 1998;19(2):94–100.
55. Gundlapalli AV, South BR, Chapman WW, Phansalkar S, Shen S, Delisle S, et al. Adaptable NLP-based surveillance methods for epidemiologic case finding and investigation using VA electronic medical records. Proc 2008 AMIA Fall Symposium (under review); 2008.
56. Callen JL, Alderton M, McIntosh J. Evaluation of electronic discharge summaries: a comparison of documentation in electronic and handwritten discharge summaries. Int J Med Inform 2008;77(9):613–20.

SUGGESTED READING

Chapman WW. Natural language processing for outbreak and disease surveillance. In Handbook of biosurveillance, Elsevier Inc., New York, NY; 2005.

Jurafsky D, Martin JH. Speech and language processing (2nd ed.). Upper Saddle River, New Jersey: Prentice-Hall, Inc.; 2008.

Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128–44.

Travers DA, Haas SW. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. J Biomed Inform 2003;36(4–5):260–70.

ONLINE RESOURCES

Biosurveillance Standards

Consensus syndrome definitions, ontology, and API: http://www.code.google.com/p/syndef/.

Data Sets

Repository of de-identified clinical notes: http://www.dbmi.pitt.edu/blulab/nlprepository.

Chief complaint set: http://www.code.google.com/p/syndef/.

Open Source NLP Tools and Algorithms

clinical Text Analysis and Knowledge Extraction System (cTAKES): http://www.ohnlp.org.

Hitex: https://www.i2b2.org/software/index.html.

NegEx and ConText: http://www.code.google.com/p/negex/.

CoCo: http://www.openrods.sourceforge.net/.

NLP toolkit in the Python programming language: http://www.nltk.org/Home.

Chapter 14

KNOWLEDGE MAPPING FOR BIOTERRORISM-RELATED LITERATURE

YAN DANG*, YULEI ZHANG, HSINCHUN CHEN,
and CATHERINE A. LARSON

CHAPTER OVERVIEW

This chapter describes major Knowledge Mapping techniques and how they are used for mapping bioterrorism-related literature. The invisible college, which consists of a small group of highly productive and networked scientists and scholars, is believed to be responsible for the growth of scientific knowledge. By analyzing scholarly publications of these researchers using select content analysis, citation network analysis, and information visualization techniques, Knowledge Mapping helps reveal this interconnected invisible college of scholars and their ideas. This chapter outlines the important techniques used in Knowledge Mapping, presents how these techniques are used for mapping bioterrorism-related literature, and shows some findings related to the productivity status, collaboration status, and emerging topics in the bioterrorism domain.

Keywords: Knowledge mapping; Invisible college; Bioterrorism-related literature

1. INTRODUCTION

In Diane Crane's seminal book on "Invisible Colleges: Diffusion of Knowledge in Scientific Communities" (Crane, 1972), she suggests that it is the

* *Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721, USA, ydang@email.arizona.edu*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_14,

"invisible college," a small group of highly productive scientists and scholars, that is responsible for the growth of scientific knowledge. Crane shows that many scientific disciplines go through similar stages of initiation, growth, expansion, maturation, and decline. The productive scientists and scholars form a network of collaborators in promoting and developing their fields of study. The presence of an invisible college or network of productive scientists linking separate groups of collaborators within a research area has been evident in many studies (Chen, 2003; Shiffrin and Börner, 2004).

"Knowledge Mapping," which is based on content analysis, citation network analysis, and information visualization, has become an active area of research that helps reveal such an interconnected, invisible college or network of scholars and their seminal publications and ideas.

According to Chaomei Chen in his *Mapping Scientific Frontiers* book (Chen, 2003), Knowledge Mapping helps "depict the spatial relations between research fronts, which are areas of significant activity. Such maps can also simply be used as a convenient means of depicting the way in which research areas are distributed and conveying added meaning of their relationships… By using a series of chronically sequential maps, one can see how knowledge advances. Mapping scientific frontiers involves several disciplines, from the philosophy and sociology of science, to information science, scientometrics, and information visualization."

Two forces are contributing to the rapid development and the overwhelming interest in Knowledge Mapping. First, the availability of online publications, from scientific Abstracts and Indexes (A&I), full-text articles, and online preprints, to digital dissertations, multimedia (e.g., video and audio files) magazine and journal articles, and multilingual Web-accessible patent filings, has made it possible to more systematically examine the scientific output produced by members of the invisible college. Secondly, recent advances in text mining, network analysis, and information visualization techniques have provided more scalable and accurate methods to understand and reveal the interconnections between scientific disciplines and scholars.

Bioterrorism attacks against civilians are usually intended to cause widespread panic and terror (Lane et al., 2001). Bioterrorism has been given a high priority in national security since 9/11 and the Anthrax attacks. The U.S. Government has attempted to monitor and regulate biomedical research labs, especially those that study bioterrorism agents/diseases. However, monitoring worldwide biomedical researchers and their work is still an issue. Given the explosive growth of literature resources in the biomedical domain and advances in text mining, network analysis, and information visualization techniques, Knowledge Mapping can be used to help monitor worldwide bioterrorism research.

The remainder of this chapter is organized as follows. The next section discusses the literature review from three perspectives: the online resources, the units of analysis, and the analysis techniques for Knowledge Mapping. The subsequent section shows the research design. The following two sections provide an explanation of the testbed and analysis results, as well as a discussion of mapping bioterrorism literature. The last section summarizes this chapter.

2. LITERATURE REVIEW

2.1 Online Resources for Knowledge Mapping

Various online resources are available for mapping scientific knowledge. They vary from formal to informal publications; from text-based to multimedia presentations; and from academic literature to industry-relevant international patents.

One type of online resources is abstracts and indexes (A&I). A&I contain abstract and index (bibliographic) information and are used to locate articles, proceedings, and occasionally books and book chapters in various subjects. Most abstracts and indexes are available electronically. Public and university libraries often subscribe to such databases and services. Only a very few biological or scientific databases are searchable for free on the Web, primarily databases generated by the National Library of Medicine (http://www.nlm.nih.gov/), such as MEDLINE (medicine) or TOXLINE (toxicology). There are A&I databases in almost every subject area, e.g., BIOSIS (biology), COMPENDEX (engineering and technology), ERIC (education), etc.

Another type of online resources includes commercial full-text journal articles and digital libraries. Many commercial publishers have made their online content available on the Web. The most prominent service of this type is provided by the Web of Science (http://scientific.thomson.com/products/wos/), a product of Thomson Scientific. The Web of Science provides seamless access to current and retrospective information from approximately 8,700 research journals from around the world. More recently, many professional societies have made their articles available through various digital libraries. For example, the ACM Digital Library (http://portal.acm.org/dl.cfm) contains 54,000 online articles from 30 journals and 900 proceedings of the Association for Computing Machinery. The IEEE Computer Society Digital Library (http://www.computer.org/portal/site/csdl/index.jsp) provides online access to 18 IEEE journals and 150 proceedings in computer science.

A third type of online resources contains free full-text articles and e-prints. In addition to commercial resources for journal articles, there is also a grassroots movement initiated by the academic community to provide free access to journals and books. For example, on the Free Medical Journals site (http://www.freemedicaljournals.com/), users can find many important academic journals made available online, free and in full-text. HighWire Press (http://highwire.stanford.edu/lists/freeart.dtl), a service affiliated with Stanford University, is believed to be the largest archive of free full-text science articles. As of December 20, 2006, it provides access to more than 1.5 million free full-text articles in many subject disciplines. In some scientific disciplines, e-prints (scientific or technical documents circulated electronically to facilitate peer exchange, including preprints and other scholarly papers) are strongly encouraged and accepted by the community. For example, the arXiv.org service (http://arxiv.org/), supported by Cornell University, provides open access to about 400,000 e-prints in Physics, Mathematics, Computer Science, and Quantitative Biology.

Citation indexing systems and services form another type of online resources. In addition to accessing bibliographic and full-text content of scientific articles, aggregated and individualized citation information is critical in the assessment of highly-cited, influential papers and authors. The Science Citation Index (http://scientific.thomson.com/products/sci/), a product of Thomson Scientific, provides access to bibliographic information, abstracts, and cited references in 3,700 scholarly science and technical journals worldwide covering more than 100 disciplines. A recent service provided by Google Scholar (http://scholar.google.com/intl/en/scholar/) also supports broad access to scholarly literature. A user can search across many disciplines and sources: peer-reviewed papers, theses, books, abstracts, and articles. The service features many advanced search functionalities, including ranking articles based on how often an article has been cited in other scholarly literature. CiteSeer (http://citeseer.ist.psu.edu/citeseer.html) is another example of an advanced search system (for computer and information science literature) that is built upon citation information. It was one of the first digital libraries to support automated citation indexing and citation linking.

In addition to formal literature published in journals, magazines, and conference proceedings, Ph.D. and Master's theses and dissertations constitute a significant part of scientific knowledge generated. University Microfilms (UMI) was founded in 1938 to collect, index, film, and republish doctoral dissertations in microfilm and print. Currently, UMI's dissertation abstract database has archived over 2.3 million dissertations and Master's theses. Some two million of them are available in print, microfilm, and digital format, via its ProQuest system (http://il.proquest.com/brand/umi.shtml). More recently, the Networked Digital Library of Theses and Dissertations (NDLTD, http://www.ndltd.org/) was formed to promote the adoption, creation, use, dissemination, and preservation of electronic analogues to the traditional paper-based theses and dissertations. Via electronic theses and dissertations (ETD), graduate students learn electronic publishing as they engage in their research and submit their own work, often in a rich multimedia format. Universities learn about digital libraries as they collect, catalog, archive, and make ETDs accessible to scholars worldwide.

Patent publications have often been used to evaluate science and technology development status worldwide (Narin, 1994). While academic literature represents fundamental scientific knowledge advancement, patents reveal scientific and technological knowledge that has a strong potential for commercialization. There are several governmental or intergovernmental patent offices that control the granting of patents in the world. The United States Patent and Trademark Office (USPTO, http://www.uspto.gov/), European Patent Office (EPO, http://www.european-patent-office.org/index.en.php), and Japan Patent Office (JPO, http://www.jpo.go.jp/) combined issue nearly 90% of the world's patents. USPTO handles over 6.5 million patents with 3,500–4,000 newly granted patents each week. EPO handles over 1.5 million patents with more than 1,000 newly granted patents each week. JPO handles over 1.7 million patents with 2,000–3,000 newly granted patents each week. All three patent offices provide search systems for Web-based access.

Business and industry articles and reports are also important for knowledge mapping. Critical science and technology knowledge eventually flows from academic literature and patents to various industries and companies. At the other end of the knowledge mapping resources are various business and industry articles and reports; some are reported in general-interest science and technology magazines and newspapers, while others can be purchased from industry-specific consulting firms. For example, timely, in-depth industry-specific or technology-specific reports are available at sites such as: Forrester (http://www.forrester.com), IDC (http://www.idc.com), and Gartner (http://www.gartner.com), among others.

In addition to the abovementioned formal publications generated by scholars, students, and industry practitioners, the Web has enabled virtually anyone to become an online publisher. There is potentially interesting scientific, product, and marketing information that has been generated and disseminated by various platforms over the Web, such as Web pages, forums, chat rooms, blogs, multimedia sites, social networking sites, and virtual worlds. However, the diversity and quality of such information varies significantly. It is often quite difficult to use such Web-based, self-produced information for technology assessment or knowledge mapping.

2.2 Units of Analysis for Knowledge Mapping

For knowledge mapping analysis, pre-processing of raw online resources is needed. Each article, patent, or report needs to be processed to identify key indicators for further analysis and comparison. Among the most common units of analysis for knowledge mapping are: authors or inventors, publications and publication outlets, institutions (companies or universities), countries or regions, subject and topic areas (broad categories or specific topics), and timeline (publication date).

Authors or inventors: The most critical unit of analysis for knowledge mapping consists of the researchers, authors, and inventors who are the productive members in the invisible college. Extracting the author or inventor field from various knowledge sources is a non-trivial task. Although HTML, XML, and structured database representations have made automatic name identification easier (than in the paper-based format), author name extraction and identification is difficult in different cultural contexts (e.g., recognizing Chinese names), especially when a publication does not contain complete first and last names. For example, common names such as "W. Zhang" and "L. Liu" abound in the Chinese Academy of Sciences (one of the most productive and largest academic research institutions in the world).

Publications and publication outlets: Different academic publications have different levels of prestige; most are measured based on their Impact Factor (an aggregate, normalized number based on citation counts). For example, the Impact Factor of *Science* was 30.927 in 2005; while the *Journal of Computational Biology* Impact Factor was 2.446. There are many other publications that do not even have an Impact Factor score. In order to determine the value and impact of a researcher's work, quality is more important than quantity. Quality is often determined based on the prestige of a publication outlet. The number of citations is also a major determinant. A seminal or landmark paper can often help define a person's career or a particular field. For example, while many good academic articles are cited hundreds of times, Albert Einstein's seminal paper on "Can quantum-mechanical description of physical reality be considered complete?" that appeared in *Physical Review* in 1935 was cited 3,753 times (based on a search on Google Scholar). Based on analysis results reported by ScienceWatch (http://www.sciencewatch.com/), the most cited paper of the past two decades (1983–2002) was: Chomczynski, N. Sacchi, "Single-step method of RNA isolation by acid guanidinium thiocyanate phenol chloroform extraction," *Analytical Biochemistry*, 162(1): 156–9, 1987. The paper received a citation count of 49,562 (based on data from Thomson Scientific's Web of Science). However, correctly parsing and identifying unique publication names is a difficult task as many databases record those names in cryptic, shorthand forms, e.g., Analyt. Biochem, Proc. Natl. Acad. Sci., J. Biol. Chem., J. Gen. Physiol., Physiol., Lond., etc. While many are easily recognizable by domain scientists, a computer program would have difficulty parsing them correctly.

Institutions: While researchers publish their research, it is often the institutions (companies or universities) that are the owners and keepers of the resulting intellectual property. An analysis based on institutional output and productivity can help depict an institution's relative strength and position in the competitive knowledge landscape. Knowledge mapping can help reveal not just the invisible college of researchers, but the "invisible college of institutions." A comparison between basic university research and applied industry invention can also foster an understanding of the progression and impact of knowledge creation.

Countries or regions: Similar to institutional analysis, it is often important to analyze publications (especially patents) based on their countries or regions (e.g., Europe vs. Asia) of origin. This kind of analysis is useful for depicting a competitive international landscape and is often relied upon for governmental research policy and funding decisions. For example, the U.S. National Nanotechnology Initiative (NNI) has performed excellent cross-regional analyses for worldwide nanotechnology research, development, and funding.

Subject and topic areas: Academics are often defined by their traditional academic boundaries in colleges or departments. However, researchers often work in several (often related) subject or topic areas. Academic publication outlets are also defined by their fields of interest and focus. While most academic journals provide a list of interested topics, some information resources are more comprehensive in their listings. For example, the USPTO provides a detailed patent classification scheme (USPC), which consists of two levels. The first level contains about 450 categories; while the second contains about 160,000 categories. In addition to these predefined subject categories, important topic-specific keywords, phrases, and concepts can be extracted from the title, abstract, and text body of an article. However, advanced Natural Language Processing (NLP) techniques are needed for such topic identification purposes.

Timeline: All scientific disciplines evolve over time. Most of the online resources for mapping scientific knowledge contain explicit publication dates. Dynamic analysis and visualization of changes in research topics and citation networks could help reveal advancements in scientific knowledge.

2.3 Analysis Techniques for Knowledge Mapping

Three types of analysis are often adopted in knowledge mapping research: *text mining, network analysis, and information visualization.*

2.3.1 Text Mining

Text mining, sometimes referred to as *text data mining*, refers generally to the process of deriving high quality information from text (according to

Wikipedia, http://en.wikipedia.org/wiki/Text_mining). Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent importation into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output (Chen and Chau, 2004). Typical text mining tasks include entity and relation extraction, text categorization, text clustering, sentiment analysis, and document summarization (Chen, 2001).

For knowledge mapping research, text mining can be used to identify critical subject and topic areas that are embedded in the title, abstract, and text body of published articles. While most structured fields (such as authors, publication outlets, dates of publication, institutions, etc.) can be parsed from online resources, extracting meanings or semantics from multimedia publications requires advanced computational techniques. Different processing algorithms are needed for different media types, e.g., text (Natural Language Processing), image (color, shape, and texture-based segmentation), audio (indexing by sound and pitch), and video (scene segmentation).

Text mining consists of two significant classes of technique: Natural Language Processing (NLP) and content analysis. In NLP, automatic indexing and information extraction techniques are effective and scalable for concept extraction. In content analysis, clustering algorithms, self-organizing map, multidimensional scaling, principal component analysis, co-word analysis, and PathFinder network are techniques often adopted for knowledge mapping analysis.

2.3.2 Network Analysis

Recent advances in *social network analysis* and *complex networks* have provided another means for studying the network of productive scholars in the invisible college.

One collection of methods that is recommended in the literature for studying networks is Social Network Analysis (SNA) (McAndrew, 1999; Sparrow, 1991; Xu and Chen, 2005a). Because SNA is designed to discover patterns of interactions between social actors in social networks, it is especially apt for co-authorship network analysis. SNA is capable of detecting sub-groups (of scholars), discovering their pattern of interactions, identifying central individuals, and uncovering network organization and structure. It has also been used to study criminal networks (Xu and Chen, 2005a, b).

Complex networks of individuals and other entities have been traditionally studied under the random graph theory (Albert and Barabasi, 2002). However, later studies suggested that real-world complex networks (such as collaboration or co-authorship networks) may not be random but may be governed by certain organizing principles. This prompted the study of real-world networks. These studies have explored the topology, evolution and growth, robustness and attack tolerance, and other properties of networks.

2.3.3 Information Visualization

The last step in the knowledge "mapping" process is to make knowledge transparent through the use of various information visualization (or mapping) techniques. *Information representation* and *user-interface interaction* are two dimensions often considered in information visualization research (Zhu and Chen, 2005).

Shneiderman (Shneiderman, 1996) proposed seven types of information representation methods including the *1D (one-dimensional), 2D, 3D, multi-dimension, tree, network,* and *temporal* approaches.

1D representation: The 1D approach represents abstract information as one-dimensional visual objects and displays them on the screen in a linear or a circular manner (Eick et al., 1992; Hearst, 1995).

2D representation: A 2D approach represents information as two-dimensional visual objects. Visualization systems based on 2D output of a self-organizing map (SOM) (Chen et al., 1996; Huang et al., 2003, 2004; Kohonen, 1995) belong to this category.

3D representation: A 3D approach represents information as three-dimensional visual objects. One example is the WebBook system (Card et al., 1996) that folds Web pages into three-dimensional books.

Multi-dimensional representation: The multi-dimensional approach represents information as multi-dimensional objects and projects them into a three-dimensional or a two-dimensional space. This approach often represents textual documents as a set of key terms that identify the theme of a textual collection. The SPIRE (Spatial Paradigm for Information Retrieval and Exploration) system presented in (Wise et al., 1995) belongs to this category.

Tree representation: The tree approach is often used to represent hierarchical relationships. The most common example is an indented text list. Other tree-based systems include the Tree-Map (Johnson and Shneiderman, 1991), the Cone Tree (Robertson et al., 1991), and the Hyperbolic Tree (Lamping et al., 1995).

Network representation: The network representation method is often applied when a simple tree structure is insufficient for representing complex relationships. Complexity may stem from citations among many academic papers (Chen and Paul, 2001; Mackinlay et al., 1995) or from interconnected Web pages on the Internet (Andrews, 1995).

Temporal representation: The temporal approach visualizes information based on temporal order. Location and animation are two commonly used visual variables to reveal the temporal aspect of information. Visual objects are usually listed along one axis according to the time when they occurred, while the other axis may be used to display the attributes of each temporal object (Eick et al., 1992; Robertson et al., 1993).

To achieve effectiveness, an effective information representation method needs to be integrated with user-interface interaction. Recent advances in hardware and software allow quick user-interface interaction, and various combinations of representation methods and user-interface interactions have been employed. Interaction between an interface and its users not only allows direct manipulation of visual objects displayed, but also allows users to select what is to be displayed and what is not (Card et al., 1999).

3. RESEARCH DESIGN

This section presents a generic framework for literature analysis of bioterrorism. The framework aims at analyzing the "invisible college" of researchers and their associated institutions, countries, and research topics. We believe the proposed framework will be useful for knowledge mapping of other scientific disciplines.

Figure 14-1 shows the research design for mapping worldwide bioterrorism research literature. The design consists of three components. The first component, data acquisition, involves gathering the bioterrorism agents/

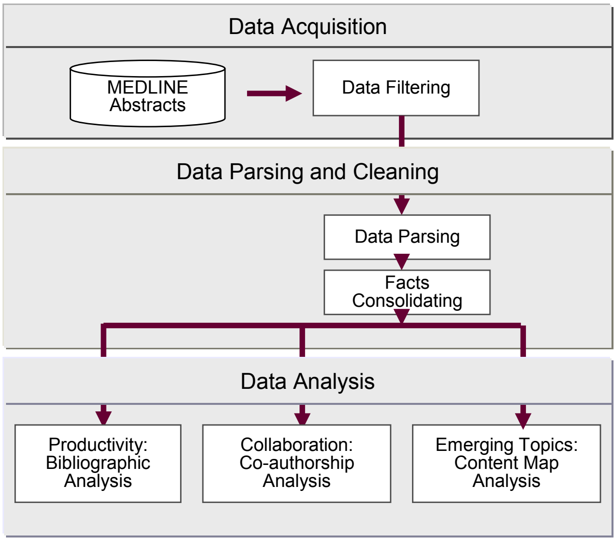

Data Acquisition
MEDLINE Abstracts
Data Filtering
Data Parsing and Cleaning
Data Parsing
Facts Consolidating
Data Analysis
Productivity: Bibliographic Analysis
Collaboration: Co-authorship Analysis
Emerging Topics: Content Map Analysis

Figure 14-1. Research design: mapping worldwide bioterrorism research literature.

diseases-related research literature from the MEDLINE database. The second component, data parsing and cleaning, contains methods to parse data into a relational database and consolidate the parsed facts. The last component, data analysis, involves identifying the productivity status, collaboration status, and emerging topics within the bioterrorism research area.

3.1 Data Acquisition

In this component, research articles were retrieved from the MEDLINE database. Compiled by the U.S. National Library of Medicine (NLM) and published on the Web by Community of Science, MEDLINE is the world's most comprehensive source of life science and biomedical bibliographic information. It contains nearly 11 million records from over 7,300 different publications from 1965 to November 16, 2005 (http://medline.cos.com/). All the related articles were collected by using keyword filtering.

3.2 Data Parsing and Cleaning

In data parsing, the title, abstract, and authors' information for each article were parsed and stored in a relational database. The institutions and countries of the authors were parsed out by using dictionaries of countries, states, cities, and institutions. All the author names of an article were parsed out, but only the first author's institution was kept for later analysis.

In facts consolidating, some variations of foreign institution names and city names were spot checked and fixed manually.

3.3 Data Analysis

In this component, we conducted three types of analysis. We used bibliographic analysis to study the productivity of authors, institutions, and countries. We also assessed the trends and evolution of bioterrorism agents/diseases research activities. We used co-authorship analysis to study collaboration between researchers. We also detected the independent or isolated research groups in the field. We used Self-Organizing-Map (SOM) to discover active research topics and identify emerging research topics in different time spans.

4. RESEARCH TESTBED

We built two sets of test data based on human- and animal-related bioterrorism agents/diseases, respectively, by retrieving related research articles from the MEDLINE database. For the human bioterrorism agents/diseases dataset, we retrieved 178,599 publication records from MEDLINE

(1964–2005) by searching article abstracts and titles using 58 keywords from the Centers for Disease Control and Prevention (CDC)'s list of agents by category (http://www.bt.cdc.gov/Agent/agentlist.asp). For the animal bioterrorism agents/diseases dataset, we retrieved 135,774 publication records from MEDLINE (1965–2005) by searching article abstracts and titles using 58 keywords from the World Organization for Animal Health (OIE)'s list of diseases by species (http://www.oie.int/eng/maladies/ en_classification.htm).

As shown in Figure 14-2, there have been an increasing number of publications in MEDLINE for both human agents/diseases research and animal agents/diseases research since 1986. Although publications in both topic areas increased rapidly, the number of publications in human agents/ diseases research is greater than the number of publications in animal agents/diseases research.

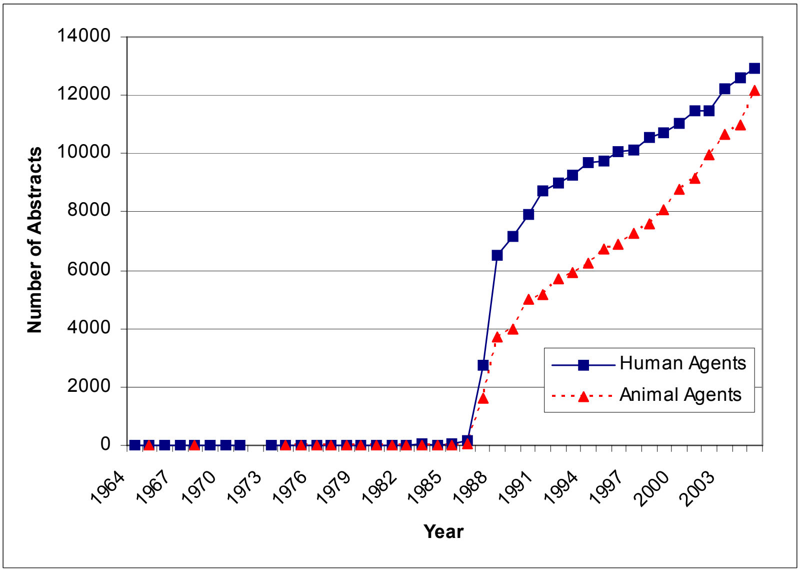


Figure 14-2. Number of publications by year.

Figures 14-3 and 14-4 show the number of publications on major human and animal agents/diseases in different years. For research related to human agents/diseases, the number of publications on Anthrax surged after 2001. Since 1987, the number of publications on Botulism has been highest among the CDC's category A agents. For research related to animal agents/diseases, the number of publications on Foot-and-mouth disease (FMD) surged after the outbreak in the UK in 2001. The number of publications on West Nile Virus (WNV) surged after an outbreak in France in 2000 and after reported

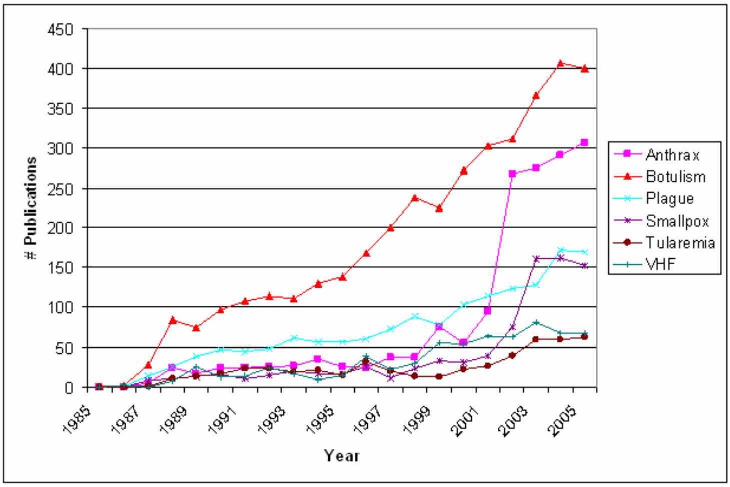


Figure 14-3. Human agents/diseases-related bioterrorism publications.

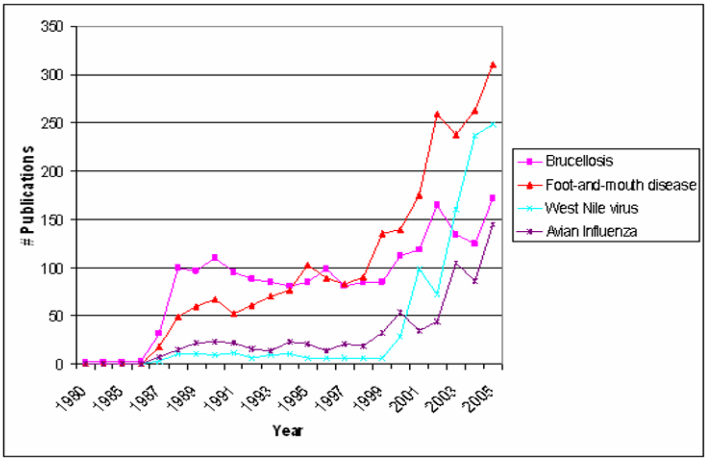


Figure 14-4. Animal agents/diseases-related bioterrorism publications.

human cases in the U.S. in 2003. The number of publications on Avian Influenza increased after an H7N2 outbreak in New York in 2003 and an H5N1 outbreak in Asia in 2004.

Tables 14-1 and 14-2 show the characteristics of the two datasets. For human agents/diseases, E. coli and Q fever, both in CDC's agents category B, had the greatest number of publications among all the agents/diseases. Botulism had the highest number of publications in category A, followed by Anthrax and Plague. There were relatively fewer publications in category C. For animal diseases/agents, most publications were about Q fever. There were also many publications on Vesicular stomatitis, Foot-and-mouth disease, and Rabies.

Table 14-1. The human bioterrorism agents/diseases dataset characteristics broken down by CDC's agents category.

Agent/disease	Number of publications	Number of unique authors	Number of unique countries
Category A	8,635	23,891	89
Botulism	3,780	9,988	56
Anthrax	1,674	5,579	54
Plague	1,504	4,169	55
Smallpox	846	2,623	43
Viral hemorrhagic fever	678	1,945	35
Tularemia	494	1,454	30
Category B[a]	170,460	356,162	157
E. coli	106,479	212,338	124
Q fever	34,312	115,136	144
Category C (Only Nipah virus and hantavirus)	919	2,974	50
Overall[b]	178,599	381,684	159

[a] Only the two most researched diseases in category B are shown
[b] Some articles mention multiple diseases

Table 14-2. The animal bioterrorism agents/diseases dataset characteristics broken down by OIE's diseases.

Agent/disease	Number of publications	Number of unique authors	Number of unique countries
Q fever	33,999	114,600	144
Vesicular stomatitis	2,374	7,281	41
Foot-and-mouth disease	2,338	7,159	63
Rabies	2,209	5,509	81
Brucellosis	1,955	5,585	77
Anthrax	1,240	4,236	50
Paratuberculosis	997	2,616	37
Japanese encephalitis	988	2,870	39
West Nile virus	944	2,086	35
Avian influenza	717	3,446	41
Overall[a]	135,774	320,630	165

[a] Only top ten diseases are shown

5. ANALYSIS RESULTS AND DISCUSSION

Significant insights were gained about the "invisible college" of researchers and their associated institutions, countries, and research topics. In this section, we discuss our analysis results for research related to human agents/diseases and research related to animal agents/diseases respectively.

5.1 Human Agents/Diseases-Related Bioterrorism Research

In the following subsections, we present the analysis results and findings on human agents/diseases-related bioterrorism research.

5.1.1 Productivity Status

Bibliographic analysis was used to identify the most productive countries, institutions, and researchers in bioterrorism research. Tables 14-3 to 14-5 list the top ten countries, institutions, and researchers with the highest numbers of human agents/diseases-related publications respectively. As shown in Table 14-3, the United States had the most publications in human agents/ diseases research, followed by Japan and the United Kingdom. At the institution level (shown in Table 14-4), Harvard University had the most publications, followed by the University of Wisconsin-Madison and Institute Pasteur-Paris. At the researcher level (shown in Table 14-5), Raoult, D., from WHO Collaborative Center for Rickettsial Reference and Research, France, had the most publications, followed by Inouye, M., from the Robert Wood Johnson Medical School in New Jersey, and Yamamoto, K. from Tohoku University in Japan. Most of these researchers who had the most publications usually performed research related to CDC's category B agents such as Q fever and E. coli.

Table 14-3. Top ten countries for human agents/diseases research.

Rank	Country	Number of publications
1	United States	65,810
2	Japan	16,023
3	United Kingdom	12,091
4	Germany	10,598
5	France	8,732
6	Canada	6,367
7	Italy	4,193
8	Sweden	3,933
9	Spain	3,847
10	India	3,589

Table 14-4. Top ten institutions for human agents/diseases research.

Rank	Institution	Number of publications
1	Harvard University[a]	1,389
2	University of Wisconsin-Madison	1,131
3	Institute Pasteur-Paris	1,125
4	University of Tokyo	883
5	Centers for Disease Control and Prevention-Atlanta	849
6	Stanford University	815
7	University of Maryland-Baltimore	813
8	Osaka University	798
9	Yale University	785
10	University of California-Davis	782

[a] Harvard University includes Harvard Medical School, the John F. Kennedy School of Government, and all other departments

Table 14-5. Top ten researchers for human agents/diseases research.

Rank	Researcher	Institution	Number of publications
1	Raoult, D.	WHO Collaborative Center for Rickettsial Reference and Research, France	220
2	Inouye, M.	Robert Wood Johnson Medical School, New Jersey	163
3	Yamamoto, K.	Tohoku University, Japan	159
4	Rowe, B.	Central Public Health Laboratory, London, UK	148
5	Peters, C.J.	University of Texas Medical Branch-Galveston	145
6	Levine, M.M.	University of Maryland-Baltimore	143
7	Dougan, G.	Imperial College London, UK	140
8	Ito, K.	Kyoto University, Japan	140
9	Kaback, H.R.	Howard Hughes Medical Institute, UCLA	136
10	Watanabe, K.	University of Tokyo, Japan	134

5.1.2 Collaboration Status

Co-authorship analysis was used to identify and visualize collaboration between researchers. We analyzed different collaboration groups based on different agents/diseases and different regions. For example, Figure 14-5 shows the collaboration status of researchers on Anthrax. The node in the network represents an individual researcher. The bigger the node, the more publications the researcher has published. The link between two researchers means that these two researchers have published one or more scientific articles together. The thicker the link, the more articles these two authors have published together. We included only researchers who published more than five articles. The largest group in the center consists of researchers from the United States. The second largest group is from France. The smaller groups are from India, Israel, Italy, and the United Kingdom. Figure 14-6 shows the collaboration status of researchers in states that sponsor terrorism for CDC's category A agents/diseases. There are six groups, all from Iran. The two largest groups with the most productive researchers are from Pasteur Institute of Iran (top) and Tehran University of Medical Sciences (bottom). Both groups focused on botulism.

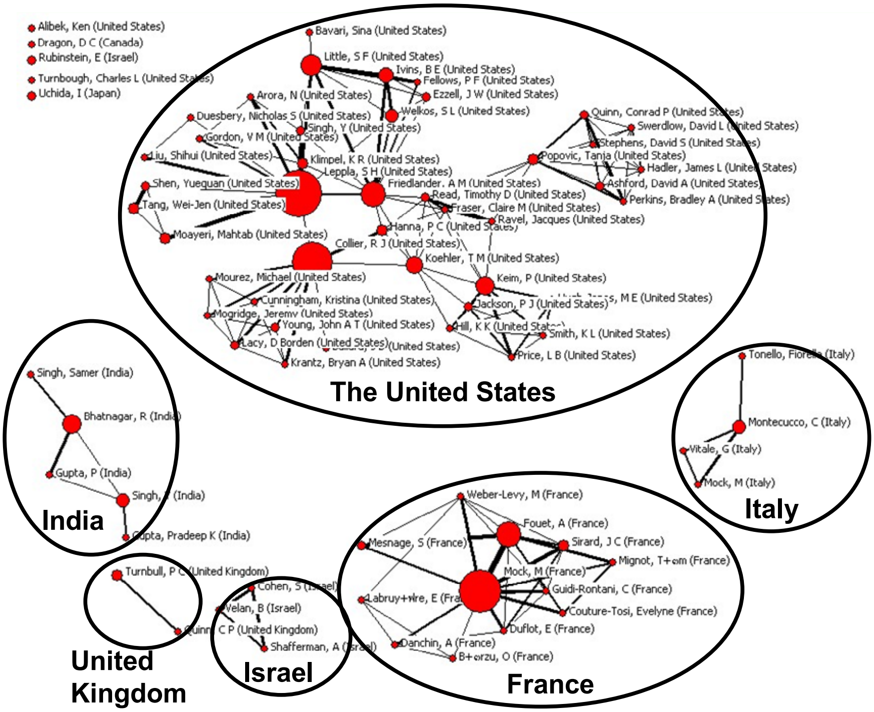

Alibek, Ken (United States)
Dragon, D C (Canada)
Rubinstein, E (Israel)
Turnbough, Charles L (United States)
Uchida, I (Japan)
Bavari, Sina (United States)
Little, S F (United States)
Ivins, B E (United States)
Fellows, P F (United States)
Ezzell, J W (United States)
Arora, N (United States)
Welkos, S L (United States)
Duesbery, Nicholas S (United States)
Singh, Y (United States)
Gordon, V M (United States)
Liu, Shihui (United States)
Klimpel, K R (United States)
Leppla, S H (United States)
Shen, Yuequan (United States)
Friedlander. A M (United States)
Read, Timothy D (United States)
Tang, Wei-Jen (United States)
Fraser, Claire M (United States)
Ravel, Jacques (United States)
Hanna, P C (United States)
Moayeri, Mahtab (United States)
Collier, R J (United States)
Koehler, T M (United States)
Quinn, Conrad P (United States)
Swerdlow, David L (United States)
Stephens, David S (United States)
Popovic, Tanja (United States)
Hadler, James L (United States)
Ashford, David A (United States)
Perkins, Bradley A (United States)
Mourez, Michael (United States)
Keim, P (United States)
s, M E (United States)
Cunningham, Kristina (United States)
Jackson, P J (United States)
Mogridge, Jeremy (United States)
Young, John A T (United States)
Hill, K K (United States)
Smith, K L (United States)
Lacy, D Borden (United States)
(United States)
Krantz, Bryan A (United States)
Price, L B (United States)
The United States
Singh, Samer (India)
Bhatnagar, R (India)
Gupta, P (India)
Singh, Y (India)
India
Gupta, Pradeep K (India)
Tonello, Fiorella (Italy)
Montecucco, C (Italy)
Vitale, G (Italy)
Mock, M (Italy)
Italy
Weber-Levy, M (France)
Fouet, A (France)
Sirard, J C (France)
Mesnage, S (France)
Mignot, T+œm (France)
Mock, M (France)
Guidi-Rontani, C (France)
Labruy+ìre, E (France)
Couture-Tosi, Evelyne (France)
Duflot, E (France)
Danchin, A (France)
B+orzu, O (France)
France
Turnbull, P C (United Kingdom)
Cohen, S (Israel)
Velan, B (Israel)
Quinn, C P (United Kingdom)
Shafferman, A (Israel)
United Kingdom
Israel

Figure 14-5. Collaboration status of researchers on Anthrax. Researchers with more than five articles are shown.

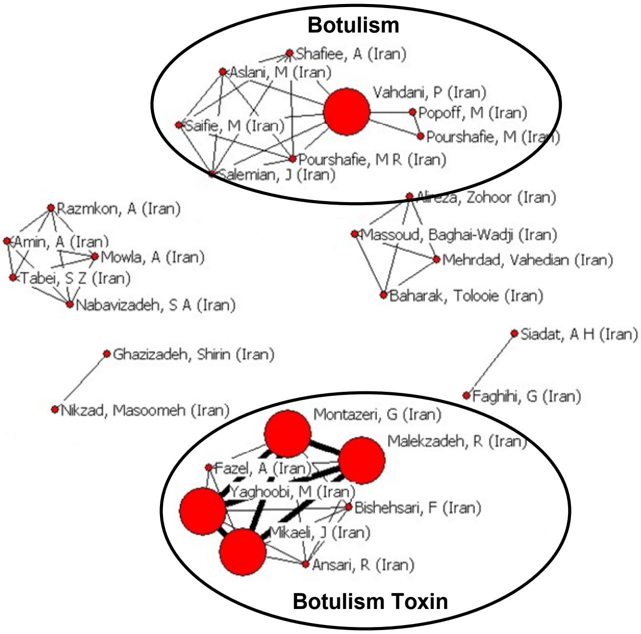


Figure 14-6. Collaboration status of researchers in states that sponsor terrorism on CDC's category A agents/diseases.

5.1.3 Emerging Topics

Content map analysis was used to identify the emerging topics and trends. Figures 14-7 and 14-8 show the evolution of major research topics in human agents/diseases literature for two time periods, 1996–2000 and 2001–2005 respectively.

The nodes in the folder tree and colored regions are topics extracted from research papers. The topics are organized by the multi-level self-organizing map algorithm. The conceptually closer technology topics (according to co-occurrence patterns) are positioned closer geographically. Numbers of papers under each topic are presented after the topic labels. The sizes of the topic regions also correspond to the number of documents assigned to the topics. Region color indicates the growth rate of the associated topic: the warmer the color, the higher the growth rate. The growth rate is defined as the number of articles published in the previous time period/the number of articles published in the following time period for a particular topic (region).

It can be observed that dominating topic regions during 1996–2000 are: “Botulinum toxin type,” “Francisella tularensis,” “Clostridium botulinum,” “Effect of Botulinum toxin,” “V Antigens,” and “Ebola viruses.” The sizes of these topic regions suggest that they were the key technology topics during the 5 years preceding 2000. Among these dominating topics, “Botulinum toxin type,” “Effect of Botulinum toxin,” “Ebola viruses,” and “V Antigens” are emerging topics. During 2001–2005, dominating topics are: “Yersinia pestis,” “Centers for Disease Control,” “Protective antigens,” “Francisella tularensis,” and “Botulinum neurotoxin.” The new important topics are: “Biological weapons,” “Anthracis spores,” and “Smallpox vaccination.” We can see a shift in research interest towards the use of Anthrax spores and biological weapons after 2000.

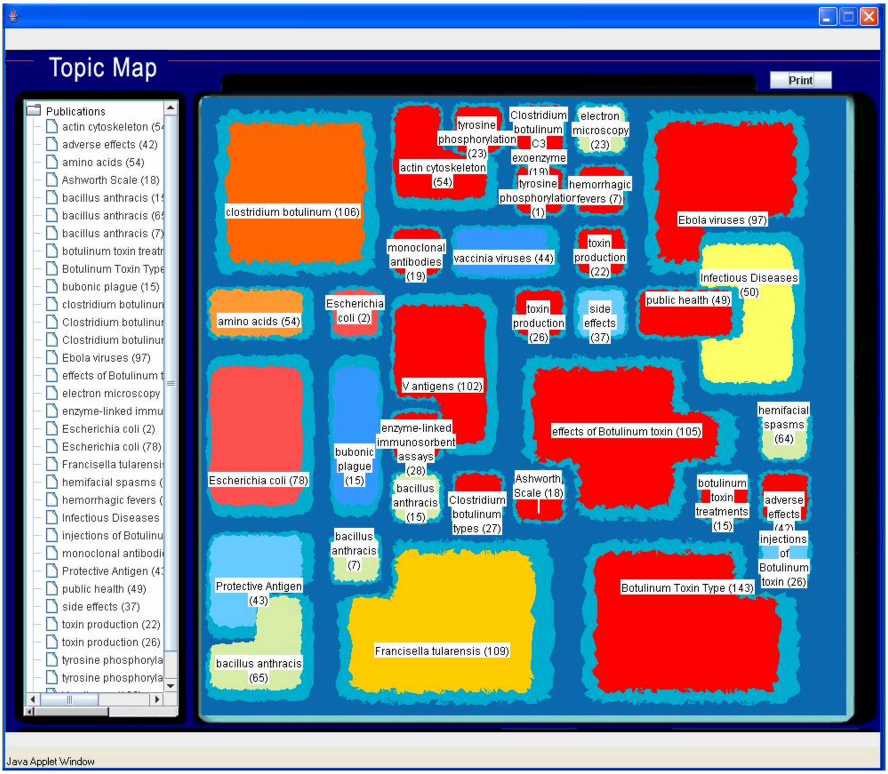


Figure 14-7. Content map for human agents/diseases literature (1996–2000).

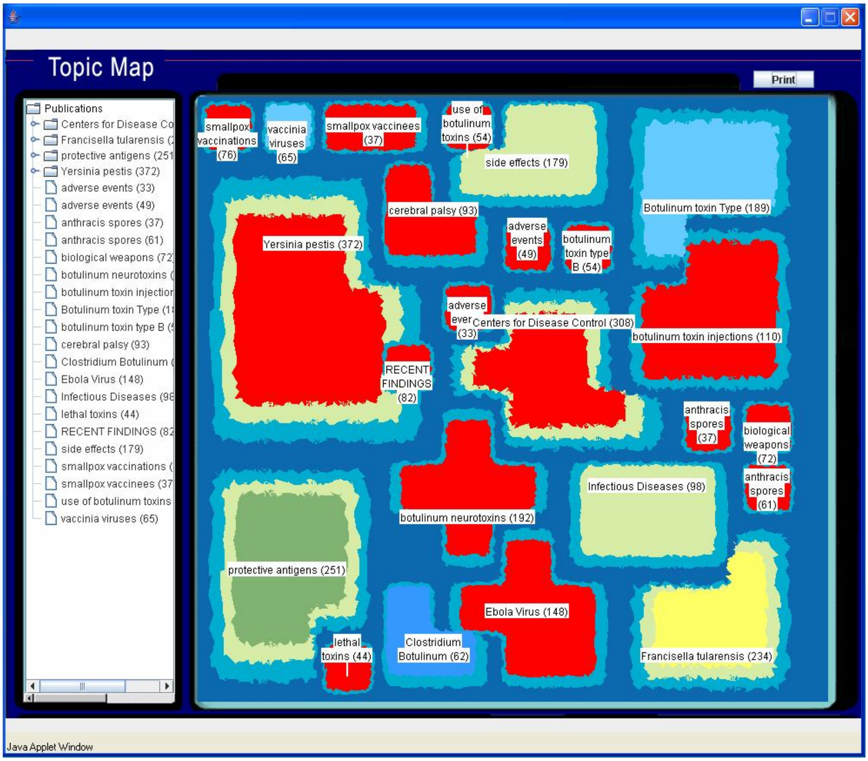


Figure 14-8. Content map for human agents/diseases literature (2001–2005).

5.2 Animal Agents/Diseases-Related Bioterrorism Research

5.2.1 Productivity Status

Tables 14-6 to 14-8 list the top ten countries, institutions, and researchers with the most numbers of animal agents/diseases-related publications respectively. As shown in Table 14-6, the United States had the most publications in animal agents/diseases research, followed by Japan and the United Kingdom. At the institution level (shown in Table 14-7), CDC-Atlanta had the most publications, almost twice as many as National Taiwan University and Institute Pasteur-Paris in the second and third rank. At the researcher level (shown in Table 14-8), Chen, D. S. from National Taiwan University had the most publications, followed by Williams, R. from King's College Hospital, London, UK, and Raoult, D. from WHO Collaborative Center for Rickettsial Reference and Research, France.

Table 14-6. Top ten countries for animal agents/diseases research.

Rank	Country	Number of publications
1	United States	39,901
2	Japan	10,392
3	United Kingdom	9,369
4	France	6,115
5	Italy	5,970
6	Germany	5,269
7	India	4,308
8	Spain	3,695
9	Canada	3,568
10	Taiwan	3,146

Table 14-7. Top ten institutions for animal agents/diseases research.

Rank	Institution	Number of publications
1	Centers for Disease Control and Prevention-Atlanta	1,300
2	National Taiwan University	685
3	Institute Pasteur-Paris	638
4	University of California, San Francisco	602
5	University of California-Davis	551
6	University of Pittsburgh	529
7	Mayo Clinic-Rochester	521
8	University of Southern California	500
9	Mahidol University (Thailand)	479
10	U.S. Department of Agriculture-Agricultural Research Service	458

Table 14-8. Top ten researchers for animal agents/diseases research.

Rank	Researcher	Institution	Number of publications
1	Chen, D.S.	National Taiwan University	209
2	Williams, R.	King’s College Hospital, London, UK	203
3	Raoult, D.	WHO Collaborative Center for Rickettsial Reference and Research, France	198
4	Lee, S.D.	Veterans General Hospital, Taipei, Taiwan	163
5	Liaw, Y.F.	Chang Gung Memorial Hospital, Taiwan	159
6	Hayashi, N.	Osaka University, Japan	151
7	Okamoto, H.	Jichi Medical School, Japan	142
8	Carreno, V.	Viral Hepatitis Research Foundation, Madrid, Spain	139
9	Prusiner, S.B.	University of California, San Francisco	137
10	Purcell, R.H.	National Institute of Allergy and Infectious Diseases, Maryland	130

5.2.2 Collaboration Status

We analyzed different collaboration groups for animal agents/diseases research. For example, Figure 14-9 shows the collaboration status of researchers on West Nile Virus. We only included researchers who published more than five articles. The largest group in the center consists of researchers from the United States.

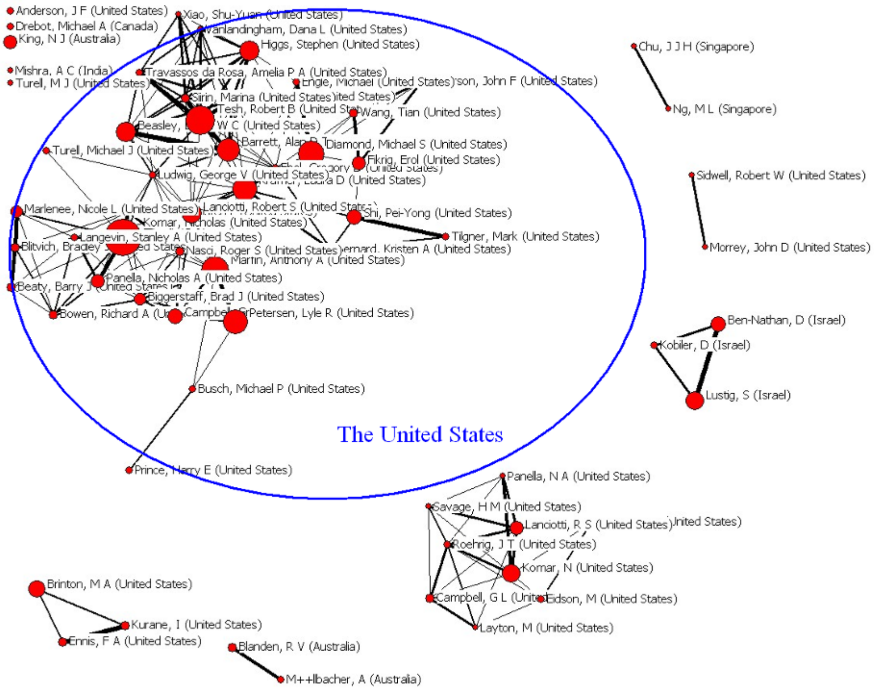

Anderson, J F (United States)
Drebot, Michael A (Canada)
King, N J (Australia)
Mishra, A C (India)
Turell, M J (United States)
Xiao, Shu-Yuan (United States)
Vanlandingham, Dana L (United States)
Higgs, Stephen (United States)
Travassos da Rosa, Amelia P A (United States)
Engle, Michael (United States)
son, John F (United States)
Sirin, Marina (United States)
ited States)
Tesh, Robert B (United Sta
Wang, Tian (United States)
Beasley,
W C (United States)
Barrett, Alan D T
Diamond, Michael S (United States)
Turell, Michael J (United States)
Fikrig, Erol (United States)
Ludwig, George V (United States)
(United States)
a D (United States)
Marlenee, Nicole L (United States)
Lanciotti, Robert S (United States)
Shi, Pei-Yong (United States)
Komar, Nicholas (United States)
Langevin, Stanley A (United States)
Tilgner, Mark (United States)
Blitvich, Bradley
ed States
Nasci, Roger S (United States)
ernard, Kristen A (United States)
Martin, Anthony A (United States)
Panella, Nicholas A (United States)
Beaty, Barry J (United States)
Biggerstaff, Brad J (United States)
Bowen, Richard A (Un
Campbe
Petersen, Lyle R (United States)
Busch, Michael P (United States)
The United States
Prince, Harry E (United States)
Chu, J J H (Singapore)
Ng, M L (Singapore)
Sidwell, Robert W (United States)
Morrey, John D (United States)
Ben-Nathan, D (Israel)
Kobiler, D (Israel)
Lustig, S (Israel)
Panella, N A (United States)
Savage, H M (United States)
Lanciotti, R S (United States)
Jnited States)
Roehrig, J T (United States)
Komar, N (United States)
Campbell, G L (Unit
Eidson, M (United States)
Layton, M (United States)
Brinton, M A (United States)
Kurane, I (United States)
Ennis, F A (United States)
Blanden, R V (Australia)
M++llbacher, A (Australia)

Figure 14-9. Collaboration status of researchers on West Nile Virus. Researchers with more than five articles are shown.

5.2.3 Emerging Topics

We analyzed the emerging research topics based on different animal agents/diseases. Foot-and-mouth Disease (FMD) is one of the most devastating diseases of farm animals. It occurs throughout the world and is a significant hazard to agriculture. The 2001 epidemic in the UK led to the loss of six million livestock. Here we use FMD as an example of the analysis of emerging topics. Figures 14-10 and 14-11 show the evolution of major research topics related to FMD for two time periods, 1996–2000 and 2001–2005 respectively.

It can be observed that dominating topic regions during 1996–2000 are: “High resolution ultrasound,” “Brachial artery diameter,” “Mean age,” “Risk factors,” “Animal products,” “Endothelium-dependent vasodilation,” and “Oxidative stress.” The sizes of these topic regions suggest that they were the key technology topics during the 5 years preceding 2000. During 2001–2005, dominating topics are: “Fibromuscular dysplasia,” “Foot-and-mouth disease virus,” “Outbreak of Foot-and-mouth disease,” “Reactive Hyperemia,” and “United Kingdom.” These dominating topics are also the emerging ones. These topics are due to the outbreak of FMD in the UK in 2001.

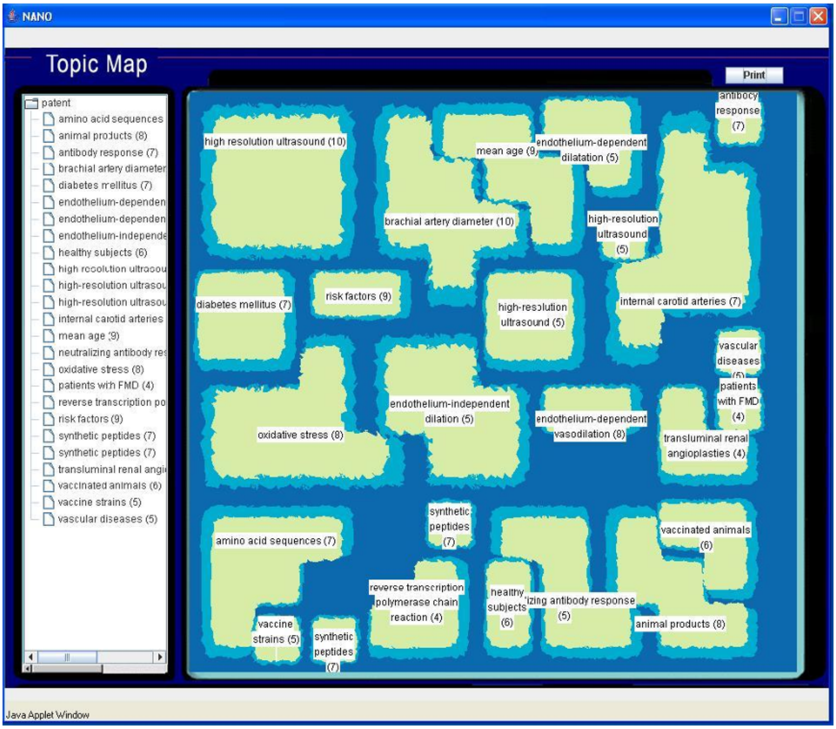


Figure 14-10. Content map for FMD-related literature (1996–2000).

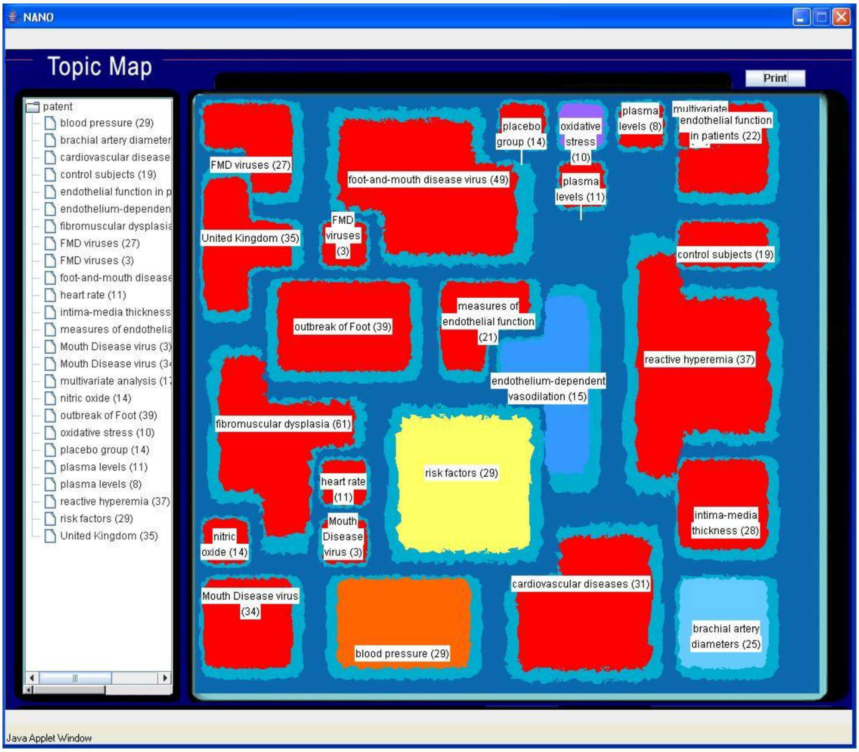


Figure 14-11. Content map for FMD-related literature (2001–2005).

6. CONCLUSION

Knowledge Mapping, aimed at the processes of charting, mining, analyzing, sorting, enabling navigation of, and displaying knowledge, helps reveal the interconnected, invisible college or network of scholars and their seminal publications and ideas. In this chapter, we discussed different types of online resources often used for Knowledge Mapping, the various units of analysis for Knowledge Mapping, and the three major types of Knowledge Mapping analysis techniques: Text Mining, Network Analysis, and Information Visualization. We applied these techniques to mapping worldwide bioterrorism literature, and studied the productivity of researchers, institutions, and countries, the collaborations among researchers, and the emerging research topics and trends.

ACKNOWLEDGEMENTS

Special thanks to all the members of the University of Arizona Artificial Intelligence Lab for their support and assistance of this study. This work is supported by the NSF Information Technology Research (ITR) Program, "BioPortal: A National Center of Excellence for Infectious Disease Informatics," (IIS-0428241), August 2004–July 2009.

QUESTIONS FOR DISCUSSION

1. What are other additional informal online resources, such as Web sites, forums, chat rooms, blogs, multimedia sites, and/or social networking sites, which could be used for Knowledge Mapping?
2. How can non-text information produced by scholars, such as images, photos, audios, and videos, be analyzed to reveal knowledge generated in a scientific discipline?
3. How can we develop an interactive, user-controlled, highly-visualized system for knowledge mapping using different online resources?
4. What are the most important countries, institutions and researchers in the bioterrorism domain in the past decade?
5. What are some of the hottest bioterrorism topics or research areas in the past 5 years?

REFERENCES

Albert, R., and Barabasi, A.-L. (2002). "Statistical mechanics of complex networks," *Reviews of Modern Physics, 74*(1), 47–97.

Andrews, K. (1995). "Visualizing cyberspace: Information visualization in the Harmony Internet browser," *Proceedings of IEEE Symposium on Information Visualization (InfoVis'95)*, 97–104.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann.

Card, S. K., Robertson, G. G., and York, W. (1996). "The WebBook and the WebForager: An information workspace for the World Wide Web," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'96)*, 111–117.

Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. New York: Springer Verlag.

Chen, C., and Paul, R. J. (2001). "Visualizing a knowledge domain's intellectual structure," *Computer, 34*(3), 65–71.

Chen, H. (2001). *Knowledge Management Systems: A Text Mining Perspective*. Tucson, Arizona: University of Arizona.

Chen, H., and Chau, M. (2004). "Web mining: Machine learning for web applications," *Annual Review of Information Science and Technology (ARIST), 38*, 289–329.

Chen, H., Schuffels, C., and Orwig, R. (1996). "Internet categorization and search: A self-organizing approach," *Journal of Visual Communication and Image Representation, 7*(1), 88–102.

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: The University of Chicago Press.

Eick, S. G., Steffen, J. L., and Sumner, E. E. (1992). "Seesoft: A tool for visualizing line-oriented software," *IEEE Transactions on Software Engineering, 18*(11), 11–18.

Hearst, M. (1995). "TileBars: Visualization of term distribution information in full text information access," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 59–66.

Huang, Z., Chen, H., Yip, A., Ng, G., Guo, F., Chen, Z.-K., et al. (2003). "Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field," *Journal of Nanoparticle Research, 5*, 333–363.

Huang, Z., Chung, W., and Chen, H. (2004). "Graph model for E-commerce recommender systems," *Journal of the Amercian Society for Information Science and Technology (JASIST), 55*(3), 259–274.

Johnson, B., and Shneiderman, B. (1991). "Tree-maps: A space-filling approach to the visualization of hierarchical information structures," *Proceedings of IEEE Visualization'91 Conference*, 284–291.

Kohonen, T. (1995). *Self-organizing Maps*. Berlin: Springer-Verlag.

Lamping, J., Rao, R., and Pirolli, P. (1995). "A focus + context technique based on hyperbolic geometry for visualizing large hierarchies," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 401–408.

Lane, H. C., LaMontagne, J., and Fauci, A. S. (2001). "Bioterrorism: A clear and present danger," *Nature Medicine, 7*(12), 1271–1273.

Mackinlay, J. D., Rao, R., and Card, S. K. (1995). "An organic user interface for searching citation links," *Proceedings of CHI'95, ACM Conference on Human Factors in Computing Systems, New York*, 67–73.

McAndrew, D. (1999). *The Structural Analysis of Criminal Networks*. Paper presented at The Social Psychology of Crime: Groups, Teams, and Networks, Offender Profiling Series, Aldershot: Dartmouth.

Narin, F. (1994). "Patent bibliometrics," *Scientometrics, 30*(1), 147–155.

Robertson, G. G., Card, S. K., and Mackinlay, J. D. (1993). "Information visualization using 3D interactive animation," *Communications of the ACM, 36*(4), 56–71.

Robertson, G. G., Mackinlay, J. D., and Card, S. K. (1991). "Cone Trees: Animated 3D visualizations of hierarchical information," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 189–194.

Shiffrin, R. M., and Börner, K. (2004). "Mapping knowledge domains," *Proceedings of the National Academy of Sciences of the United States of America, 101*, 5183–5185.

Shneiderman, B. (1996). "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," *Proceedings of the IEEE Symposium on Visual Languages*, Washington.

Sparrow, M. K. (1991). "The application of network analysis to criminal intelligence: An assessment of the prospects," *Social Networks, 13*, 251–274.

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., et al. (1995). “Visualizing the non-visual: Spatial analysis and interaction with information from text documents,” *Proceedings of InfoVis’95, IEEE Symposium on Information Visualization*, 51–58.

Xu, J., and Chen, H. (2005a). “CrimeNet explorer: A framework for criminal network knowledge discovery,” *ACM Transactions on Information Systems, 23*(2), 201–226.

Xu, J. J., and Chen, H. (2005b). “Criminal network analysis and visualization,” *Communications of the ACM, 48*(6), 101–107.

Zhu, B., and Chen, H. (2005). “Information visualization,” *Annual Review of Information Science and Technology (ARIST), 39*, 139–178.

SUGGESTED READING

Chen, H., and Roco, M. (2008). *Mapping Nanotechnology Innovations and Knowledge: Global, Longitudinal Patent and Literature Analysis*. New York: Springer.

This book is about Mapping Nanotechnology Innovations and Knowledge. It shows the systematic and automated knowledge mapping methodology to collect, analyze and report nanotechnology research on a global basis. The result of these analyses is a systematic presentation of the state of the art of nanotechnology, which includes basic analysis, content analysis, and citation network analysis of comprehensive nanotechnology findings across technology domains, inventors, institutions, and countries.

Chen, C. (2003). *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. New York: Springer Verlag.

This book examines the history and the latest developments in the quest for knowledge visualization from an interdisciplinary perspective, ranging from theories of invisible colleges and competing paradigms, to practical applications of visualization techniques for capturing intellectual structures, and the rise and fall of scientific paradigms. Containing simple and easy to follow diagrams for modeling and visualization procedures, as well as detailed case studies and real-world examples, this is a valuable reference source for researchers and practitioners.

ONLINE RESOURCES

Various online resources are available for mapping scientific knowledge. Abstracts and Indexes:

The primary databases generated by the National Library of Medicine (such as MEDLINE or TOXLINE) (http://www.nlm.nih.gov/)

Commercial full-text journal articles and digital libraries:

Web of Science (http://scientific.thomson.com/products/wos/)
The ACM Digital Library (http://portal.acm.org/dl.cfm)
The IEEE Computer Society Digital Library (http://www.computer.org/portal/site/csdl/index.jsp)

Free full-text articles and e-prints:

The Free Medical Journals site (http://www.freemedicaljournals.com/)
HighWire Press (http://highwire.stanford.edu/lists/freeart.dtl)
arXiv.org service (http://arxiv.org/)

Citation indexing systems and services:

The Science Citation Index (http://scientific.thomson.com/products/sci/)
Google Scholar (http://scholar.google.com/intl/en/scholar/)
CiteSeer (http://citeseer.ist.psu.edu/citeseer.html)

Electronic Theses and Dissertations (ETD):

ProQuest system (http://il.proquest.com/brand/umi.shtml)
The Networked Digital Library of Theses and Dissertations (NDLTD, http://www.ndltd.org/)

Patents:

United States Patent and Trademark Office (USPTO, http://www.uspto.gov/)
European Patent Office (EPO, http://www.european-patent-office.org/index.en.php)
Japan Patent Office (JPO, http://www.jpo.go.jp/)

Business and industry articles and reports:

Forrester (http://www.forrester.com)
IDC (http://www.idc.com)
Gartner (http://www.gartner.com)

Besides the above resources, there are also a lot of Web sites, forums, chat rooms, blogs, multimedia sites, social networking sites, and virtual worlds that can be used for mapping scientific knowledge.

Chapter 15

SOCIAL NETWORK ANALYSIS FOR CONTACT TRACING

YI-DA CHEN[1,*], HSINCHUN CHEN[1], and CHWAN-CHUEN KING[2]

CHAPTER OVERVIEW

Contact tracing is an important control measure in the fight against infectious disease. Healthcare workers deduce potential disease pathways and propose corresponding containment strategies from collecting and reviewing patients' contact history. Social Network Analysis (SNA) provides healthcare workers with a network approach for integrating and analyzing all collected contact records via a simple network graph, called a contact network. Through SNA, they are able to identify prominent individuals in disease pathways as well as study the dynamics of disease transmission. In this chapter, we review the role of SNA in supplementing contact tracing and present a case study of the Taiwan SARS outbreak in 2003 to demonstrate the usefulness of geographical contacts in disease investigation.

Keywords: Contact tracing, Social network analysis, Core group identification, Dynamics of disease transmission, SARS

[1*] *Artificial Intelligence Lab, Department of Management Information Systems, Eller College of Management, The University of Arizona, Tucson, AZ 85721, USA,* *ydchenb@email.arizona.edu.*

[2] *Graduate Institute of Epidemiology,* College of Public Health *National Taiwan University, 17 Xu-Zhou Road, Taipei (100), Taiwan*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_15,

1. INTRODUCTION

Contact tracing is a public health tool used in the fight against infectious disease, and is based on the assumption that disease is transmitted via close personal contact. From patients' contact history, healthcare workers attempt to break the chain of transmission by first tracing the source of infection and then identifying other potential patients exposed to the disease so that they may be monitored and, if necessary, treated (Eames and Keeling, 2003; Rothenberg et al., 2003). Since contact tracing requires intensive manual effort in interviewing patients and collecting their contact records, contact tracing is most effective when the number of infected cases or reproductive ratio of the disease is low (Eames and Keeling, 2003). Contact tracing has been applied to the control of Sexually Transmitted Diseases (STDs), Tuberculosis (TB), and some newly emerging diseases, such as Severe Acute Respiratory Syndrome (SARS) in 2003.

The effect of social networks on STD transmission has long been recognized and has triggered the development of control measures for STDs. In the 1960s, for instance, Dr. Havlak suggested that if several syphilis patients share a common sexual contact, their contact tracing records should be kept in one folder and analyzed as a unit or lot (Rothenberg et al., 2003). This "lot system" has facilitated the identification of potential STD patients for target screening, and its basic premise is similar to the concept of clusters in Social Network Analysis (SNA) (Rothenberg and Narramore, 1996). However, the consideration of using SNA to enhance contact tracing wasn't begun until the emergence of Acquired Immunodeficiency Syndrome (AIDS) in the 1980s; its rapid spread was believed to be related to fast growing sexual networks augmented by the ease of long distance travel. In 1984, Auerbach et al. (1984) initiated a contact investigation of 19 patients in California to assess the role of sexual relationships in AIDS transmission. They eventually linked 40 patients across ten cities in the USA in a network graph and supported the long held hypothesis that AIDS is transmitted via pathogens.

In 1985, Klovdahl (1985) formally established the connection between contact tracing and SNA, using the same dataset from the Auerbach et al. study to demonstrate how SNA could be applied to examine two causal criteria of transmission: exposure and temporality. In addition, he recapped the relationship between an STD's spread and the structure of social networks, and he introduced the potential usage of centrality measures in SNA to identify prominent individuals in STD transmission. In 1994, Klovdahl et al. (1994) further proved the concept of incorporating SNA into disease investigation with a large scale study in Colorado Springs, Colorado, in which over 600 individuals were directly or indirectly connected to each other in one network.

For more than 20 years following Klovdahl's 1985 paper, SNA has been successfully applied to the studies of several STD outbreaks. The epidemiological insights that SNA can provide have also evolved from the static identification of core groups to the investigation of transmission dynamics. In this chapter, we review the development of SNA in the field of epidemiology and present a case study of the Taiwan SARS outbreak in 2003 to discuss the role of geographical contacts in disease investigation.

The remainder of this chapter is organized as follows. We first review two important SNA tools for contact tracing: network visualization and measures. Then we discuss how SNA is applied in order to identify prominent individuals in disease pathways and study the dynamics of disease transmission. Finally, we present the case study and conclusions.

2. NETWORK VISUALIZATION AND MEASURES IN SNA

In any society, individuals develop their relationships with others and form their own personal networks through social activities. From these networks, they may seek advice for important decisions, obtain resources useful for their jobs, and create alliances for supporting their beliefs. Based on the observation of how individuals act in a society, instead of supporting the idea that people are autonomous, SNA proposes that people's behavior is better explained by seeing them as embedded in a network of relationships. By reconstructing a social network, SNA researchers seek to understand people's behavior and organizational structures from their linkages with each other.

In SNA, the relationship of individuals is described as a socio-matrix (Scott, 2000; Wasserman and Faust, 1994). It creates a one-to-one mapping between participants, and each cell indicates whether a relationship exists between its row and column persons (1 for existence and 0 otherwise). A socio-matrix can also be visualized as a socio-gram or social network in which individuals are symbolized as nodes and connected to each other with edges or ties for their relationships. Figure 15-1 shows a sample friendship network of ten individuals. In this network, Persons A and G are considered the most active or "popular" persons since they are linked to the largest number of people. Person F is also important although he/she doesn't have as many connections as Persons A and G: Person F bridges two different groups of friends. Without Person F, these two groups of people may not have the chance to establish relationships with each other in the future. In SNA, these three people are said to be central or prominent within the sample network.

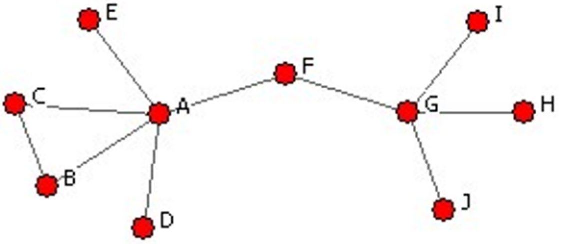


Figure 15-1. A sample friendship network of ten individuals.

Centrality measures are quantitative indicators for finding those “central” individuals from a network, originally developed in communication scenarios. From a topological perspective, people who are able to receive or control the mainstream of message flow typically stand in a position similar to the central point of a star (Freeman, 1978/79), such as the location of Person A in the network above. Various centrality measures, such as degree and betweenness, can be employed to determine the importance of a node within a network. For example, the degree is the number of edges that a node has. Since the central point of a star has the largest number of edges connecting it to the other nodes, a node with a higher degree is topologically considered to be more central to its network (Freeman, 1978/79; Wasserman and Faust, 1994). The betweenness measures “the extent to which a particular node lies between the various other nodes” (Scott, 2000) because the central point also sits between pairs. The higher betweenness a node has, the more potential it has to be a gatekeeper controlling the connections (such as communications) between the others (Scott, 2000). Table 15-1 lists the degree and betweenness of nodes in our sample friendship network. From this table we can see how these measures can reveal the prominence of people in a network.

The centrality measures are categorized as micro-level measures and focus on the status of individual nodes in a social network. In contrast, macro-level measures reflect a network’s overall structure and are usually used for network-to-network comparison, such as the number of components and network density. A component in graph theory is defined as a maximal-connected sub-graph. Two nodes belong to the same connected component if they are connected directly with an edge or indirectly through other nodes. The number of components consequently shows the number of connected sub-graphs and reflects the degree to which people are grouped in a network (Scott, 2000). The number of components in our sample friendship network is 1. If we remove Person F from the network, its number of components would become 2. Network density is calculated with the proportion of existing edges to the maximum possible edges among nodes (Wasserman and Faust, 1994). If two social networks have the same number of nodes, the network

Table 15-1. Degree and betweenness of nodes in the sample friendship network.

Node	Degree	Betweenness
A	5	25
G	4	21
F	2	20
B, C	2	0
D, E, H, I, J	1	0

density can differentiate their interaction intensity. According to combinatorics, the maximum possible edges of our sample network totals (10 × 9)/2 = 45. Its existing edges are 10. Therefore, its network density is 10/45 = 0.2222. The frequently used macro- and micro-level measures are summarized in Table 15-2. It is noted that in some occasions the average value of a micro-level measure can also serve as a macro-level measure. For example, the average degree of nodes can also indicate network participants' interaction intensity and replace the network density in usage.

Table 15-2. Summary of frequently used network measures.

Measure	Type	Description
Degree	Micro	The total number of other nodes adjacent to a node
Betweenness	Micro	The degree to which a node lies between various other nodes
Closeness	Micro	The degree to which a node is close to the other nodes
Information Centrality	Micro	The extent to which the information flowing in all paths comes from a specific node
Number of Components	Macro	The number of connected sub-graphs in a network
Density	Macro	The proportion of existing edges to the maximum possible edges
Number of N-Clique	Macro	The number of maximal sub-graphs in which any two nodes have a geodesic distance no greater than N

3. SNA IN EPIDEMIOLOGY

When applied to epidemiology, a social network is called a contact network. It represents accumulated linkages among patients with their potential contacts of infection in a period of time. Therefore, unlike the actual route of transmission which is a one-to-one mapping between patients for their infection, a contact network typically depicts a many-to-many relationship. From a contact network, disease investigators can visualize the potential scenarios or social factors that triggered an outbreak and propose corresponding containment strategies to control it.

3.1 Static Analysis of Linkage in a Contact Network for STDs

The main strength of SNA in disease analysis is its ability, through centrality measures, to identify key individuals in an outbreak. For STDs, those key individuals are referred to as the core group and bridges (Thomas and Tucker, 1996; Wasserheit and Aral, 1996). The concept of a core group was introduced by Yorke et al. (1978) in the 1970s and postulates that epidemics or endemics of an STD are maintained by a small group of sexually active individuals who persistently infect other healthy people. Because of their active sexual life, those core group members inevitably behave like the central point of a star connecting to a large number of others in a contact network and exhibit high values in the degree and betweenness measures. However, the wide spread of an STD requires individuals who, acting as bridges, transfer the disease from one subpopulation to another (Rothenberg and Narramore, 1996; Wasserheit and Aral, 1996). These bridge people may not have many sexual partners, but accidentally channel the disease to a different class of subpopulation (e.g., different economic class) via their purchase of sexual services. Therefore, they may exhibit low degree values but high betweenness.

In epidemiology, the central questions in SNA studies usually surround which *group* rather than which *person* facilitates a disease's spread. Therefore, investigators need to categorize patients into several groups according to their demographic characteristics and then calculate the average values of centrality measures for each group. In the Colorado Springs study, Rothenberg et al. (1995) estimated the relationship of centrality rankings with the perceived risk of AIDS and categorized the behaviors of their participants into six categories: prostitutes, paying and nonpaying partners, injection drug users and their partners, and other. They reported that prostitutes and nonpaying partners who ranked highest in information centrality were more likely to engage in high-risk sexual activities, such as anal sex, and know someone with AIDS. In a separate study of a syphilis outbreak, Rothenberg et al. (1998b) found that people with syphilis were more central within the outbreak network based on their significantly higher betweenness. From the network visualization, they further uncovered that a group of young girls served as the core group of the outbreak by connecting two different ethnic groups of men.

3.2 Transmission Dynamics of STDs

A contact network is analogous to a snapshot which captures the process of disease distribution within a given period of time. Comparing a series of contact networks with macro-measures enables the study of transmission dynamics by examining the change in transmission patterns over time. In the literature, there are two major perspectives in studying transmission dynamics with SNA: risky behavior and epidemic phases. In 1998, Rothenberg et al. (1998a) presented results from a longitudinal study in Colorado Springs as an example of the risky behavior perspective. Ninety-six AIDS patients were repeatedly interviewed for 3 years about their contacts with others, including sexual contact, drug use, and needle sharing. For each type of contact, the researchers constructed three serial contact networks at 1-year intervals and compared the structure of those serial networks to assess network stability and changes in risky behavior. According to the study results, one group of patients showed a significant decrease in needle sharing based on the gradually smaller average degree and size of components in the group's contact networks.

The dynamic topology of transmission proposed by Wasserheit and Aral (1996) provides a theoretical ground for using SNA to identify the epidemic phases of STDs. Wasserheit and Aral extended the core group theory and suggested that STD transmission is determined not only by the change rate of sexual partners but also by interaction with healthcare programs. According to their dynamic topology as shown in Figure 15-2, in an early phase of transmission or a growth phase, an STD must first enter a sexual network in which the change rate of sex partners is high enough to allow the STD to establish itself and grow within a subpopulation. With a consistent increase of infected individuals, the disease eventually expands to other subpopulations via bridges: people who have sexual contact with more than one subpopulation.

When the STD starts to spread simultaneously in various subpopulations, this is described as a hyperendemic phase. At this point, healthcare workers would begin to notice the disease, initiate an investigation, and develop intervention programs and curative therapies. If these measures were effective, the number of incidents would gradually decrease, thereby transitioning to a decline phase. The STD eventually would arrive at an endemic phase and

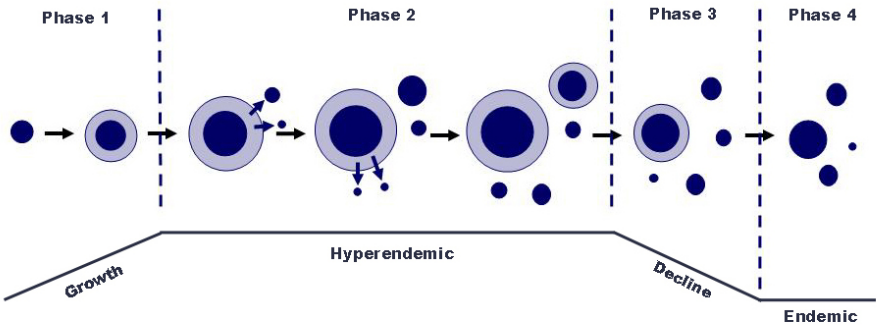


Figure 15-2. Wasserheit and Aral's dynamic topology adapted from (Wasserheit and Aral, 1996).

reside in a marginalized subpopulation where the number of sexual partners may be high but contact with healthcare systems is restricted or minimal (Wasserheit and Aral, 1996).

According to Wasserheit and Aral's topology, Potterat et al. (2002a) suggested that the structure of sexual contact networks is more accurate than secular trend data for indicating epidemic phases. To prove their concept, they constructed a sexual contact network of chlamydial patients in Colorado Springs from 1996 to 1999. They found that while the number of reported cases increased by 55% during this period of time, the network was relatively fragmented and lacked cyclic structures in comparison with an outbreak contact network. These circumstances indicated that the chlamydial transmission was in either a stable or a declining phase. Cunningham et al. (2004) further examined the structural characteristics of a contact network associated with epidemic phases. They compared the structures of two contact networks which respectively represented the periods during and after an epidemic. They reported that after the epidemic, the overall network centrality declined but the component density increased. This finding is consistent with Wasserheit and Aral's topology that in the decline phase the disease would be restrained in sexual networks that have intensive sexual exchange but limited access to the healthcare system.

3.3 From STDs to Tuberculosis

Before the year 2000, SNA studies for disease outbreaks all emerged from the study of STDs. One reason for this could have been the availability of contact tracing data. Compared to other infectious diseases, such as influenza, STDs are heavily dependent on personal connections for transmission and hence can be controlled by contact tracing and taking appropriate intervention actions. Another reason may be related to the capability of network presentation. Since SNA was originally developed to study social phenomena via person-to-person linkage, its network presentation is inherently used to portray the relationships between people and contains only individuals as actors in the network graph. This kind of presentation may be sufficient for STDs but is not sophisticated enough to describe the scenarios of indirect-contact or airborne transmission. Klovdahl et al. (2001) addressed this limitation with their investigation of a tuberculosis (TB) outbreak in Houston, Texas. They first used the conventional presentation of SNA and constructed a person-to-person contact network to analyze the outbreak. However, only 12 personal contacts were identified among the 29 patients. Through further collaboration with local healthcare workers, they found that geographical contact was more important than personal contact in understanding the outbreak. By including places such as bars and restaurants in their contact network, they were finally able to connect those 29 patients directly or indirectly in a network (Klovdahl et al., 2001).

Since then, several outbreak studies have adopted the same approach of incorporating geographical contacts into SNA (Abernethy, 2005; Andre et al., 2007; De et al., 2004; McElroy et al., 2003). McElroy et al. (2003) included clubs as nodes in their networks and showed the potential connections among 17 TB patients between 1994 and 2001 in Wichita, Kansas. De et al. (2004) also found a positive relationship between attendance at a motel bar and a gonorrhea infection in Alberta, Canada, in 1999 and used a contact network with the motel bar to demonstrate this connection. Based on these studies, many researchers believe that it is important to examine the social context of disease transmission in a contact network. Geographical locations are places of aggregation and create opportunities for social interaction. Including geographical locations in contact networks can not only help to reveal potential places for indirect or casual transmission contact, but can also help to identify social context which groups people and facilitates pathogen transfer.

3.4 Summary of SNA Studies in Epidemiology

Table 15-3 summarizes several SNA epidemiology studies in chronological order. Although Klovdahl's conceptual paper was published in 1985, the application of SNA in STD investigation did not start until the Colorado Springs study in 1994. Through the Colorado Springs study, SNA not only empirically demonstrated its ability to support contact tracing but also examined structural evolution of contact networks. Since then, STD with sexual contact has been the focus of analysis. In 2001, SNA was further applied to TB. Including geographical contact in the contact network was proposed to demonstrate airborne and casual contact transmission in public places. Because of the rich insights it provides, the inclusion of geographical contacts gradually became a standard practice for both TB and STDs to show the potential connection of patients via their daily activities.

Nonetheless, SNA has some limitations just like any other analytical tools. First, the accuracy of analysis depends on the quality of contact tracing data (Blanchard, 2002; Ghani et al., 1997). If contact tracing is not well executed and some key patients are not identified, a constructed contact network could be fragmented and fail to present a complete picture of transmission scenarios. All the analyses based on the contact network consequently could be misleading. Second, the qualitative visualization and quantitative measures of SNA are just tools for disease investigators to explore the phenomenon. To understand an outbreak with SNA, the investigators still need to consider many factors, including: environmental and social contexts, patient demographics, disease pathogen characteristics, etc. In addition, they need to interpret those data with their own domain expertise and insights (Rothenberg and Narramore, 1996).

Table 15-3. Summary of SNA studies in epidemiology.

Study	Disease	Type of Analysis	Location
Klovdahl et al., 1994	STD (AIDS)	Static	Colorado Springs
Woodhouse et al., 1994	STD (AIDS)	Static	Colorado Springs
Rothenberg et al., 1995	STD (AIDS)	Static	Colorado Springs
Latkin et al., 1995	STD (AIDS)	Static	Baltimore
Rothenberg et al., 1998b	STD (Syphilis)	Static	Atlanta
Rothenberg et al., 1998a	STD (AIDS)	Dynamic (Risk)	Colorado Springs
Potterat et al., 1999	STD (AIDS)	Dynamic (Risk)	Colorado Springs
Klovdahl et al., 2001	TB	Static	Houston
Potterat et al., 2002a	STD (Chlamydia)	Dynamic (Phase)	Colorado Springs
Potterat et al., 2002b	STD (AIDS)	Dynamic (Phase)	Colorado Springs
McElroy et al., 2003	TB	Static	Wichita-Sedgwick
Cunningham et al., 2004	STD (Syphilis)	Dynamic (Phase)	Baltimore
De et al., 2004	STD (Gonorrhea)	Static	Alberta
Andre et al., 2007	TB	Static	Oklahoma City

4. A CASE STUDY: THE SARS OUTBREAK IN TAIWAN

For our case study, we investigated the role of geographical contacts in disease analysis. In this section, we first review the Taiwan SARS outbreak of 2003 and introduce its contact tracing dataset. Then we present the two analyses, connectivity and topology analyses, used in our investigation.

4.1 Taiwan SARS Outbreak and Contact Tracing Dataset

SARS is an infectious disease caused by a novel coronavirus named SARS-associated coronavirus (SARS-CoV) (CDC, 2003; Lipsitch et al., 2003). Its first human case was identified in Guangdong Province, China, on November 16, 2002 (Chu et al., 2005). In February 2003, a medical doctor from Guangdong Province went to Hong Kong and infected at least 17 other guests during his stay at a hotel, initiating a global epidemic of SARS (Donnelly et al., 2003; Peiris et al., 2003). The epidemic ended in July 2003, with more than 24 countries reporting suspected or probable cases, including Canada, Singapore, and Taiwan.

SARS caused great public health concerns because of its rapid international spread, high case fatality rate, and unusual nosocomial infection. The majority of SARS patients were infected in healthcare and hospital settings (Peiris et al., 2003). SARS is highly contagious and transmitted primarily via close personal contact, through exposure to infectious respiratory droplets or body fluids. Some studies have also suggested that SARS may be transmitted via indirect contact based on infection incidents in transportation vehicles, hospitals, or communities (Chen et al., 2004; Peiris et al., 2003; Yu et al., 2004).

In Taiwan, a series of hospital outbreaks caused the number of SARS cases to dramatically increase to over 300 between April to June 2003 (Chu et al., 2005). The outbreaks started when a municipal hospital in Taipei received a SARS patient without a known source of infection in the middle of April. A week after her admission several healthcare workers gradually developed symptoms. The hospital was reported as having a hospital outbreak on April 22 and closed on April 24. Seven hospitals subsequently reported incidents of nosocomial infection and some suspended their emergency room operations, including a teaching hospital in Taipei. This series of outbreaks were suspected to have been triggered by inter-hospital transfer and the movement of SARS patients (Chu et al., 2005). On July 5, 2003, Taiwan was officially removed from a World Health Organization (WHO) list of SARS-affected areas.

The Taiwan SARS data was collected by the Graduate Institute of Epidemiology at National Taiwan University during the SARS period. It contains the contact tracing records of 961 suspected and confirmed SARS patients in Taiwan and their treatment histories. The records are comprised of two main categories, personal and geographical contacts. The personal contacts are those recognized interactions with known SARS patients in household, workplace, and hospital settings. The geographical contacts include visits to high-risk areas of infection, such as SARS-affected countries and hospitals. Table 15-4 summarizes the numbers of records and patients involved in each type of contact. It should be noted that a patient may have multiple records in a type and across types of contacts.

Table 15-4. Summary of the Taiwan SARS databaset.

Main category	Type of contact	Record	Suspected patients	Confirmed patients
Personal	Family member	177	48	63
	Roommate	18	11	15
	Colleague	40	26	23
	Close contact	11	10	12
Geographical	Foreign travel	162	100	27
	Hospital visit	215	110	79
	Hospital admission	622	401	153
	Hospital workplace	142	22	120
	High-risk area	38	30	7
Total		1425	638	323

4.2 Contact Network Construction

In order to present both personal and geographical contacts at one time, we adopted a two-mode network approach to construct a SARS contact network. This kind of approach has been taken in several studies, such as the Houston tuberculosis study by Klovdahl et al. (2001) and the Alberta gonorrhea study by De et al. (2004). The network contains two types of nodes, patients and geographical locations. We linked two patient nodes with an edge if they were family members or had an identified interaction. We connected a patient node to a location node, such as a hospital or foreign country, if the patient had been there during the SARS period. The construction of a contact network is demonstrated in Figure 15-3.

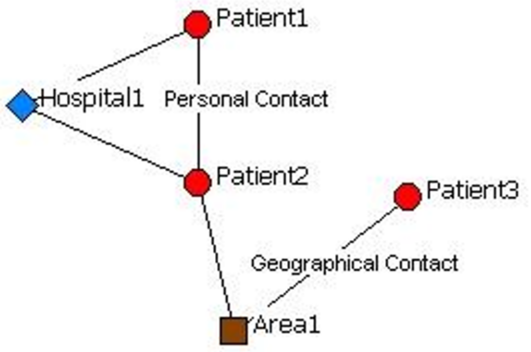


Figure 15-3. Example of contact network construction.

4.3 Connectivity Analysis

Connectivity is the degree to which a contact type can link individual patients in a network which can then be measured by the number of components. In order to understand how SARS spreads, connectivity analysis can be used to show the relative importance of geographical contacts, based on their ability to connect patients. If a type of contact has relatively high connectivity, it should significantly decrease the number of components from the total number of patient nodes. The types of contacts we investigated in this analysis are listed in Table 15-5.

Table 15-6 shows our results for the two main categories of contacts. After applying all available records, we can reduce the number of components in the network from 961 to 10. If we use the personal contacts alone for construction, the number of components decreases to 847 and the network is too sparse to get a comprehensive picture of how SARS spread in those patients. In contrast, the geographical contacts reduce the number of components to 82. This suggests that the majority of patients had been to the same place or places before the onset of their symptoms, indicating that knowing and analyzing the geographical contacts is important for understanding this outbreak.

Table 15-5. Types of contacts in the investigation.

Personal contacts	Geographical contacts
Family members	Foreign travel
Roommates	Hospital visits
Colleagues	Hospital admissions
Close contacts	Hospital workplaces
	High-risk areas

Table 15-6. Results of connectivity analysis for main categories.

	Number of components
Personal contacts	847
Geographical contacts	82
Personal + geographical	10

We further examined the connectivity of each type of contact, with Table 15-7 showing the results. Hospital-related contacts are the top 3 contacts in connectivity, consistent with the fact that SARS patients were primarily infected in the hospital setting.

Table 15-7. Connectivity analysis of the nine types of contacts.

Main category	Type of contacts	Number of components
Personal	Family member	893
	Roommate	946
	Colleague	943
	Close contact	949
Geographical	Foreign travel	943
	Hospital visit	753
	Hospital admission	409
	Hospital workplace	823
	High-risk area	924

4.4 Topology Analysis

A traditional social network, or one-mode network, is comprised of only one set of nodes and describes person-to-person relationships. A two-mode network, on the other hand, has the ability to portray micro and macro relations simultaneously. In topology analysis, the goal is to investigate the value of a two-mode contact network for deducing potential disease pathways.

Since a two-mode network contains two sets of nodes with different layers, personal and geographical, it emphasizes the relationships between patients and their visits to high-risk locations. Figure 15-4 shows the large number of patients whom have had contact with hospitals with outbreaks of nosocomial infection, such as Heping Hospital; the nodes representing patients surround each hospital. Through patients' visits and admissions, there are unusually complex linkages formed among the hospitals. These linkages may explain the series of hospital outbreaks in Taiwan.

Since a one-mode network is comprised of only patient nodes, we have to degrade geographical relations to person-to-person ones. To do this, we connect two patients together if they have been to the same geographical location. Figure 15-5 shows the transformed one-mode network. Generally, geographical contacts are collected to indicate potential occasions for infection when personal contacts are not traceable. After degrading, the linkage among patients was unnecessarily amplified to such a degree that meaningful patterns from the contact network could no longer be identified. In contrast, a two-mode contact network preserves important clues about the outbreaks from both person-to-person and person-to-location relations, even when hundreds of patients are involved in the graph.

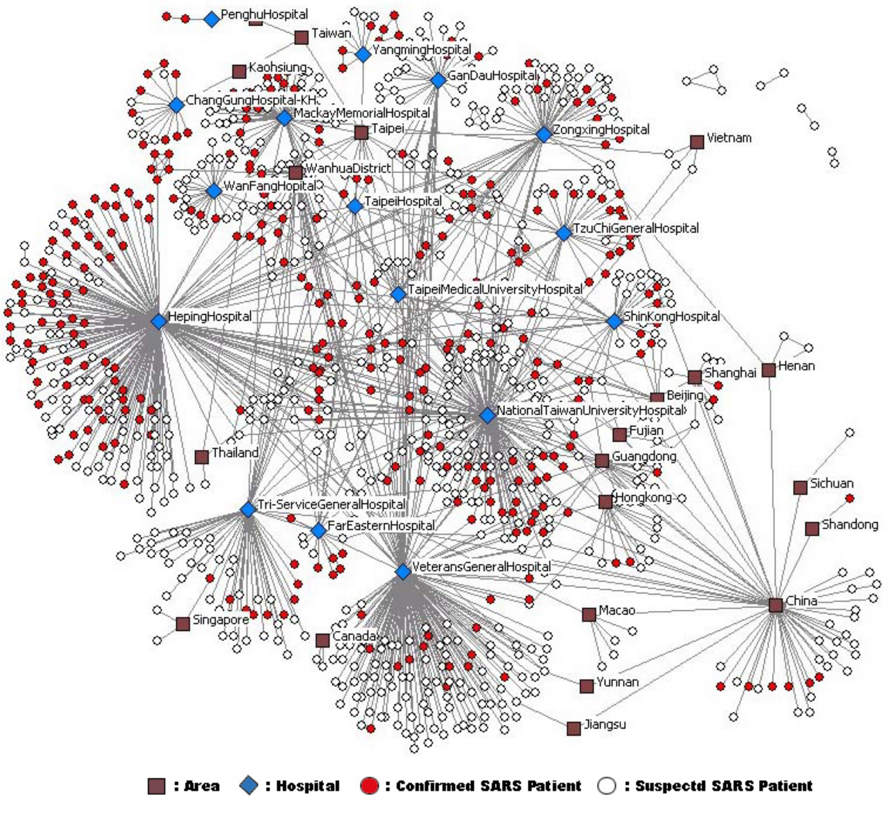


Figure 15-4. Two-mode SARS contact network.

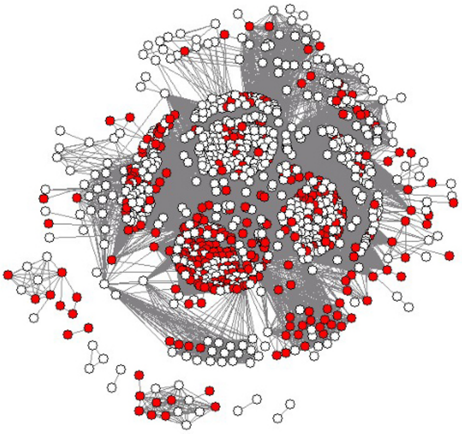


Figure 15-5. One-mode SARS contact network.

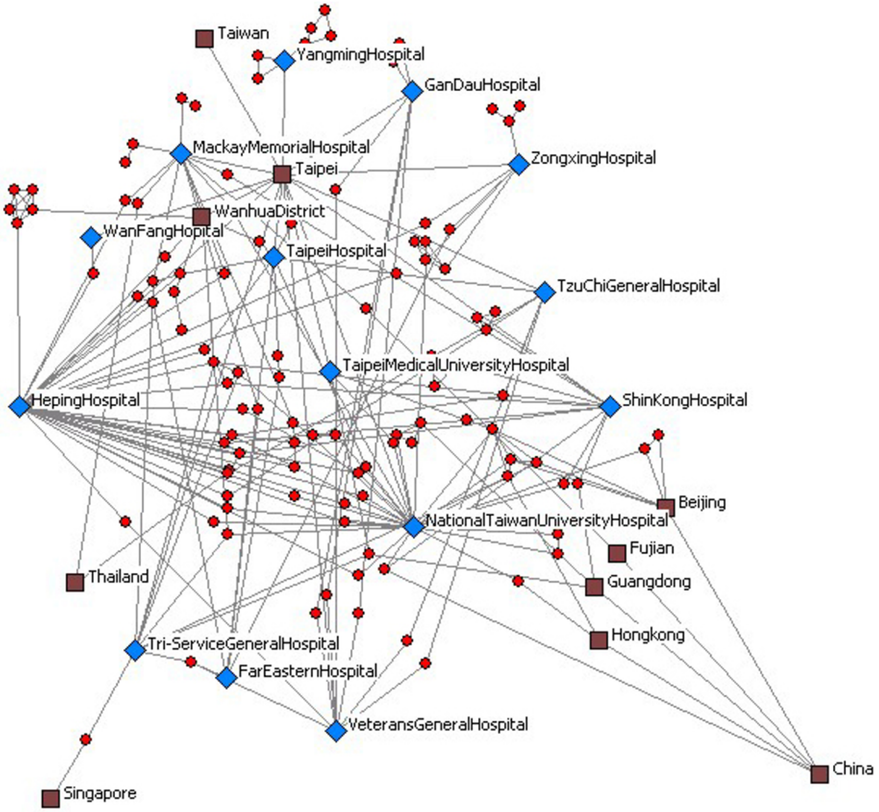


Figure 15-6. Potential bridges among hospitals and households.

The two-mode network stresses person-to-location relationships and presents patients as clusters around high-risk areas. In this type of layout, patients acting as bridges among major clusters are easily seen and identified. Figure 15-6 shows the potential bridges among the major hospitals with nosocomial infection.

When investigating a hospital outbreak, including geographical contacts in the network is also useful for seeing possible disease transmission scenarios. Figure 15-7 demonstrates the evolution of a small contact network at Heping Hospital through the onset dates of symptoms. On April 16, Mr. L., a laundry worker in Heping Hospital, had a fever and was reported as a suspected SARS patient. On April 16 and 17, Nurse C took care of Mr. L. On April 21, Ms. N, another laundry worker, and Nurse C began to have symptoms. On April 24, Heping Hospital was reported to have a hospital outbreak. On May 1, Nurse C's daughter had a fever. From the evolution of the network, development of the hospital outbreak can be readily discerned.

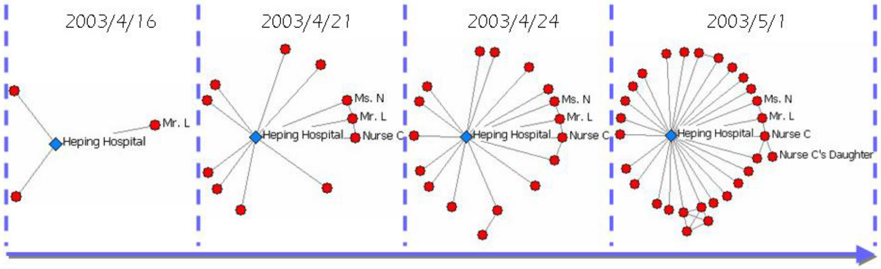


Figure 15-7. Example of network evolution through the onset dates of symptoms.

5. CONCLUSIONS

SNA has been demonstrated to be a good supplemental tool in the investigation of contact tracing. Compared to the traditional process of reviewing contact records one by one, SNA provides healthcare workers with a more efficient method of integrating and visualizing the relevant records in a contact network to discern potential linkages among patients, thus revealing disease pathways. Network measures, especially centrality measures, enable investigators to examine the context of transmission and develop effective intervention programs by identifying important individuals who may cause or exacerbate an outbreak. In addition, some studies have used SNA to study the transmission of disease dynamics, demonstrating that the structure of a contact network is a more accurate indicator of epidemic phases than the traditional secular trend data.

Incorporating geographical contact information in SNA allows disease investigators to analyze infectious diseases other than STDs. While personal contact provides direct evidence for the causality of infection, geographical contact captures the factors of human aggregation in disease transmission and provides potential leads to indirect or casual infection. In our case study, the role of a type of contact in disease transmission can be potentially identified by its ability to join patients together. Including geographical locations can significantly aid in establishing linkages among patients. Because these locations can play an important role in facilitating the transfer of pathogens, they require the attention of epidemiologists and other investigators of infectious disease.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation Information Technology Research Program, ITR, through Grant # IIS-0428241.

QUESTIONS FOR DISCUSSION

1. Contact tracing is an important control measure in the fight against an infectious disease. If you want to use contract tracing to control a developing outbreak, what kinds of data will you collect during the interview with confirmed patients? Discuss the question from two perspectives: disease control and outbreak analysis.
2. A contact network depicts the potential pathways of disease propagation among patients. Discuss the strengths and weaknesses of a contact networks in outbreak investigations.
3. Assume that you have a set of STD contact tracing data. It includes patients' sexual contacts, patronized bars and motels, and demographic information, such as patients' residency, gender, age, occupation, and income level. Discuss the kinds of analysis that can potentially be performed with this dataset and list your steps to investigate them using SNA.
4. Geographical contact information provides additional insights but can also create some problems when you include it in your disease analysis. Discuss the downsides of including geographical contacts in disease analysis and ways to reduce or eliminate them.

REFERENCES

Abernethy, N. F., 2005, Automating Social Network Models for Tuberculosis Contact Investigation. Dissertation, Stanford University, Stanford.

Andre, M., Ijaz, K., Tillinghast, J. D., Krebs, V. E., Diem, L. A., Metchock, B., Crisp, T., and McElroy, P. D., 2007, Transmission Network Analysis to Complement Routine Tuberculosis Contact Investigations, *Am. J. Public Health Nations Health* 97:470–477.

Auerbach, D. M., Darrow, W. W., Jaffe, H. W., and Curran, J. W., 1984, Cluster of Cases of the Acquired Immune Deficiency Syndrome: Patients Linked by Sexual Contact, *Am. J. Med.* 76:487–492.

Blanchard, J. F., 2002, Populations, Pathogens, and Epidemic Phases: Closing the Gap between Theory and Practice in the Prevention of Sexually Transmitted Diseases, *Sex. Transm. Infect.* 78:i183–i188.

CDC, 2003, Severe Acute Respiratory Syndrome - Singapore, 2003, *Morb. Mortal. Wkly. Rep.* 52:405–411.

Chen, Y. C., Huang, L. M., Chan, C. C., Su, C. P., Chang, S. C., Chang, Y. Y., Chen, M. L., Hung, C. C., Chen, W. J., Lin, F. Y., and Lee, Y. T., 2004, SARS in Hospital Emergency Room, *Emerg. Infect. Dis.* 10:782–788.

Chu, Y.-T., Shih, F.-Y., Hsu, H.-L. C., Wu, T.-S. J., Hu, F.-C., Lin, N. H., and King, C.-C., 2005, A Retrospective Review on the 2003 Multinational Outbreaks of SARS and the Preventive Measures of Its Nosocomial Infections, *Epidemiol. Bull.* 21:163–198.

Cunningham, S. D., Michaud, J. M., Johnson, S. M., Rompalo, A., and Ellen, J. M., 2004, Phase-Specific Network Differences Associated with the Syphilis Epidemic in Baltimore City, 1996–2000, *Sex. Transm. Dis.* 31:611–615.

De, P., Singh, A. E., Wong, T., Yacoub, W., and Jolly, A. M., 2004, Sexual Network Analysis of a Gonorrhoea Outbreak, *Sex. Transm. Infect.* 80:280–285.

Donnelly, C. A., Ghani, A. C., Leung, G. M., Hedley, A. J., Fraser, C., Riley, S., Abu-Raddad, L. J., Ho, L. M., Thach, T. Q., and Chau, P., 2003, Epidemiological Determinants of Spread of Causal Agent of Severe Acute Respiratory Syndrome in Hong Kong, *Lancet* 361:1761–1766.

Eames, K. T. D., and Keeling, M. J., 2003, Contact Tracing and Disease Control, *Proc. Biol. Sci.* 270:2565–2571.

Freeman, L. C., 1978/79, Centrality in Social Networks: Conceptual Clarification, *Soc. Networks* 1:215–239.

Ghani, A. C., Swinton, J., and Garnett, G. P., 1997, The Role of Sexual Partnership Networks in the Epidemiology of Gonorrhea, *Sex. Transm. Dis.* 24:45–56.

Klovdahl, A. S., 1985, Social Networks and the Spread of Infectious Diseases: The AIDS Example, *Soc. Sci. Med.* 21:1203–1216.

Klovdahl, A. S., Graviss, E. A., Yaganehdoost, A., Ross, M. W., Wanger, A., Adams, G. J., and Musser, J. M., 2001, Networks and Tuberculosis: An Undetected Community Outbreak Involving Public Places, *Soc. Sci. Med.* 52:681–694.

Klovdahl, A. S., Potterat, J. J., Woodhouse, D. E., Muth, J. B., Muth, S. Q., and Darrow, W. W., 1994, Social Networks and Infectious Disease: The Colorado Springs Study, *Soc. Sci. Med.* 38:79–88.

Latkin, C., Mandell, W., Oziemkowska, M., Celentano, D., Vlahov, D., Ensminger, M., and Knowlton, A., 1995, Using Social Network Analysis to Study Patterns of Drug Use among Urban Drug Users at High Risk for HIV/AIDS, *Drug Alcohol Depend.* 38:1–9.

Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, C., Chew, S. K., Tan, C. C., Samore, M. H., Fisman, D., and Murray, M., 2003, Transmission Dynamics and Control of Severe Acute Respiratory Syndrome, *Science* 300:1966–1970.

McElroy, P. D., Rothenberg, R. B., Varghese, R., Woodruff, R., Minns, G. O., Muth, S. Q., Lambert, L. A., and Ridzon, R., 2003, A Network-Informed Approach to Investigating a Tuberculosis Outbreak: Implications for Enhancing Contact Investigations, *Int. J. Tuberc. Lung Dis.* 7:S486–S493.

Peiris, J. S. M., Yuen, K. Y., Osterhaus, A. D. M. E., and Stohr, K., 2003, The Severe Acute Respiratory Syndrome, *N. Engl. J. Med.* 349:2431–2441.

Potterat, J. J., Muth, S. Q., Rothenberg, R. B., Zimmerman-Rogers, H., Green, D. L., Taylor, J. E., Bonney, M. S., and White, H. A., 2002a, Sexual Network Structure as an Indicator of Epidemic Phase, *Sex. Transm. Infect.* 78:152i–158i.

Potterat, J. J., Phillips-Plummer, L., Muth, S. Q., Rothenberg, R. B., Woodhouse, D. E., Maldonado-Long, T. S., Zimmerman, H. P., and Muth, J. B., 2002b, Risk Network Structure in the Early Epidemic Phase of HIV Transmission in Colorado Springs, *Sex. Transm. Infect.* 78:i159–i163.

Potterat, J. J., Rothenberg, R. B., and Muth, S. Q., 1999, Network Structural Dynamics and Infectious Disease Propagation, *Int. J. STD AIDS* 10:182–185.

Rothenberg, R. B., McElroy, P. D., Wilce, M. A., and Muth, S. Q., 2003, Contact Tracing: Comparing the Approaches for Sexually Transmitted Diseases and Tuberculosis, *Int. J. Tuberc. Lung Dis.* 7:S342–S348.

Rothenberg, R. B., and Narramore, J., 1996, Commentary: The Relevance of Social Network Concepts to Sexually Transmitted Disease Control, *Sex. Transm. Dis.* 23:24–29.

Rothenberg, R. B., Potterat, J. J., Woodhouse, D. E., Muth, S. Q., Darrow, W. W., and Klovdahl, A. S., 1998a, Social Network Dynamics and Hiv Transmission, *AIDS* 12:1529–1536.

Rothenberg, R. B., Sterk, C., Toomey, K. E., Potterat, J. J., Johnson, D., Schrader, M., and Hatch, S., 1998b, Using Social Network and Ethnographic Tools to Evaluate Syphilis Transmission, *Sex. Transm. Dis.* 25:154–160

Rothenberg, R. B., Woodhouse, D. E., Potterat, J. J., Muth, S. Q., Darrow, W. W., and Klovdahl, A. S., 1995, Social Networks in Disease Transmission: The Colorado Springs Study, *NIDA Res. Monogr.* 151:3–19.

Scott, J., 2000, Social Network Analysis: A Handbook. Sage Publications Inc, London.

Thomas, J. C., and Tucker, M. J., 1996, The Development and Use of the Concept of a Sexually Transmitted Disease Core, *J. Infect. Dis.* 174:S134–S143.

Wasserheit, J. N., and Aral, S. O., 1996, The Dynamic Topology of Sexually Transmitted Disease Epidemics: Implications for Prevention Strategies, *J. Infect. Dis.* 174 S201–S213.

Wasserman, S., and Faust, K., 1994, Social Network Analysis: Methods and Applications, Cambridge University Press, New York.

Woodhouse, D. E., Rothenberg, R. B., Potterat, J. J., Darrow, W. W., Muth, S. Q., Klovdahl, A. S., Zimmerman, H. P., Rogers, H. L., Maldonado, T. S., Muth, J. B., and Reynolds, J. U., 1994, Mapping a Social Network of Heterosexuals at High Risk for HIV Infection, *AIDS* 8:1331–1336.

Yorke, J. A., Hethcote, H. W., and Nold, A., 1978, Dynamics and Control of the Transmission of Gonorrhea, *Sex. Transm. Dis.* 5:51–56.

Yu, I. T. S., Li, Y., Wong, T. W., Tam, W., Chan, A. T., Lee, J. H. W., Leung, D. Y. C., and Ho, T., 2004, Evidence of Airborne Transmission of the Severe Acute Respiratory Syndrome Virus, *N. Engl. J. Med.* 350:1731–1739.

SUGGESTED READINGS

Klovdahl, A. S., 1985, Social Networks and the Spread of Infectious Diseases: The AIDS Example, *Soc. Sci. Med*. 21:1203–1216.

This is Klovdahl's 1985 conceptual paper which set out the program for SNA to support contact tracing analysis. It comprehensively discusses the relationships between disease transmission and the structure of social networks and points out the directions that SNA can contribute to the disease investigation, such as using centrality measures to find key individuals.

Scott, J., 2000, *Social Network Analysis: A Handbook*. Sage Publications Inc, London.

This is a good book for beginning reading about SNA. Rather than providing extensive details for each measure or method, it summarizes the key concepts of SNA and introduces them in a comprehensive way with clear examples. We (strongly) recommend reading this book before reading Wasserman and Faust's "Social Network Analysis: Methods and Applications."

ONLINE RESOURCES

UCINET is a popular SNA software by Analytic Technologies. It can be downloaded at http://www.analytictech.com/ with a 100-day trial. It provides both SNA visualization and analysis modules.

An online book for SNA is available at http://faculty.ucr.edu/~hanneman/nettext/. It contains good examples which illustrate basic concepts and methods of SNA and shows readers how to perform SNA with UCINET.

UNIT III: EMERGENCY RESPONSE, AND CASE STUDIES

Chapter 16

MULTI-AGENT MODELING OF BIOLOGICAL AND CHEMICAL THREATS

KATHLEEN M. CARLEY*, ERIC MALLOY, and NEAL ALTMAN

CHAPTER OVERVIEW

When pandemics, chemical spills, and bio-warfare attacks occur cities must respond quickly to mitigate loss of life. Which interventions should be used? How can we assess intervention policies for novel and low frequency events? Reasoning about such events is difficult for people due to the high level of complexity and the multitude of interacting factors. Computational models, however, are a particularly useful tool for reasoning about complex systems. In this paper, we describe a multi-agent dynamic-network model and demonstrate its use for policy assessment. BioWar is a city-level multi-agent dynamic-network model of the impact of epidemiological events on a city's population. Herein, we describe BioWar and then use it to examine the impact of school closures and quarantine on the spread and impact of pandemic influenza. Key aspects of the model include utilization of census data to set population characteristics, imputed social networks among agents, and flexible disease modeling at the symptom level. This research demonstrates that high-fidelity models can be effectively used to assess policies.

Keywords: BioWar, Multi-agent simulation, Pandemics, Dynamic networks

* *CASOS – Center for Computational Analysis of Social and Organizational Systems, ISR – Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_16,

1. INTRODUCTION

Evaluating the potential impact of social policies and estimating the unintended impacts of catastrophic events are key needs in the socio-policy arena. Most approaches to addressing these issues suffer from an inability to think through a wide range of alternatives and an inability to consider multiple interactions that are ever present in complex socio-cultural arenas. Herein, we demonstrate the value of high-fidelity multi-agent simulations for supporting such decision making.

We describe BioWar (Carley et al., 2006), a city-level dynamic-network multi-agent simulation system for reasoning about the impact of epidemiological events on city populations. Key aspects of the model are described. Then we illustrate the power of this model by using it to examine the potential impact of school closures and quarantines on the dispersion and health impact of pandemic influenza. Key validation challenges are discussed.

2. WHY MULTI-AGENT MODELING

Multi-agent models are increasingly described as the test-bed of choice for examining complex systems, particularly complex human social behavioral systems (Maxwell and Carley, 2009). The promise of these models is that they enable the researcher to focus on the activities of individuals and to derive social outcomes from the activities of and interactions among large numbers of heterogeneous agents. Such models, in comparison to laboratory or field experiments, are more ethical when dealing with life-threatening conditions, and typically are more cost-effective and time-effective for collecting data and reasoning about the impacts of policy (Carley, 2009).

Also, from a policy evaluation perspective, one of the key issues is veridicality. The veridicality of the model drives its design, assessment and validation (Carley, 1996). In general, the higher the level of veridicality the greater the problem domain flexibility, and the less likely it will be to validate the model in its entirety, and the greater the need for some validation. Consequently, developers either build simple models that in theory could be validated in full but for which there is no reason to validate them; or highly veridical models that can be used to address a plethora of policy issues but which can never be completely validated. Another key tradeoff is that as the Model Social Agent becomes more sophisticated, the number of agents that can interact within the model and that data can be gathered on with the same computational resources decreases. Consequently, there are models that have highly sophisticated and accurate agent models with only a handful of such agents or models with extremely simple agents represented by a handful of rules but millions of such agents.

In this paper we present BioWar, which takes a middle of the road approach. The BioWar model is sufficiently veridical that it can only be validated in parts, but can be used to address a wide number of policy issues. In addition, the agents in BioWar are sufficiently detailed and accurate that they make many human-level mistakes resulting in populations of hundreds of thousands of agents, rather than millions that can be simulated in a reasonable amount of time and space.

3. BIOWAR

BioWar is a scalable city-wide simulation, capable of simultaneously simulating the impact of background diseases, natural disease outbreaks and bioterrorism attacks on the population's behavior at the level of the individual. The simulator incorporates social and institutional networks, weather and climate conditions, and the physical, economic, technological, communication, health, and governmental infrastructures which modulate disease outbreaks and individual behavior. Individual behaviors include health seeking, entertainment and work/school patterns. A wide variety of output reports are generated based on the user's needs including absenteeism patterns, pharmaceutical purchases, doctor's office insurance claims reports, and hospital/emergency room reports (Figure 16-1).

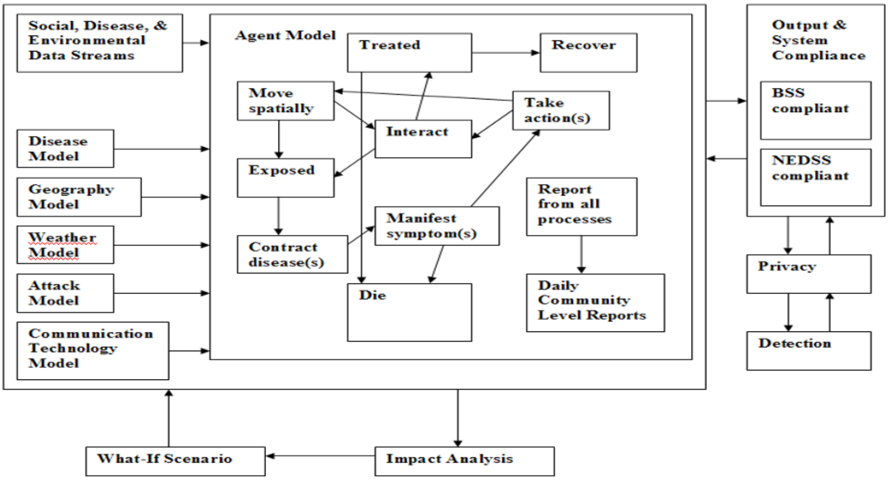


Figure 16-1. BioWar system diagram.

3.1 Agents

Agents are simulated individuals and are a fundamental unit of interest in BioWar. Agents are individually differentiated by demographics, location, medical condition and occupation. While agents move and interact within a metropolitan area, they in turn form the environment within which diseases are nurtured and spread.

BioWar was constituted to support simulations for a metropolitan area at a 1:1 ratio, so agents are structured to be fairly lightweight in terms of computational resources and memory requirements. This allows simulation of populations in excess of 1,000,000 agents employing what are currently considered moderately sized workstations.

3.1.1 Agent Characteristics

Simulated agents consist of a data structure and a set of algorithms to determine agent behavior. Agent characteristics such as age, sex and marital status are initialized to conform to the census demographics reported for the target metropolitan area. Agents also have a set of agent-to-agent connections (the social network) that defines strong social links between agents that is initialized based on social network research and a knowledge vector that helps define affinity between agents who come in contact.

An agent contains individual information about:

- Demographics: age, gender, race, occupation.
- Customary locations: home, work or school.
- Current location.
- Current and past diseases: disease type, state, symptoms.
- Social network: family, friends, coworkers, other strong ties.
- Temporary behavior modifiers ("interventions," e.g., shelter in place).

Agents are generated prior to simulation using demographic data for specific U.S. cities (primarily drawn from population and economic census information) along with templates on how agents should be located and grouped:

- Residency by location (aggregations of census tracts).
- School district geography.
- Social network partner types, interaction rates and network sizes.
- Family size and composition.

Because agents are generated separately from the simulation proper, the same population may be reused as an initial starting point for multiple simulations if desired. Equally, a fresh population can always be drawn up if desired. Populations can be scaled at any ratio between 0 and 100% of the base population data.

3.1.2 Agent Behavior

Agent activities include location-based movement and interaction with other agents with a corresponding possibility of infection. When agents visit locations as customers or are absent from their jobs and schools, they generate indicator data such as medical diagnoses, purchases of over-the-counter drugs, visits to medical information web sites and absentee reports as well as additional reports based on perfect knowledge (for example, the simulator knows with perfect certainty an agent's health status, while an agent generates indicators based on perceivable symptoms in "deciding" if they should visit a pharmacy and what to buy while there).

Agent characteristics are used as inputs to a population-specific set of algorithms, algorithms which influence their behavior in key areas, particularly healthcare seeking:

- Probability of absence.
- Recreational preferences.
- Rates of preexisting medical conditions.
- Likelihood of seeking medical assistance, by type of treatment.
- Workday duration, workweek and holiday schedules.

These algorithms were intended to emulate a modern technical society and in particular the USA in the early twenty-first century. These values may reside in parameter files, affording modification or localization, or within the program source.

3.1.3 Daily Agent Cycle

BioWar advances on a tick by tick basis. Ticks are resolved separately, but the simulator takes the time of day, day of week and holiday schedule into account when determining agent activities for each tick. The basic daily cycle for agents starts at midnight with two ticks spent at home and resting, two ticks at work or school (if the agent is of the correct age) and two ticks spent at home but active. Agents may break the basic cycle by being absent from home, work or school due to their health, because they choose an alternative activity (broadly referred to as recreation) or for unspecified other reasons (a residual value based on historical absentee counts). On weekends and holidays, agents do not go to work or school. This normal cycle may be interrupted by deliberately inserted interventions or attacks which are triggered by time or events in the simulation.

Agents are always placed in a geographical location appropriate for their selected activity. BioWar supports both workers and "customers" at all locations (customers are consumers of the location's service – students are a school's customer in this sense). BioWar creates locations for the simulation based on actual economic census data as to type and number and distributes them geographically within the metropolitan area using location database information where available and randomly if not. Locations are nodes of agent activity, typically structures (such as schools, businesses or homes) or places of public gathering (such as parks). While agents consider distance as a factor in selecting the next location to go to, movement between locations is highly abstract; agents do not spend time in transit but are placed at the appropriate location at the start of each tick.

3.1.4 Agent Interaction

While an individual agent's actions are largely determined independently of the other agents, agents potentially interact with each other on every tick. BioWar uses two methods to select candidate agents for interaction: social network-based and random. Once an agent is added to the interaction list, the interaction is resolved in a uniform way.

The social network represents strong ties between individuals, including family, friends, coworkers and classmates, using the University of Chicago General Social Services (GSS) survey data with the addition of "schoolmate" for younger agents, a population not covered by the GSS (2009). Because the research data on social networks emphasizes relatively strong ties, the BioWar social network size range is relatively small in relation to the total number of agents in the simulation. BioWar simulates a single metropolitan area at a time, so an agent's social network partners are artificially constrained to the agent population in the simulation.

Selection of potential partners from the social network is affected by the agent's current location. For instance, an agent at their workplace selects from agents in their social network who currently share that location, which increases interaction with coworkers during business hours.

In addition to the social network, random interactions are used to simulate casual or chance contacts (for example, with a fellow bus passenger). During each tick, BioWar selects agents from the full agent pool to add to an agent's interaction list. BioWar's random selection algorithm biases the selection towards agents who are physically close to the target agent.

The combined list of candidate interaction partners is then resolved. The probability of actual interaction is adjusted by the degree of similarity between agents, as represented by their knowledge vector. If the interaction occurs, agents can exchange knowledge and infectious diseases. After infection, the disease progresses in the infected agent according to the disease model.

3.2 Diseases

The current version of BioWar includes a default set of 66 diseases. Users may add additional diseases and customize existing diseases. Diseases include naturally occurring types as well as variants which are weaponized for biological attacks or which have been eliminated due to public health initiatives.

3.2.1 Disease Model

BioWar employs a symptom-based general disease model. Each disease has its own set of symptoms, timing of disease phases, variability in presentation based on age, gender, and race, and contagiousness. Each symptom has its own severity and progression timing. Furthermore, symptoms are assigned an "evoking strength" so that diagnoses based on symptoms will not only reflect accepted medical protocols but will also mimic the errors inherent in these protocols.

Each instance of a disease infecting an agent is individually represented and progresses through time as the agent goes about his or her daily routines. Diseases can propagate through a population, a process which is probabilistically determined by agent risk factors, the transmissibility of the disease, and the spatial and temporal proximity of uninfected agents to infected agents.

In human populations, certain demographic groups are more likely to be susceptible to particular diseases than others. These risk factors increase a person's susceptibility to diseases through either host factors or environmental factors to which that person is exposed. For example, individuals who have contact with animals (sheep shearers, for example) are more likely to contract cutaneous anthrax than other occupations. In BioWar, risk factors are distributed a priori to individuals in the population according to demographic characteristics based on age, sex, race, and disease prevalence.

In constructing our disease model, we used historical accounts of known anthrax releases (Inglesby et al., 1999), documents from the October, 2001 bioterrorism attack (Perkins et al., 2002), and disease knowledge bases (USAMRIID, 2001; West, 2001; Isada et al., 2003). We have also drawn on the experience of other medical expert systems developed to assist in diagnosis to ground our disease model in well-founded medical knowledge representations (Miller et al., 1982).

3.2.2 Disease Introduction

The current disease model supports three different means of introducing diseases into the simulation: attacks, outbreaks and background introduction. Diseases are not rigidly restricted to use one introduction method, although the default individual disease definitions were generally established with either attack/outbreak or background introduction in mind.

Attacks are under the user's control – severity, disease type, duration, and attack locations can be controlled separately. No default attack is programmed. Attacks might be created to simulate a bioterrorism event, accidental release or arrival of a new disease into a city. BioWar models airborne transport of spore-based diseases as well as agent infection.

Outbreak introduction allows diseases to be instantiated into a simulated population in a predetermined pattern, much like what might be expected over the normal course of a year. Diseases like influenza, which are seasonal and whose annual severity varies, are well suited for outbreak introduction. Although outbreak diseases can be controlled by the user, the default disease pattern would normally be acceptable except when special circumstances need to be considered.

Unlike the first two groups, background instantiation is controlled at simulation time by prevalence statistics gathered from California Department of Health data repositories. Background diseases are considered to be chronic diseases, so agents are selected to have background diseases at the start of the simulation based upon these statistics. Furthermore, background disease cases normally persist for the duration of the simulation.

3.2.3 Disease Progression

Agents experiencing disease state transitions are modeled as nondeterministic automata. As past medical history affects these transitions, this is a non-Markovian model. At any time within the duration of a state, a medical intervention can occur and the state can be changed. The state of the disease also affects the medical intervention.

Each disease instance progresses through up to five phases:

1. Incubation: the period of time before the agent begins presenting symptoms due to a bacterial or viral infection.
2. Early symptomatic (prodromal): the period of time during which an infected agent may experience mild or nondescriptive symptoms. Many diseases omit this phase, or have no known or identifiable early symptomatic period.

3. Late symptomatic (manifestation): the period of time during which an infected agent may experience severe and/or disease-specific symptoms. In many diseases, this phase may not be distinct from the early symptomatic phase.
4. Communicable: the period of time during which an infected agent may infect other agents. This phase may overlap with the above phases. Noncontagious diseases do not have this phase.
5. Recovery/death: a period of time during which an infection resolves or causes death.

In the current version of BioWar, the length of each phase except recovery/death is generally determined uniformly randomly using a range provided by expert analysis. Recovery and death of an agent, when not affected by treatment, is determined by a Bernoulli process with p equal to the death rate of the disease among untreated victims (again, determined by expert analysis). The duration of dying and recovering is likewise stochastically determined.

3.2.4 Medical Diagnosis and Treatment

BioWar employs a symptom-driven diagnosis and treatment model. Symptoms are important in BioWar on two levels. They motivate agent behavior and determine the initial diagnosis of agents entering the medical system. Agents with symptoms self-diagnose, stay home from work, visit their doctor or pharmacist, and change their patterns of interacting with others, depending on the severity of symptoms. This symptom-based disease model permits the representation of outliers and stochastic flux (not everyone with the same disease presents the same symptoms). The symptoms are assigned two different measures that influence which symptoms agents get and how that changes their behavior (Miller et al., 1982).

The first, frequency, is a qualitative measure of how frequently people with a particular disease will manifest a particular symptom. Frequency is denoted by a number between 1 and 5 that answers the question: “In patients with disease x, how often does one see symptom y?” For example, patients with the diagnosis of anthrax will have a fever frequency of 5 – nearly all patients with anthrax will have fevers at some point in the course of their disease. Second, the evoking strength is a qualitative measure of how frequently a doctor will associate a particular symptom with a particular disease.

Evoking strength is coded as a number between 0 and 5. It answers the question: “When you see symptom y, how often would a doctor think the cause is disease x?” For example, fever symptoms are not specific to any one disease – in our disease profile of anthrax, fever is given an evoking strength of 1. However, widened mediastinum is a more specific manifestation of anthrax – in patients who have a widened mediastinum, the diagnosis of anthrax should be considered thus the evoking strength for this is 5. Evoking strength is similar to specificity. Symptoms are present during both symptomatic phases with time-varying severity. Our current implementation of time-varying severity is a simple additive increase over time since a symptom was introduced.

Agents self-diagnose on the basis of visible or palpable symptoms. Medical personnel diagnose on the basis of visible symptoms and other information, which can include laboratory tests of varying accuracy (type 1 and 2 errors are possible) and report time. Due to the covert nature of weaponized biological attacks, doctors and ER personnel may or may not be anticipating the appearance of a particular bioagent, resulting in some degree of misdiagnosis. Moreover, doctors and ER personnel take time to file a report, delaying institutional realization of a bioattack.

BioWar employs a symptom-based differential diagnosis model to obtain information on the diseases infecting an agent who visits a medical facility. Our goal was not to build an error-free diagnosis model. Rather, we use differential diagnosis, as do medical doctors, which allows the possibility of initial misdiagnosis and the revision of diagnoses with additional information (e.g., lab results). We have based the model on the Internist1/QMR diagnosis model, but have augmented the results with probabilistic "switches" to help control aspects of the returned diagnosis (including rate of correct and incorrect diagnoses, and distribution of primary and secondary diagnoses by ICD-9 code). As such, our model is not a true computational diagnostic tool, but serves to control the simulator's response to diseases in a simulated population.

Initial medical diagnosis is simulated based on the apparent symptoms and their evoking strengths. To determine which disease a person has, the groups of evoking strengths of symptoms associated with potential diseases are compared and the highest one is chosen as the diagnosed disease. In other words, the disease most strongly associated with the most severe set of symptoms is chosen. This produces a certain amount of inaccuracy, mimicking the real world. The diagnosis determines whether a person is treated properly or not and whether advanced tests are ordered. Subsequent diagnosis can update the primary diagnosis based on the appearance of new symptoms and on the results of diagnostic testing. Chief complaints are not necessarily the same as discharge diagnosis, which is consistent with observed hospital performance (Begier et al., 2003). Treatment may not be immediately effective and symptoms vary in visibility and type of testing required for their detection. In the current version of BioWar, treatment is modeled as a simple time-delayed probability of a success.

3.3 Cities

BioWar creates a scaled representation of a real metropolitan area by using census, geographic and climatic data for a given geographic area, the "input deck." Data from the input deck is used to instantiate a specific city instance for use in a simulation. For example, a cross tabulation of population by sex, race and age for a given geographic area, when instantiated, becomes a list of individual agents who, in aggregate, reflect the composite statistics used to create them.

In addition to agents, BioWar also creates a set of functionally differentiated locations where agents live, work, study, shop and relax. Each location is placed within the geographical outline of the metropolitan area, avoiding water features. The number and type of locations conform to census economic data. Agents are associated with a specific home and age appropriate school and work locations. They will also visit other locations based on medical need or for other activities such as recreation. With the exception of homes, locations function as workplaces and may be open to "customers" who visit the location for services (e.g., a school's customers are students).

In all BioWar cities created to date, the basic unit used to create a BioWar input deck is the US Office of Management and Budget Metropolitan Area (MA):

Metropolitan and micropolitan statistical areas (metro and micro areas) are geographic entities defined by the US Office of Management and Budget (OMB) for use by Federal statistical agencies in collecting, tabulating, and publishing Federal statistics. A metro area contains a core urban area of 50,000 or more population, and a micro area contains an urban core of at least 10,000 (but less than 50,000) population. Each metro or micro area consists of one or more counties and includes the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core. (U.S. Census Bureau, 2008)

Two types of metropolitan areas are used for BioWar (U.S. Census Bureau, 1994):

- Metropolitan Statistical Areas (MSA) – A stand-alone metropolitan area.
- Primary Metropolitan Statistical Area (PMSA) – A predefined subunit of a metropolitan area with a population of one million or more (the parent metropolitan area is termed a "Consolidated Metropolitan Statistical Area" (CMSA)).

In many simulation runs, subsets of the MSA or PMSA are used. In this case, specific counties (or county equivalents) that are the building blocks for the metropolitan area are selected from within the defined metropolitan area and only data from those counties is used to build the input deck. In all cases, the simulation area consists of contiguous land areas and any intervening water features (rivers, lakes or bays).

3.4 Additional Features

3.4.1 Climate and Weather

BioWar includes climate and wind models which generate weather profiles typical for each of the simulated regions, including seasonal variation. These profiles are varied for each instantiation of the simulation but the averaged measures match historic records.

Weather, particularly precipitation, occasionally affects normal agent behavior through events such as school closure due to snow, but the primary purpose and utility of the climate inputs is in modeling transport of wind borne attack agents. For this purpose, the simulation provides several wind transport models which can be selected as part of the attack specification.

3.4.2 Chemical Attacks

In addition to biological agents, BioWar supports attacks using chemicals. As with biological attacks, BioWar includes default chemical definitions which the user may modify, or the user may add additional definitions. The effect of chemicals on agents is similar to diseases in the sense that the agent expresses the effect of chemical exposure through symptoms which medical personnel use to diagnose and treat the afflicted.

3.4.3 Interventions

Running in default mode, BioWar is well suited for examining how well disease progresses in a population and evaluating methods for monitoring public health and disease detection. In cases where the analyst wishes to dynamically alter simulation behavior during execution, BioWar provides interventions. From a user perspective, interventions are a specialized scripting language which changes elements of the default simulation behavior for a specified (but adjustable) duration of the simulation run. For instance, an intervention can force agents to shelter in place in response to a chemical attack.

Interventions include a triggering event, such as the occurrence of a disease, a disease diagnosis or simply a fixed time, followed by a specification of the simulation element(s) affected by the intervention (typically a subset of the agent population) and then the behavioral effect. The analyst has considerable flexibility in adjusting the parameters of any intervention although the suite of available intervention types is limited to the modifiable behaviors built into the simulation. In more complex studies, fresh interventions are often introduced to the simulation to support specific research needs.

While interventions can be completely scripted prior to simulation start for autonomous execution, they can also be created and inserted by the analyst during the simulation run. Using a simple command line interface, the analyst can pause the simulation to insert interventions, or an intervention can be created which automatically pauses the simulation on a triggering event, waiting until the analyst crafts any desired additional intervention(s) and chooses to continue the simulation.

3.4.4 Scalability and Configurability

In common with most large simulations, BioWar employs a large number of constants and parameters that affect program execution. Wherever possible, these values are made available to the user rather than hidden in the code. Based on the expected utility, two levels of access are provided:

- Configuration values – read from commented configuration files in the input directory. These contain the constant values most typically adjusted by the user, input file names and desired output reports. While all values have default settings, some are considered as place holders that the user is expected to adjust.
- Environment values and Definition files – these values are again read from files but are values which users seldom wish to change, for which the standard default values are considered sensible or which are complex to adjust. The file format and location are less accessible to the user but are available without program modification.

Commonly adjusted parameters include:

- Simulation scale: size of simulation relative to actual city, typically from 10 to 100% of actual.
- Simulation duration: duration of simulation execution, typical values range from 3 months to 2 years.
- Output: data recorded by BioWar during execution is output in the form of reports. Since the overhead in report output can be substantial, the user can limit output to include only the relevant reports.

A few inputs have no default settings:

- Attacks: injection of abnormal events is purely at the option of the user.

- Interventions: again, dynamic modification of simulation behavior is considered to be a user task.

The simulation will, however, run without problems if no attacks or interventions are provided.

4. ILLUSTRATIVE RESULTS

BioWar can be used as a standalone tool or as a utility to generate data streams for processing by other tools. For example, BioWar has been used as a data generator for blind testing and evaluation of syndromic surveillance algorithms by creating sets of hospital and clinic records, some of which contain bioterrorism incidents and some which do not.

As an illustration of its capabilities as a standalone tool, BioWar was used to evaluate two response strategies for an outbreak of avian influenza in a U.S. city (see also Lee et al., 2008). The simulation used a 1:1 representation of Norfolk, VA, as a test location. The city was instantiated based on year 2000 census data, with a total of 1,530,908 agents. BioWar then simulated the population for a year. Early in the simulation, 100 cases of a virulent strain of avian influenza were introduced in randomly selected agents. The disease strain is communicable from agent to agent but an agent can only contract the disease once.

Once the disease was established, the simulation was varied by introducing possible response strategies into the simulation. The three conditions tested were:

- Do nothing; run simulation to completion normally.
- Close all schools for 60 days when 50 active disease cases occur.
- Make agents quarantine themselves by staying home for 60 days when 50 active disease cases occur.

In each scenario, agents continue to interact with each other and can contract the diseases normally (i.e., measures such as wearing respirators or limiting interpersonal contacts were not simulated nor was vaccination attempted). When schools are closed, students will not attend school, but continue to move between other simulated locations. In the quarantine case, agent mobility is curtailed.

Each response case was executed ten times. In the do nothing cases, agent mortality is high and the majority of agents contract the disease. Closing schools had little to no effect on mortality or the number of days agents are sick. Strict quarantine measures do reduce agent mortality dramatically (Figure 16-2).

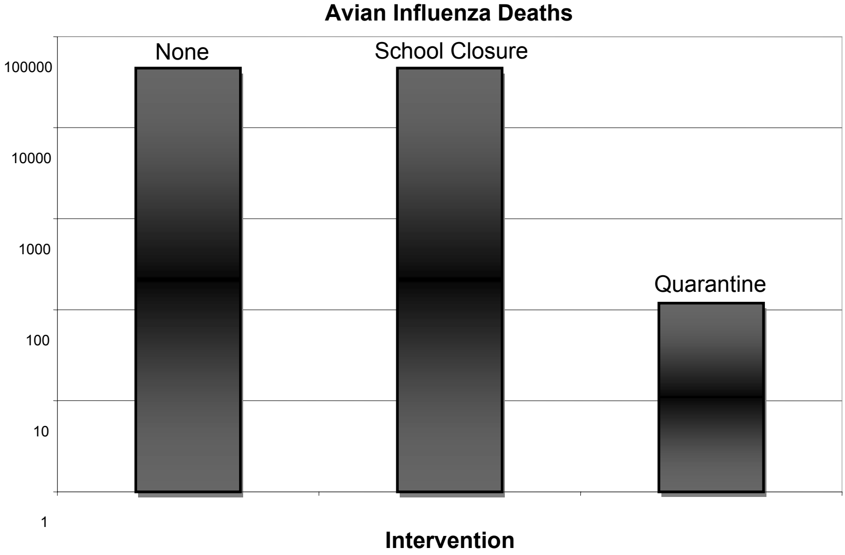


Figure 16-2. Average number of avian influenza deaths for each tested response strategy.

These runs suggest the importance of limiting the variety rather than the number of agent connections. In effect, the city divides itself into a series of smaller units for the duration of the quarantine response. Of course, promptly imposing such a long, strict quarantine is not entirely realistic, but the results immediately suggest a number of questions for further exploration:

- What is the minimum duration of strict quarantine required to cut mortality?
- How does quarantine compare with prompt isolation of patients?
- Would simply reducing the number of normal agent-to-agent contacts or lowering the probability of disease transmission have a similar effect?
- How quickly does the quarantine response need to be imposed and is there a point after which this response is no longer a viable option?

5. VALIDATION ISSUES

In general, the level and type of validation should depend on why the model was built and how the results are to be used (Burton, 1995b). In a policy context, validation of key variables and some calibration is appropriate for highly veridical models like BioWar. A typical approach to validation is input validation where a model is developed based on some theory, instantiated with data from a real situation, and then used to generate a series of results about variable “y.” These results are typically never checked against real-world data given the paucity of data and the low frequency of events; but are used to set policy directions. This approach has been taken with BioWar. In addition, in previous work, we were able to show that BioWar could generate behavior that was quantitatively and qualitatively similar to historic smallpox, anthrax, and general influenza events (Chen et al., 2006).

In general, validation, as practiced for engineering models, may be inappropriate at worst and impossible at best for highly veridical models like BioWar (given the model complexity, large number of variables and infeasibility of constructing a complete response surface). The data needed to completely validate such models does not exist. Rather, to do any type of validation we have had to collect disparate data from a wide variety of sources, from disparate time periods, populations, and under different environmental conditions, fuse the data together, and then use it collectively to validate aspects of the model. We note that the time and resources spent on BioWar validation has exceeded that used to build the model, run the virtual experiments, and analyze the results. It is also important to recognize that validation to a single historic event tends to lead to model over-fitting. To avoid this, we have used three historic scenarios: anthrax cases, smallpox cases and influenza. Finally, we have used a nontraditional validation practice called docking (Burton, 1995a; Axtell et al., 1996). For this process we demonstrated that BioWar can generate results quantitatively and qualitatively similar to those produced by standard SIR models.

6. CONCLUSION

There are several key limitations to the approach described here. First the model is very U.S. centric. As such, to use it in a foreign context would require segregating and rebuilding the U.S.-centered features such as doctors’ offices being closed on weekends and evenings and use of the emergency rooms at hospitals in those cases. A second limitation is that second order infections acquired while in a hospital are not modeled. Whether this is a significant impact on the results is not known. Third, the construction of cities is a data intensive and time consuming process. While it is relatively straightforward to build the population from census data and impute the social networks among agents, the process of building the physical landscape to match an actual city is complex. In particular, data on schools, location of public gathering spots and so on are often not readily available in machine readable form. Approximations can be used; but that means that variations due to the exact physical geography cannot be accounted for.

It should be noted that a key feature of this approach is the city modeling. The ability to construct cities from census-level data exists independent of BioWar. As such, artificial cities can be developed and used in contexts other than disease propagation.

BioWar can be used to examine a wide number of interventions and estimate the relative impact of those interventions in mitigating the population's response to infectious disease. Behavior in terms of infection, death and recovery can be explored. Since diseases are modeled at the symptom level, new diseases and those that are only hypothetical can be rapidly modeled and the repercussions of their dispersion explored. In general, this approach holds great promise for policy analysis.

ACKNOWLEDGEMENTS

This work is part of the Dynamics Networks project at the center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). BioWar is a DARPA legacy system with initial support from the program on Scalable Biosurveillance Systems. This work was supported in part by the Office of Naval Research for Bio-Surveillance Modeling (N00014-07-10696), (NSF) Integrative Graduate Education and Research Traineeship (IGERT) program in Computational Analysis of Social and Organizational Systems (IGERT 9972762), and the Air Force (FA865007C6769) for research on the use of multi-agent models to study unintended consequence. Additional support was provided by the National Science Foundation (NSF 045 2598) and the Air Force Office of Scientific Research (MURI: Cultural Modeling of the Adversary, 600322). Many of the computations were performed on the National Science Foundation Terascale Computing System at the Pittsburgh Supercomputer Center. Further, support for virtual experiments was provided in part by the National Science Foundation through TeraGrid resources provided by Purdue University. The views and proposals contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DARPA, the Office of Naval Research, the Air Force Office of Scientific Research, the National Science Foundation, or the U.S. government. The authors would like to thank Terrill Frantz and Ian McCulloh for extensive and helpful feedback on earlier drafts.

QUESTIONS FOR DISCUSSION

1. Compare statistical forecasting and multi-agent simulation from basic theory. What type of data is needed for statistically reasoning about events like pandemics? Given the paucity of such data, how can the little data that is available be effectively utilized to improve simulation models?
2. Engineering simulations are often not stochastic and so can be empirically validated against device data. In contrast socio-cultural simulations are stochastic and generate a distribution of possibilities. Hence validation against a historical event can lead to over-fitting. What are the dangers of over-fitting?
3. What are key interventions? How would you model them? What data could be gathered to make sure that the representation of the interventions is accurate?

REFERENCES

Axtell, R., Axelrod, R., Epstein, J.M., and Cohen, M.D. (1996). "Aligning Simulation Models: A Case Study and Results," *Computational and Mathematical Organizational Theory*, 1(2), 123–141.

Begier, E.M., Sockwell, D., Branch, L.M., Davies-Cole, J.O., Jones, L.H., Edwards, L., Casani, J.A., and Blythe, D. (2003). "The National Capitol Region's Emergency Department Syndromic Surveillance System: Do Chief Complaints and Discharge Diagnosis Yield Different Results?," *Emerging Infectious Diseases*, 9(3), 393–396.

Burton, R.M. (1995). "Validation and Docking: An Overview, Summary and Challenge," in *Simulating Organizations: Computational Models of Institutions and Groups*. MIT Press: Cambridge, MA, 215–228.

Burton, R.M. and Obel, B. (1995). "The validity of computational models in organization science: From model realism to purpose of the model," *Computational & Mathematical Organization Theory*, 1(1), 57–71.

Carley, K.M., Fridsma, D.B., Casman, E., et al. (2006). "BioWar: Scalable Agent-Based Model of Bioattacks," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 36, 252–65.

Carley, K.M. (2009). "Computational Modeling for Reasoning About the Social Behavior of Humans," *Computational, Mathematical and Organization Theory*, 15(1), 47–59. Available: http://springerlink.com/content/k44jr16031412578/, Retrieved: 4/2009

Carley, K.M. (1996). "Validating Computational Models," Working Paper.

Chen, L., Carley, K.M., Fridsma, D., Kaminsky, B. and Yahja, A. (2006). "Model Alignment of Anthrax Attack Simulations," *Decision Support Systems,* 41(3), 654–668.

GSS - General Social Survey (2009). http://www.norc.org/GSS+Website/, Retrieved: 4/2009.

Inglesby, T.V., et al. (1999). "Anthrax as a Biological Weapon: Medical and Public Health Management," *Journal of American Medical Association*, 281(18) (May 12, 1999): 1735–1745.

Isada, C.M., Kasten, B.L., Goldman, M.P., Gray, L.D., and Aberg, J.A. (2003). *Infectious Disease Handbook*, AACC.

Lee, B.Y., Bedford, V.L., Roberts, M.S., and Carley, K.M. (2008), "Virtual epidemic in a virtual city: simulating the spread of influenza in a US metropolitan area," *Translational Research*, 151(6), 275–87.

Maxwell D. and Carley K.M. (2009). "Principles for Effectively Representing Heterogeneous Populations in Multi-Agent Simulations," in *Complex Systems in Knowledge Based Environments*, Tolk, A. (ed.), Ch. 8, 199–228, Springer-Verlag.

Miller, R.A., Pople, H.E., and Myers, J.D. (1982). "Interist-I, An Experimental Computer-based Diagnostic Consultant for General Internal Medicine," *The New England Journal of Medicine*, 07:468–76.

Perkins, B.A., Popovic, T., and Yeskey, K. (eds.) (2002). "Bioterrorism-Related Anthrax," *Emerging Infectious Diseases*, 8(10) (special edition). http://www.cdc.gov/ncidod/EID/vol8no10/contents_v8n10.htm, Retrieved: 3/2008.

USAMRIID - US Army Medical Research Institute for Infectious Diseases (2001). *USAMRIID's Medical Management of Biological Casualties Handbook*.

U.S. Census Bureau (2008). *Metropolitan and Micropolitan Statistical Areas*. http://www.census.gov/population/www/estimates/metroarea.html, Retrieved: 3/2008.

U.S. Census Bureau (1994). *Geographic Areas Reference Manual*, Available: http://www.census.gov/geo/www/garm.html, Retrieved: 3/2008.

West, K. H. (2001). *Infectious Disease Handbook for Emergency Care Personnel*, ACGIH.

SUGGESTED READINGS

Keeling, M.J. and Rohani, P. (2007). *Modeling Infectious Diseases in Humans and Animals*, Princeton University Press: Princeton, NJ.

Provides a comprehensive introduction to infectious disease modeling and has an associated web site. R and C++ models.

Epstein, J.M. and Axtell, R. (1996). *Growing Artificial Societies: Social Science From the Bottom Up*. MIT Press/Brookings Institution: Cambridge, MA.

Early multi-agent simulation system showing the power of bottom-up reasoning even when using highly simplistic models.

http://www.econ.iastate.edu/tesfatsi/abmread.htm

Provides a good general introduction to multi-agent simulations.

ONLINE RESOURCES

Simulations of disease need to make use of existing disease descriptions and ontologies. These tend to be maintained by the military, the CDC and various professional societies:

http://www.pandemicsimulation.com/Members/admin/ontologies-for-pandemic-simulations
http://www.cdc.gov/datastatistics/
http://diseaseontology.sourceforge.net/

http://www.swarm.org/index.php/Agent-Based_Models_in_Biology_and_Medicine

General information on multi-agent modeling is described on several agent sites:

http://www.econ.iastate.edu/tesfatsi/ace.htm
http://www.casos.cs.cmu.edu/
http://www.intute.ac.uk/cgi-bin/advancedsearch.pl?field=All&term1=social+simulation&Search=Search&limit=0&subject=All&c00=Any&rank=score&fielddisplay=All

Illustrative bio and health models

http://www.brookings.edu/topics/agent-based-models.aspx
http://www.casos.cs.cmu.edu/projects/biowar/
https://www.vbi.vt.edu/public_relations/press_releases/chicago_pandemic_influenza_simulation

Chapter 17

INTEGRATED HEALTH ALERTING AND NOTIFICATION

A Case Study in New York State

LINH H. LE,[1]*, DEBRA L. SOTTOLANO[1], AND IVAN J. GOTHAM[1,2]

CHAPTER OVERVIEW

This chapter describes the evolution, design and operation of an integrated health alerting and notification system. Although modern information and communication technologies have evolved rapidly in the last decade and significantly improved the technical foundation for health alerting and notification, published literature to communicate lessons learned, best practices, system design, development and operation from real-world experiences is still limited. In this chapter, we outline the functional and technical requirements as well as architecture and components of the New York State (NYS) Integrated Health Alerting and Notification System (IHANS), address issues that affect the effectiveness and timeliness of health alerting, discuss the concept of unified messaging and the current standard alert message distribution framework, and present case studies of health alerting in emergency events and exercises. Experience from development and operation of this system has shown the important role of an integrated health alert system in public health arising from both changes in information and telecommunication technologies and new demands on the public health and healthcare systems.

[1]* *Bureau of Healthcom Network Systems Management, New York State Department of Health, Empire State Plaza, Room 148, Albany, NY 12237, USA, lhl02@health.state.ny.us.*

[2] *School of Public Health, State University at Albany, Albany, NY 12222, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_17,

Keywords: Health alert; Notification; Public health preparedness; Public health informatics; Interoperability; Unified communication; Messaging; Telecommunication

1. INTRODUCTION

Since 2001, health alerting and notification have become a critical component of public health information systems in general and especially in infectious disease prevention, detection, and management (ASTHO, 2005). This chapter is concerned with the design principles, technical foundation, messaging and data standards, and existing and future communication technologies for health alerting and notification. We will describe the architecture and operational experience of the New York State Department of Health (NYSDOH) as it developed an integrated health alerting and notification system (IHANS). We will also discuss lessons learned from our experience using the system for alerting and notification in health emergency events and exercises.

After studying this chapter, readers should understand the importance of an IHANS in all-hazard emergency preparedness and response. They should be able to describe the main design objectives of an IHANS and its functional and technical requirements. Furthermore, we hope readers will gain an understanding of the architecture and components of an IHANS and how they can be used to integrate communication technologies with the important functions of unified messaging, controlling access, and providing users with applications and data. To this end we provide, in Sect. 3.1, detailed discussions of system architecture, the unified messaging concept, communication directories, and messaging standards. These technical sections conclude with a review of the various communication methods needed to ensure delivery and receipt of health alerts in a timely manner; incorporating several communication technologies (phone calls, e-mails, faxes, etc.) into an alerting system helps to overcome the limitations of each individual method. No less important, as described in Sect. 3.2, is the framework for a wide range of emergency data exchange standards to support operations, logistics, planning, and finance that has arisen to ensure data sharing among various systems and jurisdictions.

Following these sections on system design, we conclude this chapter with a review of New York State's experiences both in developing and utilizing an IHANS (Sect. 4) and a discussion of lessons learned over the years of its operation (Sects. 5 and 6).

2. AN INFRASTRUCTURE FOR HEALTH ALERT AND NOTIFICATION SYSTEMS

The West Nile virus outbreak of 1999 and anthrax attacks of 2001 emphasized the importance of the nation's ability to prepare for and respond to bioterrorism and public health emergencies (Koblentz, 2003).

State and local health departments play an important role in public health emergency preparedness and response (CDC, 2007a). However, most state and local public health agencies lack the capacity to respond effectively to threats from emerging infectious diseases and bioterrorism (Salinsky, 2002). The Centers for Disease Control and Prevention (CDC) and the National Association of County and City Health Officials (NACCHO) conducted a joint e-mail test in 1999 to determine how quickly they could contact local health departments in the event of a health alert or bioterrorist emergency, and only 35% of these messages were delivered successfully. Another survey by the CDC in 1999 determined that only 45% of local health departments had the capacity to send alerts by fax to laboratories, physicians, state health agencies, the CDC, and others; less than 50% had high-speed continuous access to the Internet; and 20% lacked e-mail capability (Baker, 2005).

To address these deficits in health emergency communication, in 1999 the CDC began implementing the Health Alert Network (HAN) initiative to ensure that state and local health departments have rapid and timely access to emergent health information; a cadre of highly trained professional personnel; and evidence-based practices and procedures for effective public health preparedness, response, and service on a 24/7 basis (CDC, Health Alert Network). The HAN program significantly improved the communication infrastructure for response to bioterrorism and other health emergencies. Although the HAN was originally created as a national system, state and local HANs have become extremely important in responding to terrorism and urgent health threats (Baker and Porter, 2005). In 2003, 89% of local health agencies had developed continuous high-speed Internet access, more than twice the coverage (less than 40%) in 1998 (NACCHO, 2003). In 2007, all state public health departments could receive and evaluate reports of urgent health threats 24/7/365, whereas in 1999 only 12 could do so (CDC, 2008b). In addition, health alerting and notification has been included as one of the bioterrorism and public health preparedness functions, and its implementation using identified standards has been described in the Public Health Information Network Functions and Specifications, version 1.2 (CDC, 2002).

With support for the Health Alert Network program and other initiatives from the CDC, the NYSDOH has developed an enterprise information infrastructure for secure data communication over the Internet. This system was used successfully to deploy an integrated surveillance system for West Nile virus statewide within 3 months of the outbreak (Gotham et al., 2001). Utilizing this infrastructure, the NYSDOH developed an Integrated Health Alerting and Notification System (IHANS) for New York State. IHANS was designed to be fully compatible with the Public Health Information Network (PHIN) Requirements, version 2.0 (CDC, 2007b) and PHIN Preparedness Partner Communications and Alerting Functional Requirements, version 1.0 (CDC, 2005) and to meet requirements of PHIN Preparedness Key Performance Measures, version 1.0 (CDC, 2008a). IHANS is also compatible with Common Alerting Protocol (CAP) standards, version 1.1 (OASIS, 2005a). IHANS provides rapid dissemination of health alerts and communications to public health personnel and partners using multiple channels, including phone, cell phone, pager, e-mail and fax, with selective distribution based on urgency and sensitivity of the message (Loonsk et al., 2005).

3. REQUIREMENTS FOR A HEALTH ALERT AND NOTIFICATION SYSTEM

3.1 System Architecture

The purpose of an IHANS is to receive, process, manage and disseminate alerts, advisories, informational messages and other information for the public health agency and its partners, which may include but are not limited to local health departments (LHDs), healthcare providers and personnel in emergency preparedness and response. In addition to a unique problem that exists in the development of public health systems – that is, multiple local, state and federal jurisdictions need to operate in concert (Yasnoff et al., 2001) – the increasing need for health alerting from not only different program areas within a public health agency but also external partners, such as LHDs, with various levels of requirements adds significant complexity to the design and development, beyond that faced by traditional information systems. Due to the simultaneous needs for rapid receipt and dissemination of health alerts and communications using multiple channels of distribution, the system design should be based on the concept of unified messaging. This is the integration of several different communications media, allowing users to retrieve and send voice, fax, and e-mail messages from a single interface, whether it is a land-line phone, wireless phone, PC, or Internet-enabled PC (Jones, 2004).

A critical success factor of a successful IHANS is an up-to-date directory of public health contact information for key officials with whom the agency must communicate, and from whom the agency must receive communications.

A public health directory is intended to be a central repository of accurate public health contact information used for general business communications needs or for sending critical alerts and notifications in emergency circumstances. This directory's design should follow functional requirements and technical specifications in the PHIN Functions and Specifications, version 1.2, in particular function 7: Directories of Public Health and Clinical Personnel (CDC, 2002), and PHIN Preparedness Partner Communications and Alerting Functional Requirements, version 1.0 (CDC, 2005) from the CDC. With the final goal being to create an integrated health alerting and notification process among various programs with different business needs, while allowing them to take advantage of the agency enterprise communications infrastructure, it is critical to have a feasible approach to integrating health alerting functionalities in heterogeneous applications.

Extensible Markup Language (XML) is playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere (World Wide Web Consortium, 1999); it is the ideal messaging protocol to create and deploy flexible alerting integration solutions based on message-oriented middleware (MOM) architecture. Application integration may be achieved using asynchronous message queues, which draws on the features and strengths of using XML as the data representation standard (Mitch, 2006). In our implementation at the NYSDOH, this approach has been proven to be very effective because the ability to communicate via HTTP and parse XML documents is available at all major applications and the use of XML as a document-based messaging model allowed loosely coupled relationships among applications. Not only has it improved operational security, coordinated alerting activities and simplified technical operations without requiring invasive modification and service interruption of existing programs, but it has also significantly reduced costs as each program does not have to develop its own alerting application.

The NYSDOH IHANS architecture shown in Figure 17-1 describes the components that were used to integrate communication technologies into the alerting process with unified messaging, access, applications and data.

The system architecture in Figure 17-1 consists of five main layers: unified access, business logic, communication directory, communication systems and enterprise infrastructure. These are shown horizontally and connected by two vertical elements: XML-based Messaging Framework and Intelligent Load Balancing and Failover. Building foundation for this system are dedicated programs and services that provide guidance and support.

Within this architecture, via unified access by either Internet or phone to the system, communication systems link contact data from the communication directory with business logic for health alert and notification, taking advantage

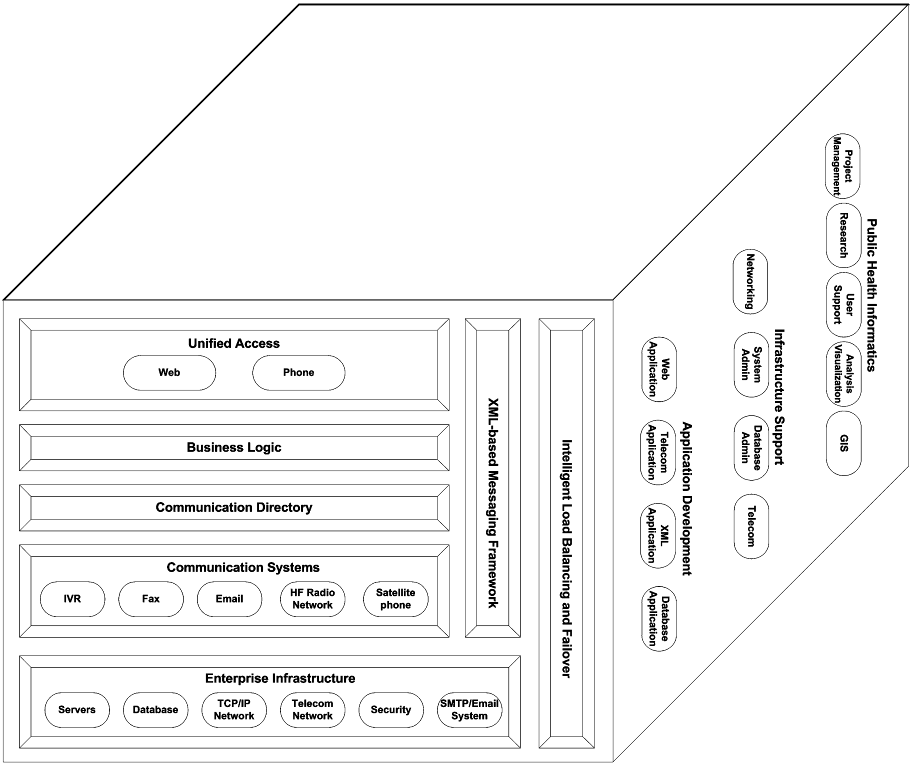


Figure 17-1. NYSDOH integrated health alert and notification system architecture.

of the enterprise infrastructure, including the underlying telecommunication network. The architecture is open, using industry standards wherever possible that enables the system to be readily integrated into a mixed platform environment. The XML-based messaging framework provides an environment that allows other programs to create new alert-enabled public health information technology applications. The intelligent load balancing and failover element provides unified high availability and system redundancy across all layers. The design of this system is guided by the public health informatics element, especially creation of technical specifications, integration of various technologies and development of data standards, policy and quality control solutions. The development, implementation and maintenance of this system are supported by application development and infrastructure support elements with talented application developers and dedicated 24/7 support services.

3.1.1 System Components

One of the main functional requirements of health alerting and notification that was implemented in the NYSDOH IHANS is the ability to send information via specified message format that allows recipients to view the message via a secure, Web-based interface. It also has the ability for threaded communication using a secure discussion forum. Message notification is delivered immediately and simultaneously through multiple methods of communication including phone, fax, numeric pager and e-mail. The alert message is directed to appropriate recipients according to role assignment in the NYSDOH Communication Directory (ComDir), described below. This system allows the alert to be sent either interactively from the NYSDOH IHANS, which is a Web-based interface for generating manual alerts, or automatically from various applications, such as electronic laboratory reporting of test results to LHD officials in response to pre-defined rules and triggers.

Various information technology and telecommunication methods under principles of public health informatics have been used in the development of this system to receive, process, manage and disseminate routine and emergency communications with public health partners and personnel (Figure 17-2).

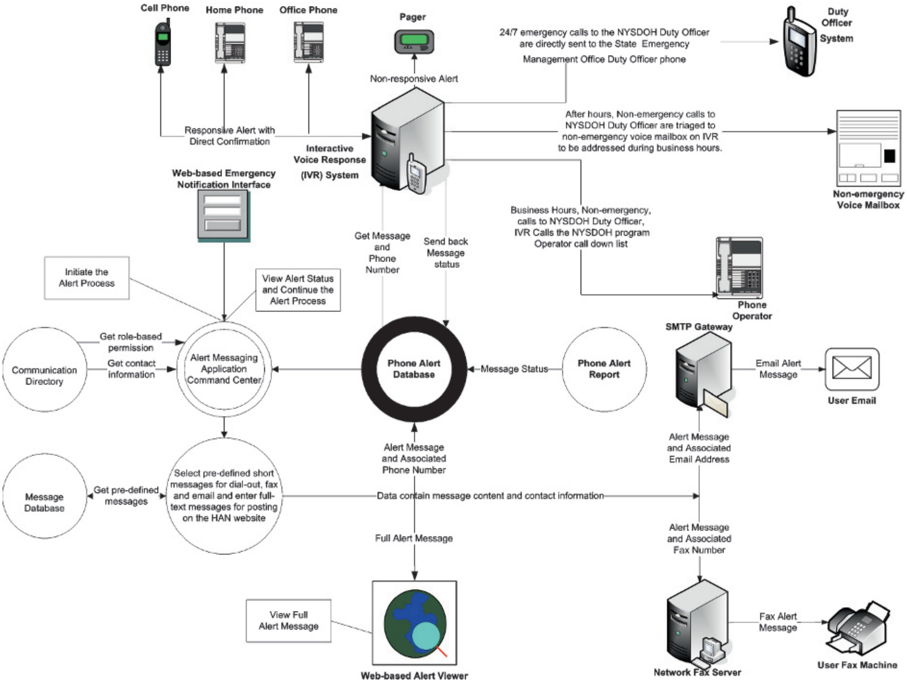

Cell Phone
Home Phone
Office Phone
Pager
24/7 emergency calls to the NYSDOH Duty Officer are directly sent to the State Emergency Management Office Duty Officer phone
Duty Officer System
Non-responsive Alert
Responsive Alert with Direct Confirmation
Interactive Voice Response (IVR) System
After hours, Non-emergency calls to NYSDOH Duty Officer are triaged to non-emergency voice mailbox on IVR to be addressed during business hours.
Non-emergency Voice Mailbox
Web-based Emergency Notification Interface
Business Hours, Non-emergency, calls to NYSDOH Duty Officer, IVR Calls the NYSDOH program Operator call down list
Get Message and Phone Number
Send back Message status
Initiate the Alert Process
View Alert Status and Continue the Alert Process
Phone Operator
SMTP Gateway
Communication Directory
Get role-based permission
Get contact information
Alert Messaging Application Command Center
Phone Alert Database
Message Status
Phone Alert Report
Email Alert Message
User Email
Alert Message and Associated Phone Number
Alert Message and Associated Email Address
Message Database
Get pre-defined messages
Select pre-defined short messages for dial-out, fax and email and enter full-text messages for posting on the HAN website
Data contain message content and contact information
Full Alert Message
Alert Message and Associated Fax Number
View Full Alert Message
Fax Alert Message
Web-based Alert Viewer
Network Fax Server
User Fax Machine

Figure 17-2. Data flow diagram for NYSDOH health alert and notification system.

3.1.2 Unified Messaging Concept

Function 8: Public Health Information Dissemination and Alerting of the PHIN Functions and Specifications, version 1.2 (CDC, 2002) and PHIN Preparedness Partner Communications and Alerting Functional Requirements, version 1.0 (CDC, 2005) require immediate distribution of messages through one or more mechanisms (phone, e-mail, fax or pager) and the ability to use standard message format and vocabulary. The concept of unified messaging provides a powerful and flexible solution to meet these requirements. Although modern communication technologies have overcome distance and time barriers so that people can communicate in real time or near-real time no matter where they are, other limitations must be overcome, such as the use of incompatible devices and forms of communication (International Engineering Consortium, 2005). The unified messaging concept addresses these limitations so that the system can communicate to anyone, anywhere, at anytime using different technologies. Unified messaging also allows an open and nonproprietary system architecture; although each component may not be from the same vendor, all can work together by industry standards.

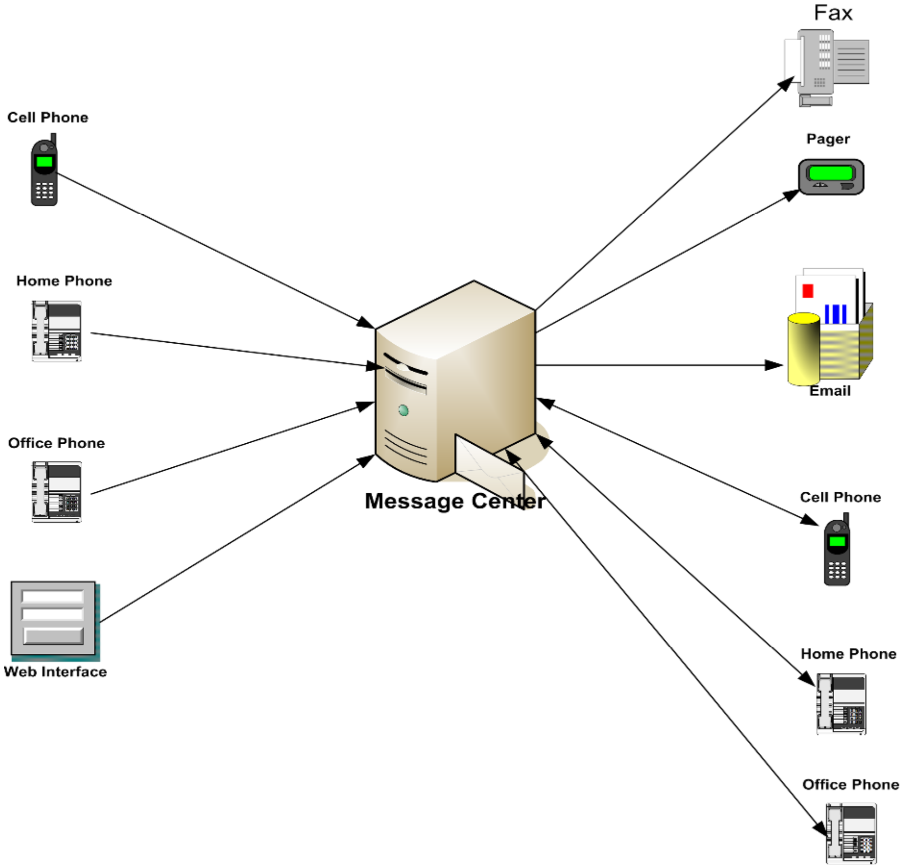


Figure 17-3. Unified messaging concept.

Within the NYSDOH IHANS, each communication component (phone, fax, e-mail, and secure Internet) was developed on a different platform, but they were all integrated successfully using unified messaging. The system has the capacity for singular delivery – that is, any message can be sent to a particular user – because it utilizes the following: a single alias; a single repository where all messages are stored for easy access/retrieval and integration, such as converting text to speech; a single point of access, where the user can find all messages through a central, intuitive Web interface; and multiple communication routes, allowing a message to be sent to the user via different methods (see Figure 17-3).

3.1.3 NYSDOH Communication Directory

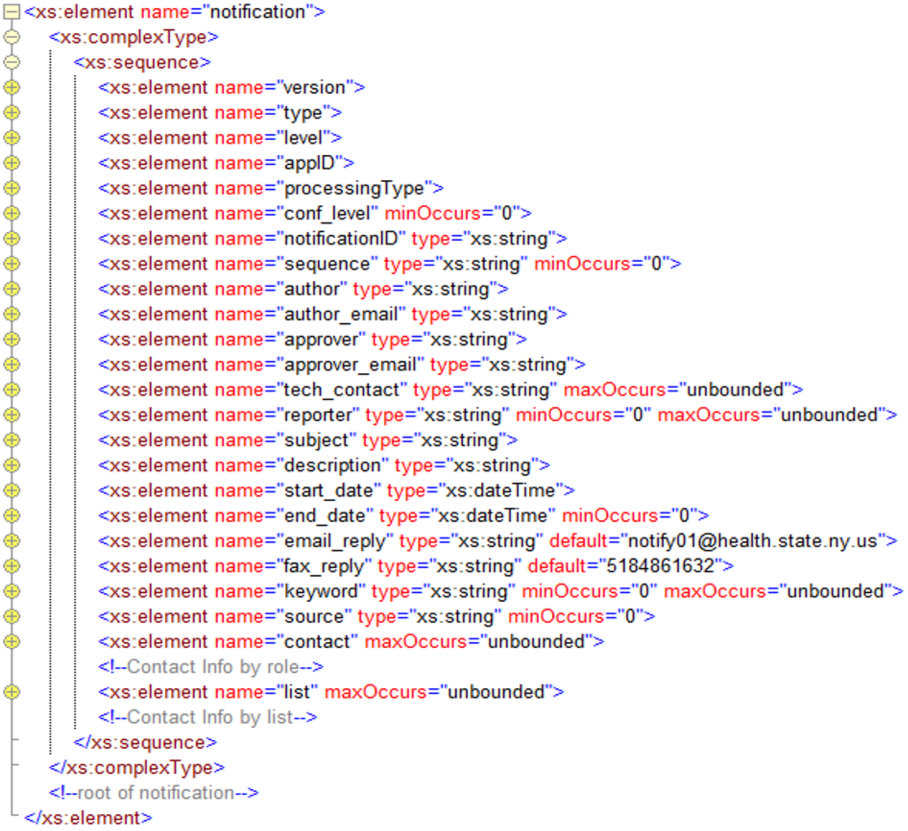
Based on the concept and functional requirements of a public health directory, the NYSDOH's Health Commerce System (HCS) ComDir is the central repository of each user organization's critical role assignments. It is a self-maintained, person-based, up-to-date listing of business and emergency contact information that is used, via integration with IHANS, for broadcasting notifications at various urgency levels using multiple modes of communication and device types. ComDir maintains three categories of roles: "Emergency" and "Business" office contacts designed to reach locations within an organization and "Person Contacts" designed to reach directly those individuals serving in specific roles or job functions in a given organization for whom the information in the notification would be most relevant (e.g., the Commissioners of Health at LHDs for an emergency operations alert or Infection Control Practitioners at hospitals for an alert about a potential disease outbreak). Role structure is flexible; there are currently over 700 roles customized specifically to the business of the nearly 200 different types of HCS user organizations. The literature stresses the importance of gathering and maintaining necessary contact information during intervals of everyday operations between emergencies, so that the directory is complete, accurate, and accessible when needed (Auf der Heide, 1989; National Science and Technology Council, 2000; CDC, 1999, 2000).

The communication directory contains categorized contact information, roles, and communication devices for every organization; access to this information is controlled by rules associated with the requestor: as an individual user, or as the holder of a role in the organization, or as an electronic data system. ComDir provides the ability to query and search contact information by person name, role, organization, organization type, and jurisdiction. As required by the PHIN Preparedness Partner Communications and Alerting Functional Requirements, version 1.0, ComDir also enables users to prioritize the dialing/contact sequence of their emergency contact numbers for both business and after hours (CDC, 2005).

In addition, this directory is used to support the authorization of access to different data systems within the NYSDOH HCS. It is fully capable of supporting the directory exchange of data with partners using standardized data exchange formats (LDIF) and protocols to support partner communications such as Directory Services Markup Language (OASIS, 2002).

3.1.4 Data Integration and Communication Using XML Messaging

As the need for health alerting increases – not only in the NYSDOH but also in its public health partners, such as local health departments and other state agencies with different messages and target audiences – it is extremely important to have transparent, secure and guaranteed integration of data for alerting across multiple programs. For this purpose, an XML alert messaging protocol and an alert messaging component was developed to provide a single interfacing solution for public health information systems requiring a health alerting function (see Figure 17-4).



```
<xs:element name="notification">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="version">
      <xs:element name="type">
      <xs:element name="level">
      <xs:element name="appID">
      <xs:element name="processingType">
      <xs:element name="conf_level" minOccurs="0">
      <xs:element name="notificationID" type="xs:string">
      <xs:element name="sequence" type="xs:string" minOccurs="0">
      <xs:element name="author" type="xs:string">
      <xs:element name="author_email" type="xs:string">
      <xs:element name="approver" type="xs:string">
      <xs:element name="approver_email" type="xs:string">
      <xs:element name="tech_contact" type="xs:string" maxOccurs="unbounded">
      <xs:element name="reporter" type="xs:string" minOccurs="0" maxOccurs="unbounded">
      <xs:element name="subject" type="xs:string">
      <xs:element name="description" type="xs:string">
      <xs:element name="start_date" type="xs:dateTime">
      <xs:element name="end_date" type="xs:dateTime" minOccurs="0">
      <xs:element name="email_reply" type="xs:string" default="notify01@health.state.ny.us">
      <xs:element name="fax_reply" type="xs:string" default="5184861632">
      <xs:element name="keyword" type="xs:string" minOccurs="0" maxOccurs="unbounded">
      <xs:element name="source" type="xs:string" minOccurs="0">
      <xs:element name="contact" maxOccurs="unbounded">
      <!--Contact Info by role-->
      <xs:element name="list" maxOccurs="unbounded">
      <!--Contact Info by list-->
    </xs:sequence>
  </xs:complexType>
  <!--root of notification-->
</xs:element>
```

Figure 17-4. Diagram of XML schema for alert messaging protocol.

It is important to realize that the role of this component is not to initiate a health alert but rather to provide a framework that combines data security, reliability, and interoperability of communication standards, such as standard vocabulary and consistent meanings. It allows NYSDOH programs, public health partners, and their technical staff to create consistent, robust and interoperable alerting ability with minimal modification and disruption of service. The system guarantees delivery of messages by queuing messages and can be configured to automatically resend messages; it also keeps all messages in persistent storage until the message is sent out. Another advanced feature of this system is advanced archiving; when a message is processed, it is automatically archived in a database for viewing and extraction at a later date. Scheduled alerting is also made possible in this system, which can send alerts out at specified dates and times. For example, a bed availability survey alert is sent out every Tuesday at 8:00 p.m.

3.1.5 Communication Methods

Emergency notification by phone is a variation of voice broadcasting, which is a recent mass communication technique that broadcasts phone messages to hundreds or thousands of call recipients at once. In a state-of-the-art voice broadcasting system, the delivery of alert messages is not only timely but also confirmed: a "press through" feature allows the call recipient to listen to the message and take an action by pressing a touch-tone phone key to confirm receipt. If the system detects a no-answer condition or busy signal, the calling business rules direct the system to retry the call up to three times, sequentially integrating attempts to other phone numbers prioritized in the person's contact record; if it detects an answering machine, the recipient is instructed to view the complete content of the notification posted on the HCS Website; however, the built-in dialing logic calls for the full set of the user's contacts to be called until one of two conditions exists: the user confirms the receipt of the call or three attempts have been made to reach each contact number the user has in the directory. While our alerts are mainly initiated from a Web-based application, the interactive voice response (IVR) system also provides alternatives using several "canned" XML message formats. The alert messages can be either pre-recorded or live-recorded by the IVR, or they can be left, via a Text-To-Speech (TTS) tool, as computer-generated synthesized speech with real voices.

Another method of emergency communication is fax broadcasting using a network digital fax server. When a fax alert message is generated, it is sent to the Simple Mail Transfer Protocol (SMTP) gateway of the fax server in the form of an e-mail message with optional attachments, which will be converted into fax format and transmitted. Using digital T1 lines, the system supports a high volume of fax traffic with a conservatively estimated capacity of 3,168 pages/h. Other features that significantly enhance the alert process include Error Correction Mode (ECM), which can correct errors in received images caused by phone line noise, guaranteeing that faxes will be sent and received correctly, as long as the sending or receiving machine also supports ECM; use of specific cover pages for different types of notifications; incorporation of a Fax Archive service with offline, long-term mode or continually active online mode; system backup and restore function for disaster recovery; and a variety of system-wide and individual notification reports.

Electronic mail (abbreviated "e-mail" or, often, "email") is a store and forward method of composing, sending, storing, and receiving messages over electronic communication systems. For emergency notification, our system uses SMTP for Internet-based e-mail to recipients outside the NYSDOH and Notes remote procedure call (NRPC), the Lotus Notes Domino network protocol for e-mail to recipients within the organization. Since e-mail messages are generally not encrypted and relatively easily intercepted and read, the e-mail notification contains only nonsensitive information with instructions for recipients to log on to the NYS Health Commerce System to read the full notification on the Web-based Notification Viewer.

Because of the sensitive nature of emergency notification, the full-text messages are posted on a secure notification viewer on the NYS HCS. Although recipients are notified of the location of each newly posted message by phone, fax, and/or e-mail, access to the posted messages may be open to all HCS users or only to target audiences. Posted messages can be deleted by the system administrator and be sorted by either *date posted, audience type,* or *keyword*.

3.2 Standard Alert Message Distribution Framework for Data Sharing Among Emergency Information Systems

As the need has grown for information sharing and data exchange across the local, state, tribal, national and nongovernmental organizations that provide emergency response and disaster management services, it has become clear that scientifically-based technical standards – including common communication and data standards – are critical to the nation's ability to prepare for, prevent, respond to, and recover from emergency incidents (DHS, 2004). Following closely the development of a growing suite of specific message standards developed by federal agencies and standard governance organizations, the NYSDOH has designed a "system to system" data communication interface using public standards, including the Common Alerting Protocol (CAP) data interchange standard (OASIS, 2005a) and the

Emergency Data Exchange Language (EDXL), a broad initiative to create an integrated framework for a wide range of emergency data exchange standards to support operations, logistics, planning and finance (OASIS, 2005b). For emergency notification, our system can utilize CAP both as a stand-alone messaging protocol and as a payload for EDXL messages.

4. CASE STUDY

The NYSDOH IHANS has been used extensively since its inception. Every user of the system is able to receive alerts and lower-level notifications by virtue of being assigned to a role targeted in that notification or by being included in a ComDir list of recipients for that alert. Additionally, there are currently 625 trained and certified users of the IHANS application who are authorized to send alerts and notifications to the HCS users applicable to their jurisdiction. A large number of notifications have been sent though the system that originated from not only the NYSDOH but many other sources, including the CDC, LHDs, New York City Department of Health and Mental Hygiene (NYCDOHMH) and NYS Department of Homeland Security (see Table 17-1).

Since the launch of the system in 2002, the usage has increased significantly every year, which shows a remarkable success of this system. It also demonstrates the ability of the system to handle a large volume of notifications, which we attribute to its sound design concepts, innovative system architecture and strong technical infrastructure (see Table 17-2).

Table 17-1. Usage of the integrated health alert and notification system in New York.

Notification	Notification Source						Total
Type	CDC	LHD	NYCDOH	NYSDHS	NYSDOH	Other	
Advisory	107	18	17	44	198	19	403
Alert	7	17	20	0	136	8	188
Drill	9	436	11	4	405	227	1,092
Informational message	7	17	16	1	385	13	439
Update	41	0	3	0	7	2	53
Total	171	488	67	49	1,131	269	2,175

CDC Centers for Disease Control and Prevention, *LHD* Local Health Department, *NYCDOHMH* New York City Department of Health and Mental Hygiene, *NYSDHS* New York State Department of Homeland Security

Table 17-2. Number of notifications by year, as of 12/31/2006.

2002[a]	2003	2004	2005	2006
2	170	293	254	1,536

[a] The system was operational on November 2002

Figure 17-5 shows the usage of the NYSDOH IHANS by type of notification and source of notification over time since the beginning of the system.

The success of an emergency notification depends not only on the ability of the system to send or receive/forward messages rapidly, it also largely relies on the ability of the notification recipients to acknowledge and act on information in the message. Especially important for these functions is a distributed model of system use across jurisdictions, so the system can be tailored for the parameters of the jurisdiction currently sending a notification. Metadata, such as keyword, approver of the notification, document templates and source, can be customized to act also as a local notification system, enhancing the credibility of notifications sent at each jurisdictional level and the usefulness of the system across the multi-jurisdictional base of user organizations. Distributed use of the IHANS adds to its embeddedness in the overall public health community in NYS and increases user familiarity and ownership of the system, creating an environment for ongoing enhancement of its usefulness.

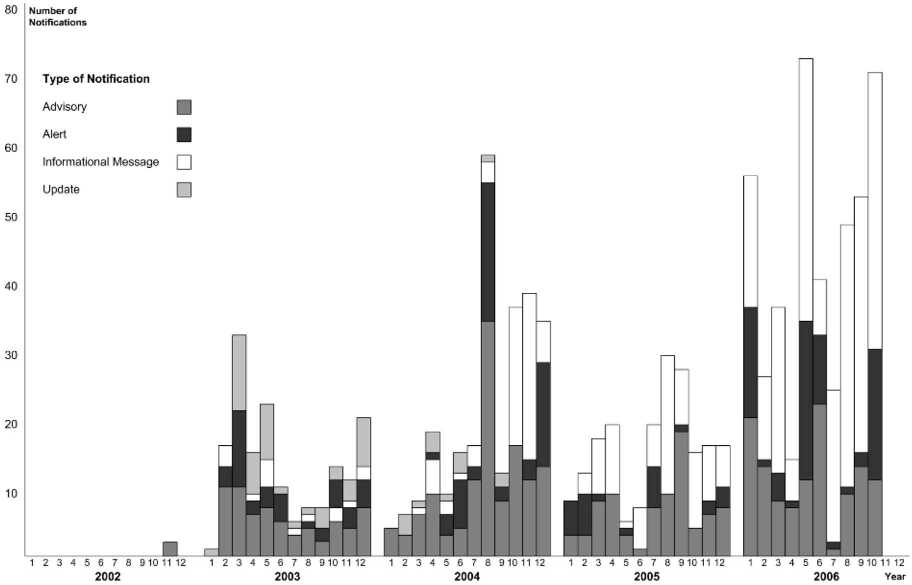


Figure 17-5. Usage of New York State integrated health alert and notification system by type of notification, by month and year.

A set of examples of system use illustrate this point. Key to measuring system success is the timeliness and completeness of message receipt by targeted recipients. Ensuring that the information gets through to the right people at the right time is essential in responding to disease conditions or other urgent public health business.

LHDs are key targeted recipients of IHANS communications sent by the NYSDOH or other LHDs. Beginning in January 2005, NYS began testing the use of its IHANS system with the CDC during its annual State Epi-X alert proficiency test drills. The purpose of these drills is to test the ability of Epi-X to transmit an alert regarding a disease situation to its registered user community. The NYSDOH agreed with CDC that the Epi-X alerts would be sent to a NYSDOH Public Health Preparedness warning point, and that the alert would be further cascaded to the appropriate NYS recipients via the NYSDOH IHANS system. In this way, upkeep of redundant sets of contact information by each agency is eliminated and notification protocols are harmonized to one system, rather than multiple, for the end user/recipient. The mode of cascading the alert from one agency to the other also allowed us to measure NYSDOH readiness to receive and act on an alert from another agency. Table 17-3 summarizes the series of annual drills conducted in January 2005, January 2006, December 2006, and December 2007, respectively referred to hereafter as Year 1, Year 2, Year 3 and Year 4.

Table 17-3. Local Health Department performance during annual CDC Epi-X alerting proficiency tests.

Date/time Epi-X alert received	1/14/05[a]; 11:118 a.m.	1/26/06; 9:15 a.m.	12/01/06; 8:46 a.m.	12/27/07; 10:58 a.m.
Time elapsed: NYSDOH receives and cascades Epi-X alert to LHDs via IHANS (minutes)	15 (11:33 a.m.)	14 (9:29 a.m.)	13 (8:59 a.m.)	10 (11:08 a.m.)
Drill start time awareness	A	U	U	U
Percent of LHDs responding (58 including NYCDOHMH)	58/58; 100%	58/58; 100.0%	54/58; 93%	57/58; 98%
Mean response time (minutes)	39.6 No outlier	W/o outlier: 7.3 With outlier: 11.1	W/o outlier: 9.5 With outlier: 11.9	W/o outlier: 9:38 With outlier: 12.36
Standard deviation (minutes)	7 No outlier	W/o outlier: 9.7 With outlier: 30.4	W/o outlier: 5.61 With outlier: 17.96	W/o outlier: 3.78 With outlier: 22.58
Modal response time (minutes)	39	2	8	7
Min. response time (minutes)	2	2	3	7
Max. response time (minutes)	64 No outlier	W/o outlier: 65 With outlier: 227	W/o outlier: 29 With outlier: 135	W/o outlier: 30 With outlier: 176

W/o without, *A* announced, *U* unannounced

[a] The January 2005 test was actually for the 2004 CDC Epi-X Proficiency Test cycle, which was delayed from December 2004. That delay pushed the 2005 next annual test to January 2006. To get the annual test cycle in sync, NYSDOH and CDC conducted a second 2006 drill in December of that year and now conducts their annual tests in December of the test year since that time.

Table 17-3 shows the analysis of samples of LHD drill response data from each of the four Epi-X drills. These results demonstrate a clear capacity for the IHANS alerting process on the part of LHDs and the NYSDOH as well as an ongoing improvement in that capacity and support for the cascade alerting model. Though the NYSDOH turnaround time of 15minutes from receipt of the Epi-X alert to sending the cascade notification to LHDs is acceptable by CDC standards (CDC, 2005), by the fourth year of drills that turnaround time was reduced by 33% to 10 min, and improvements were shown each subsequent year in between.

The Year 1 drill had the slowest mean response time by the LHDs of the four drills: 39 min, as opposed to the shortest mean response time in Year 2, which was 7.3 min. This represents an 81% reduction of the mean response time for the same group. Data from the Year 3 and Year 4 drills indicate that this improvement was persistent over time, with each drill demonstrating an approximate 75% reduction in response time from that seen in Year 1.

Additionally, examination of the data shows the Year 1 LHD sample as having the largest/slowest mean response time. After cleaning the data, (conservative removal of extreme outliers from Years 2–4 data sets with each outlier removed being >4 standard deviations from the mean), the Year 1 data showed no outliers and a fairly normal distribution around the higher mean. Examining the data in the remaining three drills discloses the narrowing of the variability in each sample when the single extreme outlier is removed from the calculations in each data set. Therefore, the LHDs were reducing their mean response time in subsequent years. Their individual performances were becoming more consistent with each other and closer to a much smaller/faster group mean response time. LHDs are also adept as users of the IHANS to send notifications. Sixty-four percent of the 625 HCS users certified and trained in using the IHANS notification sending tool are HCS users located at LHDs. Specifically for the purpose of maintaining LHD and NYSDOH skill capacities in use of the alerting and notification process, NYSDOH has conducted its own drills of IHANS with the LHDs during every grant year. During the summer of 2007, NYSDOH introduced the cascading process to LHDs and simultaneously drilled both their capacity to receive alerts effectively and to further disseminate the alert message among their own local contacts using the IHANs system. An additional aspect of the process measured was that drills were conducted both during business and after hours, to ensure that the LHD after-hours readiness to respond is equally robust. Table 17-3 and Figure 17-6 demonstrate the LHD performance during these cascading drills.

Overall, LHDs performed well on the cascading technique, though the mean turnaround time of approximately 60 min for both business hours and

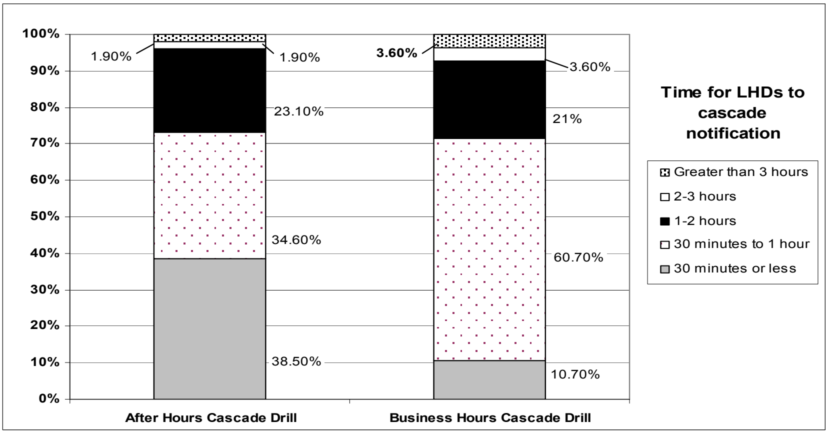


Figure 17-6. 2007 LHD cascading alert drill results.

after-hours drills, with inclusion of outliers (and approximately 48 min with removal of outliers) does show a need for improvement to comply more closely with CDC standards. At the time of writing, a second round of cascading drills with the LHDs is currently being developed for the summer of 2008, which will examine progress made toward improving the cascade capacity and achieving the goal of reducing that turnaround time to comply with CDC standards.

The NYSDOH ComDir collects contact information for the nearly 70,000 HCS users. Alerting and notifications using IHANS is being rapidly adopted throughout the agency for reliable communications regarding disease outbreaks, early warnings and distribution of guidelines and protocols with their colleagues at healthcare facilities. Hospitals are very active users of the HCS system and receive alerts and other notifications that are both manually- and system-driven through IHANS and the IVR. Recently, hospitals have been drilled specifically for readiness to act on a notification, including a surprise, after-hours alert.

The first drill was conducted to alert 147 NYS hospitals – Hospital Resources and Services Administration (HRSA) grantees – at 9:00 a.m. and have them immediately respond (within 1 h) to report data to the NYSDOH in the HCS Health Emergency Response Data System (HERDS) regarding their capacities of available bed types. Overall, hospitals performed very well and showed a very good level of preparedness in both receiving and confirming the HERDS activation alert message and especially in the critical capability of responding to the alert message with the required action of entering their data into HERDS within the 1 h required time interval. Approximately 96% confirmed receiving the alert message within 3 h of the notification being sent. Approximately 89% of the hospitals met the exercise deliverable of entering data into the survey within 1 h of being notified to do so. Ultimately, 99% of the hospitals provided their data to the HERDS system survey. Figure 17-7 illustrates hospital progress over the course of the drill in confirming receipt of the alert following its initiation.

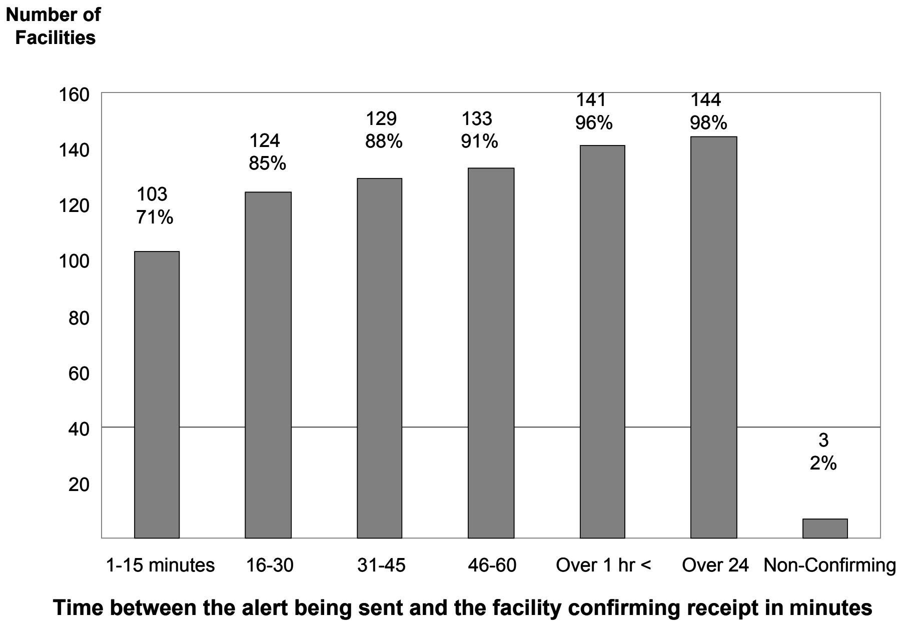


Figure 17-7. 2007 HAvBED drill facility response to alert.

The second drill called for hospital action, again to HERDS, but was sent during the evening to night hours shift (initiated at 20:18). Figure 17-8 illustrates the timeliness of the notification delivery and confirmation in responding to the alert, which was sent at the most urgent notification level by both phone and e-mail to ten different ComDir roles at 236 hospitals across the state. The total number of individuals contacted after de-duplication was 2,507. Of the individuals confirming receipt of the alert, 1,034 confirmed via e-mail and 2,030 individuals confirmed via phone, such as home phone and cell phone, given the late hour of the notification. Overall, approximately 75% of hospitals confirmed receipt of the notification within the first 3 h following the notification.

- 232/236 hospitals confirmed receiving the alert message.
- 146/147 HRSA grantee hospitals confirmed receiving the alert message.
- 155 hospitals logged onto HERDS as follow-up to the notification, as did the NYS Office of Homeland Security, LHDs and NYSDOH offices.

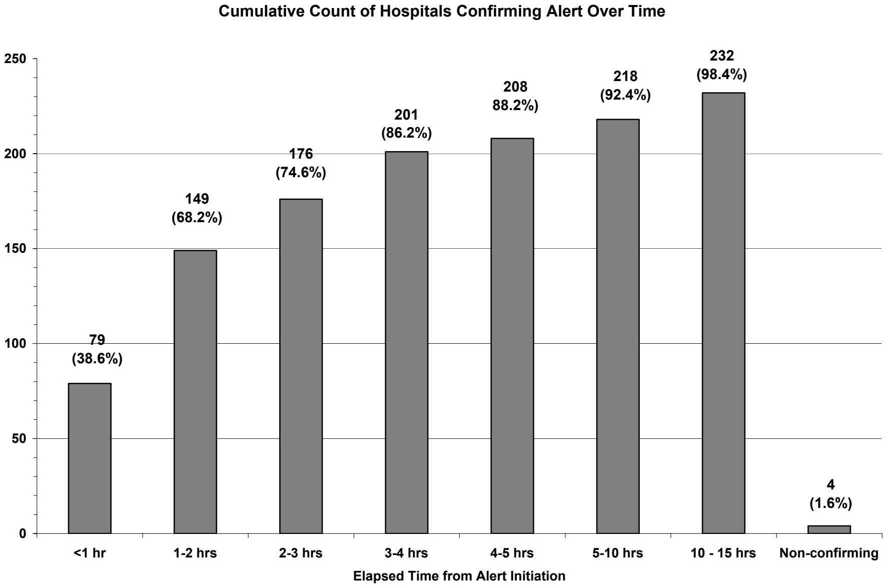


Figure 17-8. Hospital response to after-hours drill.

5. DISCUSSION

From our experience in the design and operation of New York State's Integrated Health Alert and Notification System, each current electronic communication technology has its own limitations.

Emergency notification by voice broadcasting relies heavily on answering machine detection technology and the logic to play a unique message to answering machines without message truncation; this technology has approximately a 10–15% failure rate. Operation of our voice broadcasting system is largely dependent on the traditional public switched telephone network (PSTN), which has matured and is typically reliable but still subject to outages, such as the Northeast Blackout in 2003 has showed (Beatty et al., 2006).

Fax broadcasting has its own deficiencies, including dependency on the operational status of recipient fax machines, late or lost faxes at the recipient site, poor document quality and transmission quality.

E-mail is a very old technology, which predated the inception of the Internet. Reliability is a serious issue as e-mail messages have to go through intermediate computers before reaching their destination, creating multiple possible points of failure. Another risk factor is unsolicited commercial e-mail (spamming) that can result in information overload for many computer systems. Security is another serious issue, as e-mail messages are generally unencrypted and SMTP has no ability for authentication of senders.

In our system design, the unified messaging concept allows the system to minimize the inherent limitation of current technologies and maximize the probability that the recipient will receive the notification by at least one communication pathway. Nevertheless, future development of similar systems should take advantage of new and emerging technologies to address these issues. For example, Secure/Multipurpose Internet Mail Extensions (S/MIME) can be used for end-to-end message encryption.

New alternative communication technologies should also be considered to ensure continuity of alert operation in disaster situations and improve message delivery capacity; examples are satellite phones (mobile phones that communicate directly with orbiting communications satellites) and high-frequency radio networks (which allow for both short- and long-distance voice and data communications). Distributing full content of notification messages can be made more active and timely by using Web feeds (a data format used for serving users frequently updated content). Threaded discussion for each notification would provide more interactive post-alert communication and collaboration opportunities for a broad audience, including notification recipients and senders. In an emergency situation, ability for secure instant messaging (IM), which is a form of real-time communication between two or more people based on typed text, will allow real-time, presence-enabled and easy collaboration among a group of users or direct communication between individual users.

With the introduction of multiple electronic communication technologies into an integrated health alert and notification system, it is important that a message and data integration framework is developed to ensure the consistency and accuracy of the message as well as target audiences.

A uniform, all-technologies, all-hazards messaging standards with the essential features for both existing and emerging alert systems and technologies is essential for interoperable health alert and notification among public health preparedness and emergency response communities.

6. CONCLUSIONS

In this chapter, we presented the system design, supporting technical concepts and informatics standards, and existing and future electronic communication technologies for an IHANS and its implementation at the NYSDOH. These ideas can be used to develop similar systems elsewhere. While it has been a consensus among public health agencies that development of stovepipe information systems cannot be continued, it has become more important than ever to integrate public health emergency preparedness functions, including health alerting and notification, as funding to states and localities to maintain and improve their preparedness is declining (Trust for America's Health, 2007).

However, federal and state governments will continue to provide funding categorically because of the political and governmental process. Although it is well recognized that an integrated system design and implementation alone cannot provide an inclusive solution to the current silo systems, it offers an effective way to provide shared alerting and notification capability. At the same time, implementing an integrated health alert and notification system can be problematic.

We were able to overcome several technical and programmatic challenges in the implementation at the NYSDOH with support from department executives, well-defined business rules, an effective development plan, state-of-the-art technical infrastructure, highly talented staff, and active program involvement. It is understandable that many other public health agencies might not have the same capabilities; these should consider more conservative approaches, including using a service provider for less critical or costly functionalities.

As described in Chapter 18, attaining familiarity with a system and dual use of a system, both for everyday operations and emergency operations, is a key component of a successful informatics model for integrating data systems and for the concept of "readiness to respond." This dual use is key to the success of the IHANS system, and data described in Table 17-1 demonstrate that dual use of IHANS for a variety of notifications every day, from informational messages to highest level alerts. While readiness to respond can be enhanced by system refinements, it is only ongoing practice that makes a system second nature to its users, so that during an emergency situation actions can be taken without hesitation or delay. The drill results presented here demonstrate that user experience over time has resulted in improvement in the response capacity of our constituents.

Implementation of this system over the years has demonstrated that the innovative approach to the system design and development has markedly increased the ability of the NYSDOH and its public health partners to exchange and react to secure, rapid and reliable health alerts and notifications. These findings suggest that not only the utilization of new and emerging technologies but also national messaging standards is needed to enable interoperable and effective emergency communication.

ACKNOWLEDGEMENTS

We gratefully acknowledge the leadership and support of Mr. Robert Burhans, director of the Health Emergency Preparedness Program, NYSDOH, and Dr. Dale Morse, Director of the Office of Science, NYSDOH. We acknowledge the invaluable contributions of Ronald Stamp, Carol Hirsch, Michelle Kosinski, Zhenwei Chen, Franklin Hsia, Barry Krawchuk, currently of NYSDOH, and the support of William Moyer, Ph.D., William Hulchanski, and Sean Kelly, formerly of NYSDOH, to the system design and development. We are also indebted to the hospitals and local health departments of NY State for their efforts invested in participation in the deployment and use of the system. Finally, we specifically acknowledge Kathryn Schmit for her thoughtful and excellent editing of this chapter.

This publication was supported by Cooperative Agreement Numbers U50/CCU223671 and U90/CCU216988 from the U.S. Centers for Disease Control and Prevention (CDC) and by Cooperative Agreement Number 6U3RHS05934 from the U.S. Health Resources and Services Administration (HRSA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or HRSA.

QUESTIONS FOR DISCUSSION

1. Why are health alert and notification systems important for all-hazard emergency preparedness and response?
2. List and describe the main design objectives and functional and technical requirements of an integrated health alert and notification system.
3. Describe the architecture and components of an integrated health alert and notification system and how they can be used to integrate communication technologies into the health alert and notification process with unified messaging, access, applications and data.
4. What is the unified messaging concept, and what is its importance in health alert and notification systems?
5. What is a public health directory? Describe its importance in health alert and notification systems.
6. Describe XML messaging and how it can be used for data integration and communication for health alert and notification.
7. Why did the NYSDOH decide to use XML messaging for data integration and communication for health alert and notification?
8. What are the typical communication methods for health alert and notification? How can they be used together to overcome limitations of any individual method?
9. What is the current standard alert message distribution framework for data sharing among emergency information systems?
10. What are the recommendations for design and implementation of an integrated health alert and notification system?

REFERENCES

Auf der Heide, E. (1989). Principles of preparation and coordination, The Center of Excellence in Disaster Management and Humanitarian Assistance; http://www.orgmail2.coe-dmha.org/dr/flash.htm.

Baker, E. L. (2005). The public health infrastructure and our nation's health. *Annual Review of Public Health* **26**:303–18.

Baker, E. L., and Porter, J. P. (2005). The Health Alert Network: partnerships, politics, and preparedness, *Journal of Public Health Management & Practice* **11**(6):574–576.

Beatty, M., Phelps, S., Rohner, C., et al. (2006). Blackout of 2003: public health effects and emergency response. *Public Health Reports* **121**: 36–44.

CDC (1999). Health Alert Network; http://www2a.cdc.gov/han/index.asp.

CDC (1999). Health Alert Network cooperative agreement guidance document.

CDC (2000). National electronic disease surveillance system cooperative agreement guidance document.

CDC (2002). Public health information network functions and specifications, version 1.2; http://www.cdc.gov/phin/library/documents/pdf/PHIN_Functions_Specifications_121802.pdf.

CDC (2005). Partner communications and alerting functional requirements, version 1.0; http://www.cdc.gov/phin/library/documents/pdf/PCA%20_RSv1.0.pdf.

CDC (2007a). Public health emergency response guide for state, local, and tribal public health directors; http://www.bt.cdc.gov/planning/pdf/cdcresponseguide.pdf.

CDC (2007b). Public health information network requirements, version 2.0; http://www.cdc.gov/phin/library/documents/pdf/111759_requirements.pdf.

CDC (2008a). Key performance measures, version 1.0 (Apr. 12, 2008); http://www.cdc.gov/phin/library/documents/pdf/KPM_RSv1.0.pdf.

CDC (2008b). Public health preparedness: mobilizing state by state (Apr. 12, 2008); http://www.emergency.cdc.gov/publications/feb08phprep/pdf/feb08phprep.pdf.

Gotham, I. J., et al. (2001). West Nile virus: a case study in how NY State health information infrastructure facilitates preparation and response to disease outbreaks. *Journal of Public Health Management and Practice* **7**:75–86.

Intel Corporation (2007). International 2- and 4-port voice processing boards (Mar. 21, 2007); http://www.intel.com/network/csp/products/3048web.htm.

Jones, S. (2004). The Basics of Telecommunications. International Engineering Consortium. Unified messaging; http://www.iec.org/online/tutorials/unified_mess.

Koblentz, G. (2003). Biological terrorism: understanding the threat and America's response. In *Countering Terrorism: Dimensions of Preparedness*, eds. A. M. Howitt and R. L. Pangi. Cambridge: The MIT Press.

Loonsk, J. W., McGarvey, S. R., Conn, L. A., et al. (2005). The public health information network (PHIN) preparedness initiative. *Journal of the American Medical Informatics Association* **13**(1):1–4, M1815.

Mitch, A. (2006). *XML Problem-Design – Solution.* New York: Wiley Publishing, Inc.

NACCHO (2003). National association of county and city health officials, local public health agencies better equipped to handle bioterrorist attacks. Res. Brief No. 8 (Jan. 2003). Washington, DC.

National Science and Technology Council (2000). Effective disaster warnings; http://www.sdr.gov/NDIS_rev_Oct27.pdf.

OASIS (2002). Organization for the advancement of structured information standards. Directory Services Markup Language (DSML) (Apr. 12, 2002); http://www.oasis-open.org/committees/dsml/docs/DSMLv2.doc.

OASIS (2005a). Organization for the advancement of structured information standards, common alerting protocol 1.1, OASIS Standard (Oct. 1, 2005); http://www.oasis-open.org/committees/download.php/14759/emergency-CAPv1.1.pdf.

OASIS (2005b). Organization for the advancement of structured information standards, EDXL-DE 1.0. OASIS Standard (Oct. 1, 2005); http://www.oasis-open.org/committees/download.php/18772/EDXL-DE%201.0%20Standard.pdf.

Salinsky E. (2002). *Public Health Emergency Preparedness: Fundamentals of the System*, National Health Policy Forum Background Paper. Washington, DC: George Washington University.

Trust for America's Health (2007). Ready or Not? Protecting the public's health from disease, disasters, and bioterrorism; http://www.healthyamericans.org/reports/bioterror07/BioTerrorReport2007.pdf.

U.S. Department of Homeland Security (2004). National incident management system (Mar. 1, 2004); http://www.fema.gov/pdf/emergency/nims/nims_doc_full.pdf.

World Wide Web Consortium (2010). Extensible markup language (XML); http://www.w3.org/XML/.

Yasnoff, W. A., Overhage, J. M., Humphreys, B. L., et al. (2001). A national agenda for public health informatics: summarized recommendations from the 2001 AMIA spring congress. *Journal of the American Medical Informatics Association* **8**(6):535–545.

SUGGESTED READING

Bates, R. (2006). *Voice & Data Communications Handbook*, 5th ed. Osborne: McGraw-Hill.

Mitch, A. (2006). *XML Problem – Design – Solution*. New York: Wiley Publishing, Inc.

O'Carroll, P. (2002). *Public Health Informatics and Information Systems*. New York: Springer.

Song, I. (2003). *Conceptual Modeling – ER 2003*. Berlin/Heidelberg: Springer.

Whitten, J. (2005). *Systems Analysis and Design Methods,* 7th ed. New York: McGraw-Hill/Irwin.

ONLINE RESOURCES

Public Health Information Network. http://www.cdc.gov/phin/.

Health Alert Network. http://www.phppo.cdc.gov/HAN/Index.asp.

COMCARE Data Standards. http://www.comcare.org/Data_Standards.html.

World Wide Web Consortium, Extensible Markup Language (XML). http://www.w3.org/ XML/.

National Incident Management System. http://www.fema.gov/pdf/emergency/nims.

OASIS Emergency Management Technical Committee. http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emergency.

Chapter 18

DESIGN AND PERFORMANCE OF A PUBLIC HEALTH PREPAREDNESS INFORMATICS FRAMEWORK

Evidence from an Exercise Simulating an Influenza Outbreak

IVAN J. GOTHAM[1,2,*], DEBRA L. SOTTOLANO[1], LINH H. LE[1], MICHAEL J. PRIMEAU[1], LORETTA A. SANTILLI[1], GERALDINE S. JOHNSON[1], STEPHANIE E. OSTROWSKI[1], AND MARY E. HENNESSEY[1]

CHAPTER OVERVIEW

Public Health Emergency Preparedness (PHEP) seeks to achieve and maintain a state of "readiness" within the community of response partners to detect, respond to, and mitigate health emergencies, such as large-scale infectious disease outbreaks. The activities, workflows, and information exchanges in this process are optimized when embedded as a part of routine public health practice. Information systems supporting PHEP "readiness" are also optimized when embedded within an informatics framework supporting a community of information trading partners engaged in routine (day-to-day) health information exchange. This chapter describes the attributes of a model informatics framework for support of PHEP; evaluates its performance during a full-scale exercise simulating an outbreak of a highly infectious novel strain of influenza; and discusses how the attributes of the framework contributed to the state of readiness in the response community.

[1*] *New York State Department of Health, Empire State Plaza, Albany, NY 12237, USA, ijg01@health.state.ny.us.*

[2] *School of Public Health, State University at Albany, Albany, NY 12222, USA*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_18,

Keywords: Health emergency preparedness; Informatics; Integrated information systems; Performance evaluation; Full-scale exercise; Disease outbreak

1. INTRODUCTION

Public health emergency preparedness (PHEP) has been defined as the capability of the public health and healthcare systems, communities, and individuals to prevent, protect against, quickly respond to, and recover from health emergencies, particularly those whose scale, timing, or unpredictability threatens to overwhelm routine capabilities (Nelson et al., 2007a). It is a state of sustainable "readiness to act," for all sectors and stakeholders involved in preparedness efforts (DHS, 2008), that is achieved over time as part of the essential public health activities health departments practice daily (Seid et al., 2007).

The practice of informatics supports and advances the state of PHEP. Public health informatics is defined as "the systematic application of information and computer science and technology to public health practice, research, and learning" (O'Carroll, 2003). As information is essential to support every phase of emergency planning and response, informatics can advance the state of "readiness" by assuring effective PHEP functions, workflow, and information exchange among the various partners that must participate in detection, response, and recovery during public health emergencies (Weiner and Trangenstein, 2007).

Understanding the PHEP processes, workflow, and decision structure among cross-jurisdictional and cross-discipline partners during a health emergency is fundamental for defining PHEP functions: information capture, analysis, visualization for situational awareness, decision making, information dissemination, and response (Nelson et al., 2007b). Similarities among the various workflow models in Table 18-1 and the key functions of the NYSDOH Informatics Infrastructure (Figure 18-1) illustrate commonalities in the PHEP workflows of major state, federal, and international public health agencies.

What type of informatics framework is best suited to enabling PHEP? PHEP activities and supportive functions are most effective when integrated with everyday public health activities (Baker et al., 2005; Koh et al., 2008; Nelson et al., 2007b, c; DHS, 2004). Thus an effective informatics framework for supporting PHEP should be based on information infrastructure and systems that simultaneously support both routine (normal) and emergency public health activities. The value of this "dual use" framework is that it has the potential to evolve naturally towards a sustained "readiness to act" among the key partners needed in a PHEP response.

Table 18-1. Major models of public health emergency preparedness functions and workflows.

Homeland security target capabilities (DHS, 2007a)	CDC goals & related functions (CDC, 2005)	WHO guidelines for integrated disease surveillance (Perry et al., 2007)	Bravata key decisions and tasks (Bravata et al., 2004)	NYSDOH Commerce informatics infrastructure key functions (Gotham et al., 2001, 2007)
Common capabilities: communications; intelligence & information sharing & dissemination	Timely, accurate communications; Functions: partner communications, alerting (PCA); analysis & visualization (AVR)	*Analyze & interpret:* disease patterns, data trends; create case demographics visuals. Calculate rates, action thresh-holds; describe risk factors, needed public health action	*Communication:* with first responders, public health, clinicians, officials & the public	*Alerts, communications:* role based contact via multimode alerting; *Data visualization & situational awareness*: real-time view of integrated data supports executive decisions
Prevent: gather information; recognize indicators; produce intelligence & analysis; detect CBRNE agents	*Prevent:* interventions to prevent human illness from CBRNE agents Function: Early event detection, (EED)	*Identify:* Case detection; data capture of suspect priority diseases present in inpatients, outpatients; conduct lab testing to diagnose suspected cases; maintain routine surveillance	*Diagnosis:* based on clinical symptoms; *Management:* care of exposed, acutely ill; *Prevention:* isolation & prophylaxis	*Local health disease reporting:* statewide disease case reporting; *Health facility surveillance:* report patient demographics, infectious disease admits; ED patient visits, admits, bed utilization

(*Continued*)

Table 18.1 (*Continued*)

Homeland security target capabilities (DHS, 2007a)	CDC goals & related functions (CDC, 2005)	WHO guidelines for integrated disease surveillance (Perry et al., 2007)	Bravata key decisions and tasks (Bravata et al., 2004)	NYSDOH Commerce informatics infrastructure key functions (Gotham et al., 2001, 2007)
Protect: epidemiological surveillance, investigation; lab testing	*Detect, report:* classify events; identify CBRNE agents; *Investigate:* risk factors, causes; interventions; Functions: outbreak management (OMS) & connecting lab systems (CLS)	*Report:* use of standard case definitions, when, how to report priority diseases & conditions; immediately report notifiable diseases across jurisdictions. *Investigate:* use investigation & lab results to confirm the outbreak	*Surveillance:* collect, manage, interpret data; *Reporting:* confirmed, suspect cases across jurisdictions. *Outbreak investigation:* verify cases are an outbreak	*Local health disease reporting, epidemiological investigation & response; Lab reporting:* contact tracing; transmit electronic lab test results; dynamic surveillance, surveys; health facilities report patient demographics
Respond: isolation, quarantine; mass prophylaxis; surge; critical resource logistics & distribution	*Control:* provide guidance countermeasures; Function: countermeasure & response administration (CRA)	*Respond:* treat cases & contacts using standard case management guidelines, infection control measures; involve community in response	*Outbreak control:* determine, perform out-break control measures, institute quarantine	*Health facility resource reporting, response*: Track assets, resources, surge capacity, countermeasures, bed availability; event patient admissions & tracking

Assessing the effectiveness or performance of PHEP systems presents unique challenges, as health emergencies are rare. Several efforts have been undertaken to measure the effectiveness of PHEP, including written assessments (Costich and Scutchfield, 2004) and exercises such as SNS drills (ASTHO, 2004) and the TOPOFF program (DHS, 2003a, b). While precise measurement of PHEP remains a challenge (Asch et al., 2005; Nelson et al., 2007 a, b, c), it is generally recognized that full-scale exercises are the best test of a system's ability to prepare for and respond to an emergency (Nelson, 2007). The Department of Homeland Security (DHS) National Preparedness Guidelines contains a Target Capabilities List (TCL) that identifies and defines preparedness capabilities and corresponding metrics against which achievement of a task or capability outcome can be assessed (DHS, 2007b). National Planning Scenarios (DHS, 2006) and Exercise Evaluation Programs (DHS, 2007c) provide standardized use cases for evaluation of performance metrics. One such scenario is pandemic influenza.

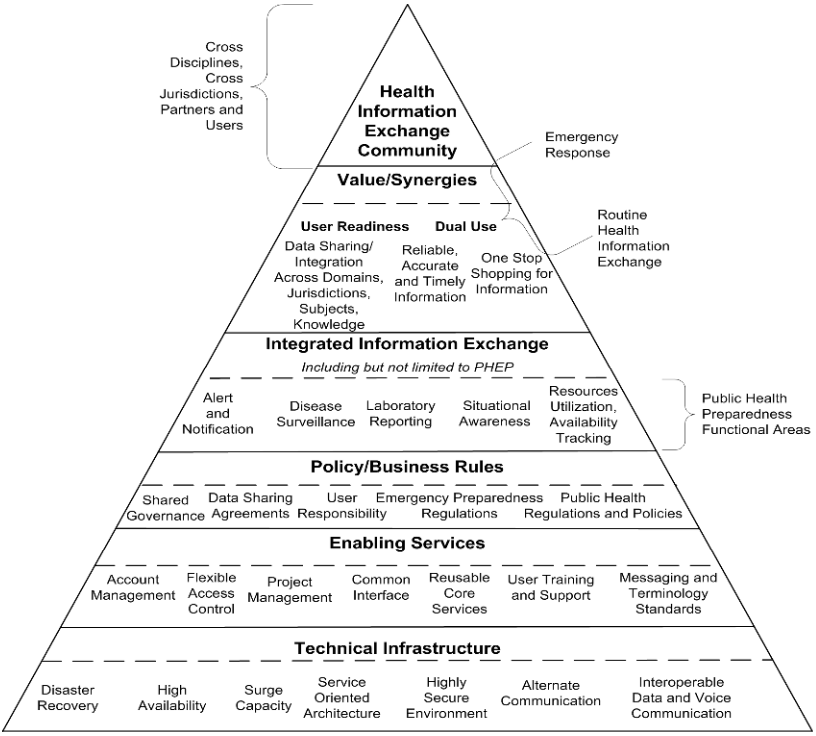

Cross Disciplines, Cross Jurisdictions, Partners and Users
Health Information Exchange Community
Emergency Response
Value/Synergies
Routine Health Information Exchange
User Readiness
Dual Use
Data Sharing/ Integration Across Domains, Jurisdictions, Subjects, Knowledge
Reliable, Accurate and Timely Information
One Stop Shopping for Information
Integrated Information Exchange
Including but not limited to PHEP
Alert and Notification
Disease Surveillance
Laboratory Reporting
Situational Awareness
Resources Utilization, Availability Tracking
Public Health Preparedness Functional Areas
Policy/Business Rules
Shared Governance
Data Sharing Agreements
User Responsibility
Emergency Preparedness Regulations
Public Health Regulations and Policies
Enabling Services
Account Management
Flexible Access Control
Project Management
Common Interface
Reusable Core Services
User Training and Support
Messaging and Terminology Standards
Technical Infrastructure
Disaster Recovery
High Availability
Surge Capacity
Service Oriented Architecture
Highly Secure Environment
Alternate Communication
Interoperable Data and Voice Communication

Figure 18-1. Framework for a model informatics infrastructure.

Over the past decade the New York State Department of Health (NYSDOH) has evolved an informatics infrastructure that enables electronic health information exchange across the state health enterprise (Gotham et al., 2001, 2002, 2007). The infrastructure encompasses the NY Health Commerce System (HCS), which is secure, integrated, interoperable, and web-based. It is in daily use for routine health information exchange applications by all NY local health departments (LHDs), healthcare organizations, and service providers (see Gotham et al., 2007). NYS PHEP functions (Table 18-1) are embedded within this dual-use infrastructure. In Spring of 2006, NY State conducted a full-scale Communicable Disease EXercise (CDEX) in the eight counties surrounding the Buffalo metropolitan area in western NY State. CDEX was data driven and designed to assess the state, regional, and local ability to identify, track, monitor, and mitigate the outbreak of a highly infectious novel strain of influenza. The generic PHEP functions used by the exercise participants to respond to the scenario included health alerting, epidemiological surveillance, healthcare response, dashboard visualization, and situational awareness (Table 18-1). The exercise provided the opportunity to measure the response rates and utilization metrics for these PHEP functions. This chapter applies these results to an assessment of the effectiveness of the informatics framework in supporting PHEP functions during a full-scale emergency.

2. MODEL INFORMATICS FRAMEWORK FOR HEALTH INFORMATION EXCHANGE

Among systems for information sharing and communications, the lack of interoperability across traditional organizational or jurisdictional boundaries has resulted in problems in responding to emergency situations (Bravata et al., 2004; Burkle, 2003) and has been noted as a deficiency in national exercises (DHS, 2007a). In response, governmental agencies, national organizations, researchers, and experts call for standards-based interoperable systems for seamless health information exchange (Baker et al., 2005; Brailer, 2004; Burkle & Hayden, 2001; CDC, 2008; Loonsk et al., 2006) and prescribe required system capabilities and performance metrics (DHS, 2007b). Key findings of the AMA/APHA Linkages Leadership Summit point to the need for PHEP components to operate within an informatics framework for health information exchange that is real-time, standards-based, secure, interoperable, integrated, reliable, highly available, and in common use by all stakeholders for both day-to-day and disaster operations, such as infectious disease outbreaks (Lyznicki et al., 2007; AMA/APHA, 2007).

Figure 18-1 depicts the framework for a model informatics infrastructure and the characteristics of each layer that provide reliability, user support, integrated information and decision support, as well as familiarity, knowledge, security, and trust in the system. The value of this framework is user and system "readiness to act, and interact" (DHS, 2007d):

- The *Technical Infrastructure* layer is the foundation for the system's reliability, availability, interoperability, efficiency, and security (see CDC, 2007; DHS, 2007a; Gotham et al., 2001, 2007; NYSDOH, 2006; Loonsk et al., 2006; OASIS, 2005a, b; HITSP, 2007).

- The *Enabling Services* layer interacts with system users and provides reusable services needed to establish defined and modifiable access, consistent vocabulary and well-defined concepts, and a user interface enabling immediate and familiar deployment of commonly needed functions, information relevance, and familiarity with system (Gartner, 2008; Popovich et al., 2002).
- The *Policy/Business Rules* layer defines protocols, requirements, responsibilities, and business rules that support a coordinated, efficient set of activities among diverse stakeholders (Gotham et al., 2001, 2007; Popovich et al., 2002).
- The *Integrated Information Exchange* layer is the first that describes a set of interoperable and essential functions – compliant with and supportive of DHS, CDC, and WHO frameworks – for accomplishing critical PHEP tasks and information exchange described earlier while simultaneously supporting the essential services of everyday public health practice (CDC, 2007; DHS, 2007a; Perry et al., 2007).
- The *Value/Synergies* and the *Health Information Exchange Community* layers represent a linked community of users committed to use of the framework as a whole, based on trust of the system's security and the reliability, accuracy, and relevance of the information created and available in that environment. The value and synergies that emerge from the opportunities for working across subject domains and from the linkage and sharing of information expand with the diversity and number of participant organizations and information exchange activities supported within the community (Baker et al., 2005; Burkle, 2001; Koh et al., 2008; Nelson et al., 2007b, c; Perry et al., 2007; Popovich et al., 2002).

3. NEW YORK STATE'S INFORMATICS FRAMEWORK FOR HEALTH INFORMATION EXCHANGE AND INTEGRAL SUPPORT OF PUBLIC HEALTH PREPAREDNESS

What are the key attributes of NY's informatics framework? The NYSDOH Health Commerce System (HCS) is a multi-tiered architecture (Figure 18-2) and is functionally compliant with national interoperability standards (e.g., Loonsk, 2007, OASIS, 2005a, b, HITSP, 2007). HCS operates within an informatics framework as described in Figure 18-1. The framework supports a diverse set of applications cross-cutting the scope of routine health information exchange, including: disease case and lab reporting, vital records, healthcare finance, healthcare utilization, managed care, medical conduct, controlled substance prescription reporting, licensed practitioner prescription pad orders, heavy metal and lead poisoning registries, malformations and cancer registry reporting, environmental and clinical laboratory proficiency reporting. In total, HCS currently supports some 200 health information applications, 75,000 users, and 15,000 participant organizations (see Gotham et al., 2007). Given this mission, the HCS architecture is highly available and includes full off-site disaster recovery capacity (Figure 18-2).

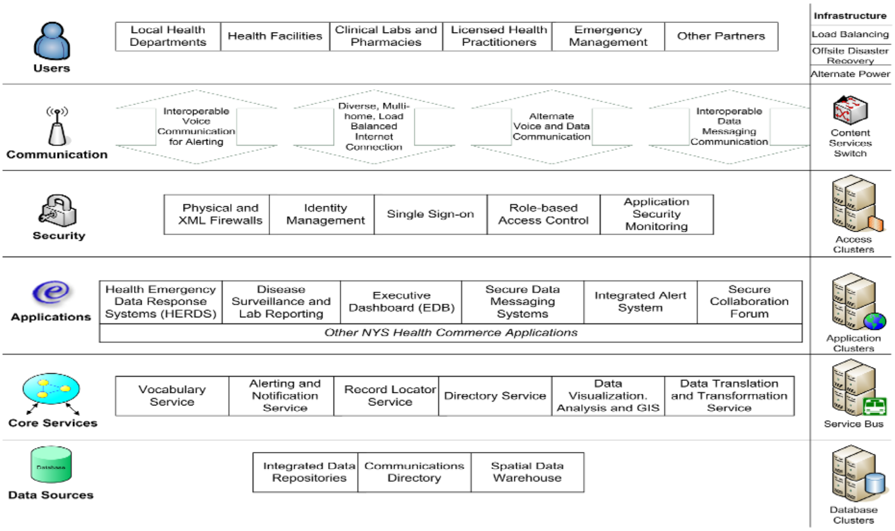

Users
Local Health Departments
Health Facilities
Clinical Labs and Pharmacies
Licensed Health Practitioners
Emergency Management
Other Partners
Infrastructure
Load Balancing
Offsite Disaster Recovery
Alternate Power
Communication
Interoperable Voice Communication for Alerting
Diverse, Multi-home, Load Balanced Internet Connection
Alternate Voice and Data Communication
Interoperable Data Messaging Communication
Content Services Switch
Security
Physical and XML Firewalls
Identity Management
Single Sign-on
Role-based Access Control
Application Security Monitoring
Access Clusters
Applications
Health Emergency Data Response Systems (HERDS)
Disease Surveillance and Lab Reporting
Executive Dashboard (EDB)
Secure Data Messaging Systems
Integrated Alert System
Secure Collaboration Forum
Other NYS Health Commerce Applications
Application Clusters
Core Services
Vocabulary Service
Alerting and Notification Service
Record Locator Service
Directory Service
Data Visualization, Analysis and GIS
Data Translation and Transformation Service
Service Bus
Data Sources
Database
Integrated Data Repositories
Communications Directory
Spatial Data Warehouse
Database Clusters

Figure 18-2. New York State Health Commerce System architecture.

Health information exchange applications share core integration services for generic functions such as notification/alerting, communications directory (ComDir) services, secure collaboration, GIS and data visualization (see Figure 18-2; also Gotham et al., 2007). ComDir is a centralized repository of role and contact information. Directory coordinators at each HCS organization assign users to functional roles within their view of ComDir and maintain contact information for both routine and emergency communications. The health alerting system uses contact information from the ComDir to notify targeted organizations and persons in specified functional roles within those organizations using automated phone calls, e-mail, fax, and secure web postings (Figure 18-2).

Access control is single sign-on and role-based. Role assignments made in ComDir by coordinators at participant HCS organizations enable access to application roles, data, and information appropriate to that organization. Access is conveyed instantaneously once a coordinator assigns a user to a ComDir role. Distributed access control allows the organization to assure appropriate access to applications and information within the organizational hierarchy. Access rules within the HCS information structure provide both open and restricted access to information, data, and applications (Gotham et al., 2003). Applications and information provision systems with restricted access use a hierarchical model to enable both vertical and horizontal flows of information appropriate to organization type and roles within that organization (see Gotham et al., 2002, 2007).

Organizations joining HCS are required to execute organization and user agreements covering security, data sharing, and nondisclosure. NY State Public Health Law requires all regulated healthcare facilities in NY State to abide by the HCS security agreements; maintain adequate levels of user accounts and coordinator designees to meet the state's disaster preparedness requirements; and maintain accurate and up-to-date role and contact information in ComDir (NYS, 2005a, b).

Oversight of the HCS system is a shared governance process with healthcare and local health department partners. Ongoing hands-on training sessions and drills reinforce skills in the use of the system. These partner groups also assist with coordination of drills, training, and requirements gathering. Online tutorials and reporting guidelines are available on the HCS training website. A dedicated accounts and help desk unit supports HCS account sign-up and user assistance.

How does NY's informatics framework enable and support PHEP? The HCS is an ideal platform for health preparedness, given its overarching informatics framework and routine use by the universe of partner organizations needed to facilitate detection and respond to a health event or emergency. As such, NY State has evolved a core set of interoperable PHEP workflows within HCS (Table 18-1) and also instantiated the capacity for automated exchange of standardized surveillance and response data with external entities, such as national infectious disease informatics portals (Zeng et al., 2004, 2005) and federal healthcare resource portals (AHRQ, 2005). The HCS preparedness systems have supported statewide response to emergent infectious disease events, emergency disaster declarations, health resource shortages, elevated national threat levels, and high-profile security events (Gotham et al., 2001, 2007). The HCS infrastructure is an integral component of NYSDOH incident management and PHEP plans (Gotham et al., 2007), including pandemic influenza (NYSDOH, 2006).

What are the core systems supporting PHEP activities and functions within the HCS informatics framework? Health alerting and other notifications occur through the HCS Integrated Health Alerting and Notification System (IHANS) (Gotham et al., 2001, 2007) using contact information derived from ComDir to notify appropriate roles and organizations that secure alert notifications have been posted on the HCS health alert network (HAN) file viewer. Notifications are sent using multiple pathways, including automated pathways, including automated phone calls, e-mail, and fax. On receiving a notification, users in the notified roles log into the HCS system, access the HAN file viewer, and download the alert document.

Electronic disease case reporting in NY State occurs through the HCS Communicable Disease Electronic Surveillance System (CDESS) (Gotham et al., 2003), where LHDs report detailed patient demographics and disease-specific supplemental data for 65 reportable disease conditions. CDESS also supports contact tracing activities and is integrated with a standards-based Electronic Clinical Laboratory Reporting System (ECLRS) (Gotham et al., 2003). Electronic reporting of test results to ECLRS by clinical labs is mandated by public health law in NY State (NYS, 2007). Positive test results for reportable disease conditions are transmitted by hospital, commercial, and public health laboratories to ECLRS and automatically routed to the LHD of jurisdiction for case establishment or confirmation. IHANS automatically alerts the county of jurisdiction on receipt of high-priority disease results. The NYSDOH Wadsworth Public Health Laboratory is a reference laboratory and part of the CDC Laboratory Response Network (LRN). Its Clinical Laboratory Information Management System (CLIMS) records and tracks test results on specimens submitted to the laboratory and also reports electronically to ECLRS.

Healthcare response in NY is supported by the HCS Health Emergency Response Data System (HERDS) (Tanielian et al., 2005; Gotham et al., 2007). HERDS is a statewide dynamic data reporting and visualization system supporting surveillance reporting (e.g., event-related patient admissions, deaths, and ED traffic) and resource and asset tracking (surge capability, bed availability, patient tracking, and medical encounter measures). Data reported into HERDS by healthcare facilities is immediately available to LHDs, as well as regional and state health jurisdictions, for planning and response. HERDS data is also available to the state and local incident command staff for planning, allocation, and distribution of state and federal stockpiled inventories of resources in an emergency. HERDS has been used in responding to emergency disaster declarations and healthcare resource shortages, in exercises for tracking bed and resource capacities, and in ongoing reporting activities, such as influenza and bed availability surveillance (Gotham et al., 2007). It is currently deployed to all hospitals, nursing homes, adult and home care facilities, clinics, and public schools statewide.

In an actual health event, NYSDOH activates a dedicated, open website within the HCS portal to provide general situational awareness to the HCS community at large, as was the case during the emergence of West Nile Virus in North America in 1999 (Gotham et al., 2001). This website provides timelines, updates, response protocols, plans, procedures, and links to data and applications needed for response. Executive-level decision support and situational awareness is achieved through an Executive DashBoard (EDB), providing summary-level visualization and integration of data feeds from response systems such as HERDS, CDESS, and ECLRS via drill-down charts, graphs, and maps. Access to the EDB is limited to key executive roles in ComDir for participating LHDs, hospitals, state and regional health offices, and other response partners. IHANS notifies HCS organizations on activation of the event-specific website and EDB.

4. EVALUATION OF FRAMEWORK RESPONSE DURING A FULL-SCALE EXERCISE

4.1 Exercise Scenario, Scope, and Extent

CDEX lasted from May 15th through June 15th, 2006. The participating organizations included 8 LHDs, 26 hospitals, and the Central and Western Regional State Health Offices. Counties participating in the exercise ranged from 43,000 to 930,000 in population. Participating hospitals ranged from 60,000/20,000 to 10/64 Emergency Department/Inpatient admissions per year. The scenario initiated with 900 passengers exposed to index cases aboard three local charter flights returning to the Buffalo region. Within 48 h of returning home an index case is admitted to a local hospital and dies, followed in close suit by family members. Specimen samples are sent by hospitals to the NYSDOH Wadsworth Laboratory where they test positive for a novel strain of influenza, "H7N2." Hospitals experience a concomitant surge of thousands of admissions and hundreds of deaths. Hospitals continue to submit specimens to Wadsworth Laboratory for confirmation. LHDs follow up with contact tracing of passengers from their jurisdictions on the airline manifests and report large numbers of disease cases and deaths of nonhospitalized cases. Hospital resources, such as Intensive Care Units, ventilators, and antiviral inventories, are overwhelmed. Hospital, county, and state emergency plans are activated. The Federal Strategic National Stockpile (SNS) is called upon to provide additional resources.

4.2 Preparedness Functions Used in the Exercise

The data flow within and between PHEP systems and organizations participating in the exercise is shown in Figure 18-3. IHANS supported all alerting functions. General situational awareness of the CDEX event as it

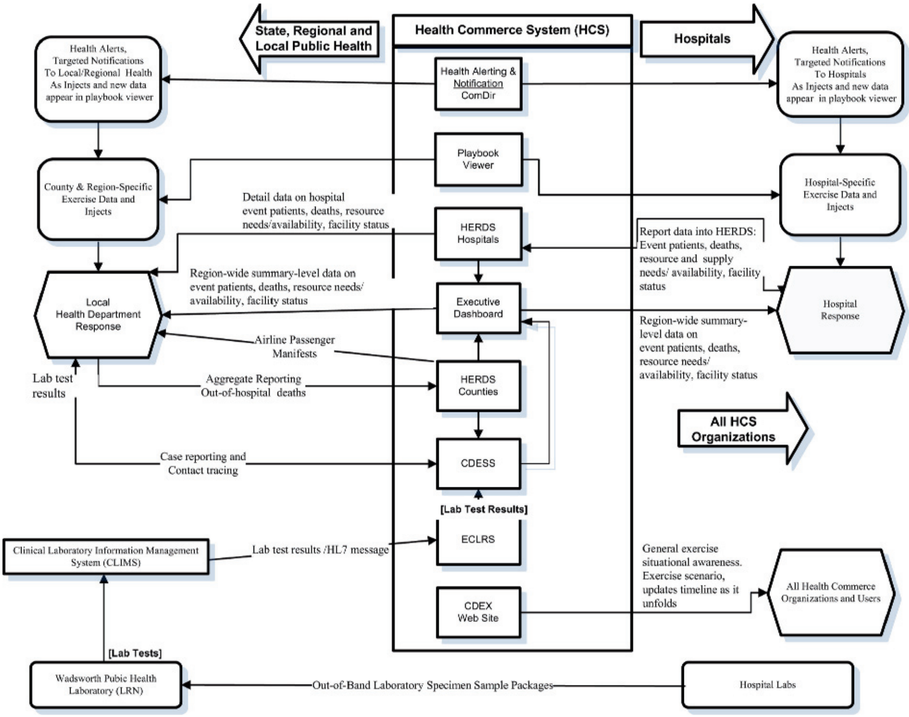

State, Regional and Local Public Health
Health Commerce System (HCS)
Hospitals
Health Alerts, Targeted Notifications To Local/Regional Health As Injects and new data appear in playbook viewer
Health Alerting & Notification ComDir
Health Alerts, Targeted Notifications To Hospitals As Injects and new data appear in playbook viewer
Playbook Viewer
County & Region-Specific Exercise Data and Injects
Hospital-Specific Exercise Data and Injects
Detail data on hospital event patients, deaths, resource needs/availability, facility status
HERDS Hospitals
Report data into HERDS: Event patients, deaths, resource and supply needs/ availability, facility status
Region-wide summary-level data on event patients, deaths, resource needs/ availability, facility status
Local Health Department Response
Executive Dashboard
Hospital Response
Region-wide summary-level data on event patients, deaths, resource needs/ availability, facility status
Airline Passenger Manifests
HERDS Counties
Lab test results
Aggregate Reporting Out-of-hospital deaths
All HCS Organizations
Case reporting and Contact tracing
CDESS
[Lab Test Results]
Clinical Laboratory Information Management System (CLIMS)
Lab test results /HL7 message
ECLRS
General exercise situational awareness. Exercise scenario, updates timeline as it unfolds
All Health Commerce Organizations and Users
CDEX Web Site
[Lab Tests]
Wadsworth Pubic Health Laboratory (LRN)
Out-of-Band Laboratory Specimen Sample Packages
Hospital Labs

Figure 18-3. CDEX inject and public health emergency preparedness data flow within HCS informatics framework.

evolved was made available to the exercise participants as well as the entire HCS community through the event-specific CDEX website. Electronic disease case reporting and contact tracing of index patients occurred through CDESS. The Wadsworth public health CLIMS system reported "H7N2" test results to ECLRS from mock specimens shipped to the Wadsworth Lab from participating hospitals. The hospital instance of HERDS (HERDS-H) was used by hospitals to report patient admissions and deaths by age group, resource needs, bed availability, emergency plan status, supply and medication inventories, ICU and ventilator needs. A county-based HERDS instance (HERDS-C) was used to push airline passenger manifest information to the LHDs of jurisdiction for follow-up contact tracing in CDESS. HERDS-C was also used for aggregate reporting of outside-of-hospital deaths by LHDs. Executive-level decision support and situational awareness were provided through the Executive DashBoard (EDB) using summary displays of data feeds from contributing PHEP systems (Figure 18-3). Access to the EDB was limited to key executive roles in participating organizations, including public health directors, hospital CEOs, lead epidemiologists, and directors of preparedness, disease control, and emergency departments.

4.3 Exercise Injects and Data Pushed to Exercise Participants

The information flow within and between HCS preparedness functions for sharing injects, alerts, and data among organizations participating in the exercise is shown in Figure 18-3. Three types of injects were used to drive the exercise scenario and concomitant response: *informational, action-request, and exercise "playbook" data.* Injects customized to each exercise participant organization (EPO) were placed in their respective secure file viewer, and IHANS was used to notify the target organization of new injects. Injects and notifications were sent unannounced and thus confirmation of access by the recipients is equivalent to a response to unannounced alerts or emergence of information about the event. Details regarding injects, organizations affected, actions, and response times are summarized in Table 18-2. Informational injects were sent to participating organizations for initiation or

Table 18-2. CDEX inject descriptions by organization, dates, response times allotted and actions requested.

Organization type	Inject ID	Date/time inject alert sent	Organizations alerted	Expected response[a]	Action step 1 requested	Action step 2 requested
LHD	LHD01.I	5/18/2006 13:26	1	asap[b]	Info. inject sent to 1 LHD	
LHD	LHD02.I	5/19/2006 15:08	1	asap	Info. inject sent to 1 LHD	
LHD	LHD03.C	5/22/2006 14:38	8	asap	Pick up airline manifest 1 from HERDS-C	Use data – contact tracking in CDESS
LHD	LHD04.H	5/23/2006 16:49	8	asap	Review hospital data in HERDS-H	
LHD	LHD05.C	5/24/2006 14:16	8	asap	Pick up airline manifest 2 from HERDS-C	Use data – contact tracking in CDESS
LHD	LHD06.L	5/26/2006 10:02	8	asap	Pick up lab test results on cases in ECLRS	
LHD	LHD07.I	5/26/2006 10:10	8	asap	Review school closing policies	
LHD	LHD08.P	6/02/2006 12:41[c]	8	76.8 h[d]	Pick up playbook data[e]	Use data – report deaths in HERDS-C
LHD	LHD09.P	6/09/2006 09:26[c]	8	80.1 h[d]	Pick up playbook data[e]	Use data – report deaths in HERDS-C
Hospital	HOSP01.I	5/18/2006 11:57	1	asap	Info. inject sent to 1 hospital	

(*Continued*)

Table 18.2 (*Continued*)

Organization type	Inject ID	Date/time inject alert sent	Organizations alerted	Expected response[a]	Action step 1 requested	Action step 2 requested
Hospital	HOSP02.I	5/19/2006 14:38	2	asap	Info. inject sent to 2 hospitals	
Hospital	HOSP03.I	5/22/2006 10:40	1	asap	Ship specimens to public health lab	
Hospital	HOSP04.I	5/23/2006 12:56	26	asap	Ship specimens to public health lab	
Hospital	HOSP05.H	6/13/2006 10:47	27	asap	Update resource requests in HERDS	
Hospital	HOSP06.P	5/19/2006 15:53[c]	27	73.6 h[d]	Pick up playbook data[f]	Operationalize data in HERDS-H
Hospital	HOSP07.P	5/26/2006 13:25[g]	27	100 h[d]	Pick up playbook data[f]	Operationalize data in HERDS-H
Hospital	HOSP08.P	6/02/2006 16:51[c]	27	72.7 h[d]	Pick up playbook data[f]	Operationalize data in HERDS-H
Hospital	HOSP09.P	6/09/2006 09:54[c]	27	79.6 h[d]	Pick up playbook data[f]	Operationalize data in HERDS-H

LHD local health department, *HCS* health commerce system, *HERDS-H* health emergency response data system, hospital instance, *HERDS-C* health emergency response data system, LHD instance, *CDESS* communicable disease electronic surveillance system, *ECLRS* electronic clinical lab reporting system (integrated with CDESS)

[a] Time allotted to pick up inject or playbook and complete action(s) requested

[b] asap – immediate action requested

[c] Friday

[d] Time given in actual time, cob next business day. Exercise time amounted to 24 h

[e] Synthetic data on out of hospital deaths

[f] Synthetic data driving hospital outbreak related admits, deaths, resource utilization

[g] Friday of Memorial Day weekend

continuance of the exercise scenario. Action-requests provided, or referenced availability of, background information and data and requested that the recipient organization take one or more steps to complete an activity. The hospital exercise playbooks contained detailed hospital-specific data for that week on event-related Emergency Department (ED) traffic, hospital admissions, and deaths by age group and event-related resource utilization, such as ICU and ventilator usage. They also provided hospital-specific baseline resource, bed, and patient care data that would normally be expected for that time of year. LHDs were also sent playbook data to enable their reporting of aggregate out-of-hospital deaths through the HERDS-C.

The hospital playbook data was used to drive hospitals' operational responses and reporting of facility status, asset needs (equipment, supplies, medications), and bed availability into HERDS-H. Playbooks also drove reporting of event-related admissions and deaths by age group and ED patient traffic into HERDS. The hospital data reported into HERDS-H was therefore available to LHDs, state and regional health offices, and incident command staff, driving their response (Figure 18-3).

Mock Health Alerts were also disseminated as the scenario evolved. These were posted on the HCS health alert viewer and made available to all HCS users. The IHANS system sent phone notifications of the postings to key roles at Exercise Participant Organizations (EPO). LHD staff, who also have access to the IHANS system for alerting within their jurisdictions, were encouraged to use it to cascade exercise alerts within their participating jurisdiction. Seven of the eight participating LHDs cascaded these alerts using IHANS.

The CDEX website was activated at the beginning of the exercise and EPOs were alerted. Other Non-Participating Organizations Alerted (NPO-A) at the same time included the remaining hospitals and LHDs in the state and also all nursing homes statewide. The process of accessing the CDEX website by other Non-Participating HCS Organizations who were Not Alerted (NPO-NA) was therefore strictly passive, their users logging on to HCS during the exercise as a matter of their routine use of the system. NPO-As were apprised of the exercise in advance and encouraged to access the CDEX website for situational awareness via informational notifications through the IHANS system. The NPO-NAs (e.g., pharmacies, individual providers) were not informed of activation of the CDEX website out of concern over misinterpretation of the exercise as an actual event.

4.4 Methodology Used in Measuring Preparedness Function Responses

Organizational responses to *informational* injects were measured as time in minutes from transmission of alert notification to the time at which the first user in that organization downloaded the inject or health alert content. Organizational responses to *action-request* inject notifications were measured as time in minutes from transmission of alert notification to the time at which the first user in that organization downloaded the inject content and, where requested by the inject, subsequently initiated an action using a PHEP function as listed in Table 18-2. Response expectations for *playbook data* injects were that the requested responses (Table 18-2) occur by close (17:00 h) of the next business day. The absolute amount of time to complete the response varied depending on when the playbook notification was transmitted. Playbook responses were measured as the number of organizations accessing playbook data and responding with data reported into HERDS within 5, 50, and 100% of the time between alert transmission and close of the next business day. User and organization utilization of situational awareness functions was tracked for the CDEX website and the EDB. IHANS used all alerting modalities for announcing injects.

4.5 Responses to Informational and PHEP Action-Request Injects

Responses to inject alerts reflect the organizations' ability to receive an unannounced notification from IHANS, access HCS to receive new information or requested actions, and immediately initiate those actions (Tables 18-1 and 18-2; Figure 18-3). For local health departments these functions involved accessing data on potentially exposed patients in their jurisdiction and initiating contact tracing; picking up positive laboratory test results on suspect cases and initiating case reports on infected patients; accessing hospital reports on event patients and deaths and resource requirements in their jurisdiction. For hospitals these functions involved reviewing event patient admissions and deaths and resource and medical countermeasure needs and immediately reporting to LHDs and NY State via HERDS-H (Figure 18-3).

Response times measured for informational and action-request injects for LHDs and hospitals are shown in Table 18-3. The overall average response time for LHDs accessing inject content was 29 min (90% CL [23,35]) from alert initiation. Responses to action step injects (accessing passenger manifests and initiating contact tracing in the CDESS reporting system) averaged 103 min (90% CL [80,127]) from alert initiation. Responses to action step injects (accessing laboratory test results on specimens from suspect cases submitted by hospitals and reported by Wadsworth Laboratory into ECLRS/CDESS) averaged 75 min (90% CL [43,107]) from time of alert initiation.

For hospital participants, the overall average response time in accessing inject content was 49 min (90% CL [43,56]) from alert initiation. One of the injects, an urgent request, required hospitals to review and update their admissions and resource data reported into the HERD-H system. The average time for hospitals to complete this request was 175 min (90% CL [161,189]) from alert initiation.

Table 18-3. LHD and hospital time responses to inject notifications and subsequent requested actions related to surveillance, data access and response.

Inject ID	Date – time inject alert sent	Number organizations alerted	Expected response[a]	Time to access inject on HCS (min) – average[b] [90%CL]	Time to initiate action step 1 (min) – average[b] [90%CL]	Time to initiate action step 2 (min) – average[b] [90%CL]
LHD01.I	5/18/2006 13:26	1	asap[c]	4	–	–
LHD02.I	5/19/2006 15:08	1	asap	3	–	–
LHD03.C	5/22/2006 14:38	8	asap	45[28,62] $P < 0.002$	65 [45,85] $P < 0.0004$	106[61,151] $P < 0.003$
LHD04.H	5/23/2006 16:49	8	asap	30 [8,53] $P < 0.037$	495[d] [167,823] $P < 0.024$	–
LHD05.C	5/24/2006 14:16	8	asap	42 [33,52] $P < 0.0001$	63 [52,73] $P < 0.0001$	100 [73,127] $P < 0.0002$
LHD06.L	5/26/2006 10:02	8	asap	17 [12,21] $P < 0.0002$	75 [43,107] $P < 0.003$	–
LHD07.I	5/26/2006 10:10	8	asap	18 [13,23] $P < 0.008$	–	–
Overall LHD response				29[23,35] $n = 42$ $P < 0.0001$	68[56,79] $n = 24$ $P < 0.0001$	103[80,127] $n = 16$ $P < 0.0001$
HOSP01.I	5/18/2006 11:57	1	asap	56	–	–
HOSP02.I	5/19/2006 14:38	2	asap	41.5(± sd 3.5)	–	–
HOSP03.I	5/22/2006 10:40	1	asap	11	–	–
HOSP04.I	5/23/2006 12:56	26	asap	67 [58,76] $P < 0.0001$	–	–
HOSP05.H	6/13/2006 10:47	27	asap	34 [27,41] $P < 0.0001$	175 [161,189] $P < 0.001$	
Overall hospital response				49[43,56] $n = 57$ $P < 0.0001$	–	–

See Table 18-2 for inject descriptions

LHD local health department

[a] Time allotted from transmission of inject alert notification to accessing inject information and taking action as requested on HCS

[b] Average elapsed time for all organizations from transmission of inject alert to accessing inject information and taking actions as requested on HCS

[c] As soon as possible (immediate responses requested)

[d] Mean is significantly different from other LHD response action steps ($P < 0.003$). While all LHDs picked up the LHD04.H inject within 109 min, apparently due to alert being sent at 4:50 p.m., half (4) of the LHDs accessed HERDS-H in less than 60 min, while the remaining half (4) waited until the following business day to access HERDS-H. This response was excluded from the overall average

In sum, on average LHDs and hospitals were able to receive an ad hoc urgent notification, access HCS, and retrieve the content within 50 min from initiation of the notification. LHDs were able to initiate accessing disease surveillance and lab reports within 1.75 h of notification. On urgent request, hospitals were able review, update, and report resource and medical countermeasure requirements and event patients and deaths within 3 h of alert initiation.

4.6 Responses to Operational Data Injects

The response to hospital playbook injects reflects the facilities' ability to receive a request, gather information, and report patient admissions and resource impact within a requested deadline. Reporting resource needs and facility status via HERDS supports state and local incident command ability to distribute state and federal stockpiles as needed. Hospital-based reporting of event-related patient admissions and deaths by age group provides state and local epidemiologists with detailed information on the extent, severity, and age-specificity of the emergent virus. LHD reports of out-of-hospital deaths provide additional epidemiological information to neighboring LHDs and the NYSDOH.

Measures of responses to playbook injects by LHDs and hospitals are shown in Table 18-4. All LHDs accessed and downloaded the playbook injects within 5% of the time allotted, and on average 81% (±8) recorded the requested information into HERDS-C within the same interval. Seventy-five percent (±20) of hospitals accessed and downloaded the playbook injects within 5% of the time allotted, and on average 47% (±26) completed reporting of the requested data into HERDS-H in the same time interval. Eighty percent of LHDs completed the response function within 5% of the requested time, and the same percentage of hospitals completed the activity within 50% of the requested time. All organizations were able to complete these activities within the requested timeframes (Table 18-4).

Table 18-4. LHD and hospital time responses to inject notifications as to updates in exercise playbook data and actions taken to operationalize data and report in HERDS.

Inject ID	Date/time playbook notification sent	Number alerted	Expected response[a]	Cumulative number accessing playbook data within percent time given to respond: Number/(% total organizations)			Cumulative number accessing and reporting playbook data to HERDS system within percent time given to respond: Number (% total organizations)		
				5%	50%	100%	5%	50%	100%
LHD08.P	6/02/2006 12:41[b]	8	76.8 h	8/(100)	8/(100)	8/(100)	6/(75)	6/(75)	8/(100)
LHD09.P	6/09/2006 09:26[b]	8	80.1 h	8/(100)	8/(100)	8/(100)	7/(87)	7/(87)	8/(100)
Average percent LHDs responding ± SD				100	100	100	81 ± 8	81 ± 8	100
HOSP06.P	5/19/2006 15:53[b]	27	73.6 h	19/(70)	22/(81)	27/(100)	10(37)	10/(37)	27/(100)
HOSP07.P	5/26/2006 13:25[c]	27	100 h	21/(78)	22/(81)	27/(100)	15/(55)	15(55)	27/(100)
HOSP08.P	6/02/2006 16:51[b]	27	72.7 h	14/(52)	16/(59)	27/(100)	5/(18)	8/(30)	27/(100)
HOSP09.P	6/09/2006 09:54[b]	27	79.6 h	27/(100)	27(100)	27/(100)	21/(78)	21/(78)	27/(100)
Average percent hospitals responding ± SD				75 ± 20	80 ± 17	100	47 ± 26	50 ± 21	100

See Table 18-2 for inject descriptions

LHD local health department, *HERDS* health emergency response data system, *HCS* health commerce system

[a] Time allotted from transmission of inject notification to accessing playbook data on HCS, operationalizing data as appropriate and entering data into HERDS. Time given in actual time, cob next business day. Exercise time amounted to 24 h

[b] Friday

[c] Friday of Memorial Day weekend

4.7 Accessing Health Alert Postings by Key Roles at Local Health Departments and Hospitals

Notifications that an alert had been posted on HCS were transmitted by NYSDOH IHANS to specific roles at the EPOs using phone, fax, and e-mail. On receipt of the notification, staff assigned to those roles were expected to log on to HCS, access the secure alert system viewer, and download the alert content. Table 18-5 details alert content, timing, and roles notified as well as average response rates for LHDs and hospitals. Because of the diversity of functional roles in ComDir, the event-related communiqués both informed the epidemiological response and supported local and healthcare risk communications (Table 18-5). The overall average time to access all exercise health alert postings by key roles at LHDs (71 min) was significantly shorter than that observed for roles (194 min) at hospitals ($P < 0.0005$).

4.8 Usage of CDEX Event-Specific Website for Situational Awareness

During the exercise the CDEX website received 177,690 access hits by 15,795 unique users. Eighty-seven percent of all users accessing HCS during the exercise also accessed the CDEX website. Details of usage statistics for the CDEX website by CDEX participants and non-participants are shown in Tables 18-6 and 18-7, respectively. EPOs accounted for 48% of the total usage of the site. Nearly 1,900 users from the EPOs accessed the site, of which 63% (±11), on average, were repeat users. With the exception of one hospital, all EPOs accessed the site at least once per weekday over the duration of the exercise (Table 18-6).

Non-participant organizations accounted for 52% of the access hits, with nearly 14,000 users distributed among 7,500 distinct organizations accessing the site (Table 18-7). All NPO-As statewide (LHDs, hospitals, and nursing homes) accessed the site, accounting for 70% of hits from non-participants. Large numbers of NPO-NAs also accessed the site, including adult and home care facilities, pharmacies, clinical labs, and clinics (Table 18-7). On average, 85% (±11) of NPO-NAs that actually logged on to HCS also accessed the CDEX website. Of the non-participating organizations, those alerted also showed the highest percentages of repeat users (t-test alpha = 0.05, $t = 3.5$, $p = 0.008$, df = 8).

Table 18-5. Response rates to health alerts and updates posted during CDEX exercise by participating LHDs and hospitals.

Date/time sent	Content	Roles alerted	LHD response (min)	Hospital response (min)
			Average[a] [90%CL]	Average[a] [90%CL]
5/17/2006 12:13	Influenza A (H7N2) Advisory: WHO declares pandemic alert phase 5	LHDs – commissioner/director of public health, director disease control, lead epidemiologist, PHEP contact; Hospitals – ICP, BT coordinator	17[0,35] $P < 0.11$	114[54,176] $P < 0.004$
5/19/2006 13:24	Press release: H7N2 Influenza virus identified by Wadsworth Center Laboratory	LHDs – Commissioner/director of public health, preparedness risk communications coordinator Hospitals – PIO, BT coordinator	21[7,35] $P < 0.026$	184[71,296] $P < 0.010$
5/22/2006 16:12	Influenza A (H7N2) Alert: First case reported in New York. WHO declares pandemic alert phase 6	LHDs – commissioner/director of public health, director disease control, lead epidemiologist, PHEP contact Hospitals – ICP, BT coordinator	166[107,224] $P < 0.001$	274[224,324] $P < 0.0001$
6/12/2006 16:08	PIO message maps regarding outbreak	LHDs – commissioner/director of public health, preparedness risk communications coordinator Hospitals – BT coordinator, PIO	102[−13,217] $P < 0.008$	360[288,434] $P < 0.0001$
6/13/2006 12:57	Public health alert: prioritization of antiviral medications	LHDs – commissioner/director of public health, director disease control, lead epidemiologist, PHEP contact Hospitals – ICP, BT coordinator	11[4,17] $P < 0.015$	53[−4,110] $P < 0.125$
6/13/2006 13:28	PIO message maps regarding antiviral prioritization	LHDs – commissioner/director of public health, PHEP contact, risk communications coordinator Hospitals – PIO	108[15,200] $P < 0.065$	160[88,232] $P < 0.001$
Overall response rates			71[44,97] $n = 48$ $P < 0.0001$	194[161,226] $n = 162$ $P < 0.0001$

LHD local health department, *CD* communicable disease, public information officer, *BT* bioterrorism (disease event preparedness coordinator), *ICP* infection control practitioner

[a] Average time required from transmission of notification to first access of alert content by each participating county or hospital on HCS

Table 18-6. CDEX web site access by HCS organizations participating in the CDEX exercise.

Organization type	Number participating organizations	Distinct HCS organizations within each type accessing CDEX website during exercise[a]		Percent total access hits	Distinct users from participant organizations accessing CDEX website			Response rate on activation (h)[c]	Site access frequency[d] by participant organizations
		Number	% Total		Percent total distinct users	Percent repeat users[b] within organization	Average hits/user		
LHDs	8	8	100%	30%	14%	74%	98	16.6	100%
Hospitals	27	27	100%	30%	18%	62%	75	40.2	96%
State and regional health	Regional = 4[e] Central = 17[f]	Regional = 4 Central = 17	Regional = 100% Central = 100%	40%	68%	52%	27	42.8	100%
Average ± SD	–	–	100%	–	–	63 ± 11	67 ± 36	33 ± 14	98.7 ± 2.3
Overall totals	Organizations 56	Organizations 56	na	Hits 84,701	Users 1,865	na	na	na	Na

[a] Duration of CDEX exercise 15 May 2006–15 June 2006

[b] Three or more visits on separate days during the exercise

[c] Time required for at least 95% of organizations within their organization type to first access dashboard after activation. Participant organizations alerted as to dashboard availability on Monday 15 May 2006:16:35 h

[d] Percent organizations within organization type accessing the site at a rate of at lease once every weekday during exercise

[e] Number of regional health offices involved in the exercise

[f] Number of central office program areas involved in exercise

Table 18-7. CDEX web site access by HCS organizations not directly participating in the CDEX exercise.

Organization type	Alerted on activation of website	Distinct HCS organizations within each type accessing CDEX website during exercise[a]			Percent total access hits	Distinct users from non-participant organizations accessing CDEX website		
		Number	% Registered with HCS	% Accessing HCS		% Total users	Percent repeat users[b]	Hits per user
Local health department[c]	Y	50	100%	100%	36.20%	9.83%	66%	25
Hospital[d]	Y	405	100%	100%	23.40%	18.76%	56%	8
Nursing home[e]	Y	779	100%	100%	10.35%	11.20%	40%	6
MDs and practices	N	2,667	18%	87%	7.02%	23.59%	8.5%	2
Other practitioners and practices[f]	N	1,452	17%	86%	3.35%	10.74%	11%	2
Adult & home care	N	742	36%	97%	9.32%	10.11%	43%	6
Pharmacy	N	615	41%	63%	1.33%	4.65%	10%	2
Clinics	N	301	43%	93%	2.03%	2.96%	31%	5
Clinical or environmental lab	N	252	14%	81%	2.58%	2.87%	28%	2
Other organizations[g]	N	263	44%	86%	4.42%	5.28%	26%	6
Overall totals		Organizations 7,536	na		Hits 92,989	Users 13,930	na	na

[a] Duration of CDEX exercise 15 May 2006–15 June 2006

[b] Three or more visits on separate days during the exercise

[c] Includes NY City Department of Health and Mental Hygiene

[d] Includes 239 physical hospital facilities in addition to their operating certificate business entities

[e] Includes 663 physical nursing home facilities in addition to their operating certificate business entities

[f] Includes dentists, nurse practitioners, physician assistants, veterinarians

[g] Includes schools, other state agencies, law enforcement, emergency management, etc.

Absent alerting or direct participation in the exercise, large numbers of organizations encountered the event website and thus gained situational awareness of CDEX during the course of their routine activities on HCS. In an infectious disease event, the universe of participant organizations (see Table 18-7) would look to HCS for situational awareness, many also having the potential to serve as partners assisting in response or resource provision to facilitate the response (e.g., pharmacies).

4.9 Executive DashBoard Usage for Situational Awareness by Key Decision Makers

The PHEP systems (Figure 18-3) distributed detailed data on surveillance, case reporting, and healthcare resource status to subject matter experts involved in the exercise. The integrated data displayed via the EDB (Figure 18.3) provided executives with summary visuals of key information needed for incident management, including event-related patient admissions, ED traffic, bed availability (e.g., ICU beds needed and available), equipment needs (e.g., adult and pediatric ventilators), medical countermeasure needs (e.g., antiviral inventories), and general supplies (e.g., PPE). The HCS data sharing and access rules allowed access to key executive roles at response partner organizations, including hospitals and local, regional, and central state health offices.

Usage of the EDB by executive decision makers within EPOs is detailed in Table 18-8. EPOs were alerted via the IHANS system when the EDB was activated on 25 May 2006:15:00 h. Seventy-five percent of all EPOs accessed the EDB within 21 (±4) h and 95% within 61 (±39) h. On average 53% (±13) of all users in key roles within EPOs who accessed the EDB were repeat users. At the organization level, 98% (±4) of EPOs had at least one executive user who accessed the EBD at least every other weekday, with 77% (±21) accessing it at least once every weekday. In sum, executive decision makers at organizations participating in the exercise returned to the EDB on a regular basis.

Table 18-8. Executive dashboard (EDB) usage by HCS organizations participating in the CDEX exercise[a].

Organization type	Distinct HCS organizations within each type accessing EDB over duration of its availability[b]		Percent total access hits on EDB by organization[b]	Distinct users from participant organizations accessing EDB			Organization response rate on activation (h)[d]		Site access frequency by organization[e]	
	Number accessing EDB	Percent total organizations of this type		% Total distinct users accessing dashboard	% Users within organization type who were repeat users[c]	Average hits/user	95%	75%	At least once per weekday	At least every other weekday
LHD	8	100%	22%	26%	65%	42	19	17	100%	100%
Hospitals	27	100%	32%	30%	54%	55	67	22	60%	93%
State and regional health[f]	Regional = 4 Central = 17	Regional = 100% Central = 100%	46%	44%	40%	50	96	24	70%	100%
Total numbers	56	–	20,131	418	–	–	–	–	–	–

LHD local health departments

[a] Access restricted to key executive decision-maker roles (e.g. public health director, lead epidemiologist, hospital CEO) within organizations participating in exercise

[b] Duration of dashboard availability: 25 May 2006–15 June 2006

[c] Three or more visits on separate days

[d] Time required for at least 75% or 95% of organizations within their organization type to first access dashboard after being alerted of its activation on Thursday 25 May 2006:15:00 h

[e] Percent organizations within organization type accessing the site at a rate of at lease once every weekday or every other weekday

[f] Four regional health offices and 17 participating central office program areas

5. DISCUSSION AND LESSONS LEARNED

5.1 How Well Did the Observed Exercise Responses Meet Expectations?

From the perspective of the overall informatics framework for PHEP response, the functions and response times observed in this exercise agree well with the metrics published in the DHS Target Capabilities List (DHS, 2007b) and the CDC's target capabilities and critical tasks (CDC, 2008). The results of the CDEX exercise also agree well with functions supported, and responses observed, in actual health events and other drill scenarios (Gotham et al., 2001, 2007).

During the 30 days of CDEX's operation, the PHEP functions on HCS were in continual and repeated use by both the exercise participants as well as the general HCS health information exchange (HIE) community. The organizations participating in the exercise demonstrated the ability to receive alerts and update emergent information in a timely manner, immediately accessing and subsequently taking action via the electronic PHEP workflows supported within the HCS. Healthcare facilities were able to receive and assimilate complex exercise scenario data, operationalize it, and report their resource needs and patient admissions well within the expected response times. Key executive decision makers demonstrated repeated use of the integrated dashboard visualization tool for situational awareness throughout the exercise. The organizations within the HCS HIE community not directly involved in the exercise also demonstrated repeated use of the event-specific website for situational awareness. The alert system, supported by a broad-based communications directory, demonstrated rapid and simultaneous support of multiple PHEP activities across organizations responding to an emergent event. These activities include the clinical as well as the risk communications response across local public health and healthcare sectors. The responses observed in the CDEX exercise – along with those observed in actual health events and other drill scenarios (Gotham et al., 2001, 2007) – support the conclusion that the NYS HIE community is in a state of "readiness" to use the PHEP functions within the HCS informatics framework for the detection of and sustained response to a public health emergency.

5.2 How Does the HCS Informatics Framework Enable a State of PHEP "Readiness"?

Value and synergies gained through dual use (Figure 18-1). The CDEX scenario simulated an outbreak of a novel and highly infectious disease unfolding over a month. During this time participating organizations engaged in sustained, continuous use of the HCS PHEP functions in their response to the scenario. The response partners were also able to continue to engage in their day-to-day health information exchange activities on the HCS. At the same time, the global HIE community was informed of event status through their access to the open event website. The effectiveness of the "dual use" of HCS – facilitating both event-related situational awareness and routine information exchange – is illustrated by the fact that every local health department, hospital, and nursing home in the state accessed the site during the exercise. Experience with actual health events in the past has shown that the HCS HIE community sees the State's system as the authoritative source of information about health events (Gotham et al., 2001, 2007).

The diversity, activity, and breadth of the HCS HIE community enhances both readiness and capacity for response by enabling the incident command process to contact and increase the circle of engaged response partners as the situation requires. The intensity of CDEX website access by HCS organizations not directly involved in the exercise indicates a high degree of interest in the subject of pandemic disease events. By accessing the site and downloading its contents, these organizations were also provided the opportunity to improve their knowledge and increase their awareness of response plans and surveillance processes for such an event. *The value and synergies realized from dual-use informatics frameworks is significant. Embedding PHEP functions within an existing informatics framework (Figure 18-1) that also supports a broad set of routine information health exchange activities for a large, diverse HIE community is a powerful enabling factor in establishing and sustaining a state of "readiness" and the capacity for response to health events within that community.*

Integrated information exchange (Figure 18-1). The electronic systems integrated within the HCS informatics framework support workflows essential for PHEP response to health events (Table 18-1; Figures 18-1 to 18-3) as well as day-to-day public health activities. In either mode, the electronic systems provide a continual flow of information between and within the horizontal and vertical hierarchy of healthcare organizations and state, regional, and local public health agencies. As these electronic systems are integrated within the same framework, their data feeds are also easily integrated, summarized, and presented in a single visual interface for decision support and situational awareness across HCS organizations. *Embedded PHEP systems built on interoperability standards establish the substrate for integrated information exchange. However, well-defined policies and access rules, shared governance, data sharing and non-disclosure agreements – along with flexible, role-based access control distributed to participant organizations – are essential components for assuring that the "Exchange" in HIE actually happens in a timely manner. Having these policies, processes, and business rules in place for ongoing routine information exchange is absolutely critical to maintaining a state of "readiness" for rapid, unfettered information exchange required for effective event detection, response, and mitigation.*

Enabling services (Figure 18-1). The response rates to ad hoc informational and action-request inject alerts (Tables 18-2 and 18-3; Figure 18-3) observed in the exercise are indicative of a pre-existing state of "user-readiness" within the participating organizations. This state is predicated on essential preconditions: respondents' contact and role information in the communications directory must be accurate and up-to-date; respondents' accounts must be active and users must know their authentication information; respondents must be familiar with the response protocols for alerts; and each respondent must be intimately familiar with accessing and effectively using PHEP electronic systems. *The key enabling services (Figure 18-1) promoting and supporting user familiarity and usage – such as training, help desk and account maintenance, common usability interface and terminology, and flexible access control – must be in routine daily operation well in advance of a health event in order to assure that preconditions for user-readiness are met.*

Technical Infrastructure (Figures 18-1 and 18-2). The attributes of the underlying technical infrastructure supporting a dual-use infrastructure for routine health information exchange and PHEP must also reflect a state of "readiness". The response rates observed in this study would not have been achievable without the capacity to support dual use as well as surge usage. The infrastructure must therefore have sufficient communication and computation *capacity* to support both in an emergency. It must be a *trusted and reliable site* that provides a secure and highly available environment with information for disaster recovery and alternate communications. The application architecture must be *agile, flexible, and extensible* and be able to leverage reusable and interoperable core services to rapidly assemble interconnected PHEP functions to meet emergent circumstances. Finally, it must be able to support rapid dissemination and bi-directional exchange of data, information, and intelligence with data providers or external agencies using open interoperability standards for health information exchange.

6. CONCLUSIONS

This chapter provides evidence from a full-scale exercise on how a model informatics framework can enable and support an ongoing state of Public Health Emergency Preparedness (PHEP). A foundational principle of this framework is that information systems supporting PHEP "readiness" are optimized when embedded within a dual-use information infrastructure that supports a community of information trading partners engaged in routine (day-to-day) health information exchange, including the routine practice of public health. NYSDOH is currently pursuing two changes to the HCS system that will provide further research opportunities. The system's user interface and organization is being redesigned to operate within a portalized environment with assistance from experts in human usability design. The changes will enable the opportunity to compare the impact of human interface design improvement on workflow improvement and information access by the user community. The HCS system is also being revised to enable bi-directional, standards-based and automated exchange of health information with Regional Health Clinical Information exchange organizations emerging in NY State. These activities will enable NY to assess the efficacy of these entities in improving both public health and clinical practice through improved situational awareness and decision support.

ACKNOWLEDGEMENTS

We are indebted to Robert L. Burhans, Director of Health Preparedness, and Dr. Dale Morse, MD, Director of the Office of Science, NYSDOH, for their leadership and support. We also acknowledge the Western NY Health Alliance, New York Association of County Health Officials, hospitals, and local health departments of NY State for their contributions to the success of CDEX and HCS. This publication was supported by Cooperative Agreement Numbers U50/CCU223671 and U90/CCU216988 from the U.S. Centers for Disease Control and Prevention (CDC) and by Cooperative Agreement Number 6U3RHS05934 from the U.S. Health Resources and Services Administration (HRSA). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC or HRSA.

QUESTIONS FOR DISCUSSION

1. What are the key attributes of a model informatics framework for health information exchange (HIE)? How do they enable and support Public Health Emergency Preparedness (PHEP)? Discuss.
2. From your readings in this chapter, list and prioritize five technical, or architectural, attributes of a model informatics framework that must be in place to achieve and sustain an operational HIE. Classify each as a foundational element or as an element that depends upon and extends the functionality of foundational elements and discuss how they interoperate with each other to improve PHEP.
3. Now list and prioritize five key non-technical attributes of a model informatics framework that must be in place to achieve and sustain an operational HIE. Discuss the inherent difficulties you expect would be encountered in implementing them. Recognizing that implementing these attributes might require a great deal of inter- and intra-organizational change, what are the overarching benefits that would persuade decision makers to undertake the necessary changes to assure their implementation?

REFERENCES

Agency for Healthcare Research and Quality. National Hospital Available Beds for Emergencies and Disasters (HAvBED) System: Final Report, AHRQ Publication No. 05-0103, Rockville, MD. December 2005; http://www.ahrq.gov/prep/havbed/.

Asch, S. M., Stoto, M., Mendes, M., Valdez, R., Burciaga, G., Meghan, E., Halverson, P. and Lurie, N., 2005. A review of instruments assessing public health preparedness, *Public Health Reports*. **120**(5): 532–542.

Assoc. State Territ. Health Off (ASTHO). Exercising the stockpile: lessons learned and tools for application. 2004; http://www.astho.org/pubs/Exercisingthestockpile.pdf.

Baker, E. L., Potter, M. A., Jones, D. L., Mercer, S. L., Cioffi, J. P., Green, L. W., Halverson, P. K., Lichtveld, M. Y., Fleming, D. W., 2005. The public health infrastructure and our nation's health, *Annual Review of Public Health*. **26**: 303–318.

Brailer, D., Development and Adoption of a National Health Information Network. Department of Health and Human Services, November 9, 2004.

Bravata, D. M., McDonald, K. M., Szeto, H., Smith, W. M., Rydzak, C., and Owens, D. K., 2004. A conceptual framework for evaluating information technologies and decision support systems for bioterrorism preparedness and response, *Medical Decision Making*. **24**(2): 192–206.

Burkle, F. M., 2003. Measures of effectiveness in large-scale bioterrorism events, *Prehospital Disaster Medicine*. **18**(3): 258–262.

Burkle, F. M. Jr., Hayden, R., 2001. The concept of assisted management of large-scale disasters by horizontal organizations, *Prehospital and Disaster Medicine*. **16**(3): 87–96.

Centers for Disease Control and Prevention (CDC). Continuation guidance for cooperative agreement on public health preparedness and response for bioterrorism – budget year five. 2005; http://www.bt.cdc.gov/planning/coopagreement/#fy05.

Centers for Disease Control and Prevention (CDC). Public Health Information Network (PHIN) Requirements V2.0. June 22, 2007.

Centers for Disease Control and Prevention (CDC). Public Health Preparedness: Mobilizing State by State, A CDC Report on the Public Health Emergency Preparedness Cooperative Agreement. February, 2008.

Centers for Disease Control and Prevention (CDC). The 10 Essential Services of Public Health, http://www.cdc.gov/od/ocphp/nphpsp/EssentialPublicHealthServices.htm#es1. Last accessed October 1, 2010.

Costich, J. F., Scutchfield, F. D., 2004. Public health preparedness and response capacity inventory validity study, *Journal of Public Health Management and Practice*. **10**(3): 225–233.

Department of Homeland Security (DHS). Top Off (TOPOFF), Exercise series: after action summary report for public release. 2003a.

Department of Homeland Security (DHS). Homeland Security Presidential Directive/HSPD 8. December 17, 2003b; http://www.whitehouse.gov/news/releases/2003/12/20031217-6.html.

Department of Homeland Security (DHS). The National Incident Management System (NIMS). March 1, 2004.

Department of Homeland Security (DHS). The National Planning Scenarios, Created for Use in National, Federal, State and Local Homeland Security Preparedness Activities. March, 2006.

Department of Homeland Security (DHS). Top Officials 4 (TOPOFF 4) After Action Quick Look Report (AAR/QL), National Exercise Program (NEP). November 19, 2007a. http://www.mipt.org/pdf/TOPOFF2AfterActionRpt.pdf.

Department of Homeland Security (DHS). Target Capabilities List, A Companion to the National Preparedness Guidelines. September, 2007b.

Department of Homeland Security (DHS). Homeland Security Exercise and Evaluation Program, Volume I: HSEEP Overview and Exercise Program Management. Revised February 2007c.

Department of Homeland Security (DHS). The National Preparedness Guidelines. September 13, 2007d; http://www.dhs.gov/xlibrary/assets/National_Preparedness_Guidelines.pdf.

Department of Homeland Security (DHS). HSEEP Volume V: Prevention Exercises. March 28, 2008.

FEMA Emergency Management Institute. Independent Study program IS-120.a: An Introduction to Exercises October 2010. http://training.fema.gov/IS/docs/IS%20Brochure.pdf.

Gartner. Key Issues for SOA Governance Technologies. 2008; http://www.gartner.com/DisplayDocument?ref=g_search&id=603407&subref=simplesearch.

Gotham, I., Eidson, M., White, D., et al., 2001. West Nile virus: A case study in how New York State health information infrastructure facilitates preparation and response to disease outbreaks, *Journal of Public Health Management Practice*. **7**(5): 75–86.

Gotham, I., Smith, P. F., Birkhead, G. S., Davisson, M. C., 2002. Policy Issues in Developing Information Systems for Public Health Surveillance of Communicable Diseases, In: O'Carroll, P. W., Yasnoff, W. A., Ward, M. E., Ripp, L. H., Martin, E. L., *Public Health Informatics and Information Systems*, Springer, New York, pp. 537–573.

Gotham, I., Smith, P., Birkhead, G., Davisson, M., 2003. Policy issues in developing information systems for public health surveillance of communicable diseases. In: *Public Health Informatics and Information Systems*. Springer-Verlag, New York, pp. 537–573.

Gotham, I., Sottolano, D., Hennessy, N., Napoli, J., Dobkins, G., Le, L., Burhans, R., Fage, B., 2007. An integrated information system for all hazards health preparedness and response, New York State Health Emergency Response Data System (HERDS), *Journal of Public Health Management and Practice*. 13(5): 487–497.

Health Information Technology Standards Panel (HITSP). Biosurveillance-Connecting to Clinical Care (HITSP/IS-02). ANSI website, October 1, 2007; http://publicaa.ansi.org.

AMA/APHA. Improving health system preparedness for terrorism and mass casualty events, Recommendations for action. July 2007. A Consensus Report from the AMA/APHA Linkages Leadership Summit, Chicago, July 7–8, 2006, New Orleans; http://www.ama-assn.org/ama1/pub/upload/mm/415/final_summit_report.pdf.

Koh, H. K., Elqura, L. J., Judge, C. M., and Stoto, M. A., 2008. Regionalization of local public health systems in the era of preparedness, *Annual Review of Public Health*. **29**: 205–218. .

Loonsk, J. W., McCarvey, S. R., Conn, L. A., and Johnson, J., 2006. The public health information network (PHIN) preparedness initiative, *Journal of the American Medical Informatics Association*. **13(**1): 1–4.

Lyznicki, J., Subbarao, I., Benjamin, G. C., and James, J. J., 2007. Developing a consensus framework for an effective and efficient disaster response health system: A national call to action, *Disaster Medicine and Public Health Preparedness*. **1**(Suppl 1): S51–S54.

Nelson, C., Lurie, N., Wasserman, J., 2007a, Assessing public health emergency preparedness: Concepts, tools, and challenges, *Annual Review of Public Health*. **28**: 1–18.

Nelson, C., Lurie, N, Wasserman, J, Zakowski,.S., 2007b, Conceptualizing and defining public health emergency preparedness, *American Journal of Public Health*. **97**(Suppl 1): S9–S11.

Nelson, C., Lurie, N., Wasserman, J., and Zakowski, S., 2007c, Editorial: Conceptualizing and defining public health emergency preparedness, *American Journal of Public Health*. **97**(Suppl 1): S1.

New York State Department of Health. Pandemic Influenza Plan. February 2006; http://www.health.state.ny.us/diseases/communicable/influenza/pandemic/docs/pandemic_influenza_plan.pdf. Accessed May 2007.

NYS Health Laws and Regulations. 2005a. Health Provider Network Access and Reporting Requirements NYCRR Title 10b; http://www.health.state.ny.us/nysdoh/phforum/nycrr10.htm.

NYS Health Laws and Regulations. 2005b. Disaster and Emergency Planning NYCRR Title 18b; http://www.health.state.ny.us/nysdoh/phforum/nycrr10.htm.

NYS Public health law 576-C. 2007. Electronic reporting of disease and specimen submission; http://public.leginfo.state.ny.us/menugetf.cgi.

O'Carroll, P., 2003. Introduction to public health informatics, policy issues in developing information systems for public health surveillance of communicable diseases. In: *Public Health Informatics and Information Systems*. Springer-Verlag, New York, pp. 3–15.

Organization for the Advancement of Structured Information Standards (OASIS). CAP 1.1 OASIS Standard, OASIS website. October 1, 2005a; http://www.oasis-open.org/committees/download.php/14759/emergency-CAPv1.1.pdf.

Organization for the Advancement of Structured Information Standards (OASIS). EDXL-DE 1.0 OASIS Standard, OASIS website. October 1, 2005b; http://www.oasis-open.org/committees/download.php/18772/EDXL-DE%201.0%20Standard.pdf.

Perry, H. N., McDonnell, S. M., Wondimagegnehu, Al., Nsubug, P., Chungong, S., Otten, Jr., M. W., Lusambadikassa, P. S., Thacker, S. B., 2007. Planning an integrated disease surveillance and response system: a matrix of skills and activities, *PMC Medicine*. **5**: 24.

Popovich, M., Henderson, J., Stinn, J., 2002. Information technology in the age of emergency public health response, *IEEE Engineering in Medicine*. **21**(5): 48–55.

Seid, M., Lotstein, D., Williams, V. L., Nelson, C., Leuschner, K. J., Diamant, A., Stern, S., Wasserman, J., and Lurie, N., 2007. Quality improvement in public health preparedness, *Annual Review of Public Health*. **28**: 19–31.

Tanielian, T., Ricci, K., Stoto, M., Dausy, D., Davis, L., Myers, S., et al., 2005. *Exemplary Practices in Public Health Preparedness*. Sponsored by the US Department of Health and Human Services Office of the Assistant Secretary for Public Health Emergency Preparedness, Contract No.: 282-00-000-T11, RAND Corporation, RAND Center for Domestic and International Health Security, Santa Monica, California.

Weiner, E. E., Trangenstein, P. A., 2007. Informatics solutions for emergency planning and response, *Medinfo*. **12**(Pt 2): 1164–1168.

Zeng, C. H. C., Larson, C., Eidson, M., Gotham, I., Lynch, C., Asher, M., WNV-BOT Portal project summary, ACM International Conference Proceeding Series, Volume 262, Proceedings of the 2004 annual national conference on Digital Government Research, Seattle, Washington, May 24–26, 2004.

Zeng, D., Chen, H., Lynch, C., Eidson, M., Gotham, I., 2005. Infectious Disease Informatics and Outbreak Detection, In: Chen, H., Fuller, S., Friedman, C., Hersh, W. *Medical Informatics: Knowledge Management and Data Mining in Biomedicine,* Springer, New York, pp. 359–395.

SUGGESTED READING

O'Carroll, P. 2002. Public Health Informatics and Information Systems. Springer, New York.

Chen, H. 2005. Medical Informatics – Knowledge Management and Data Mining in Biomedicine. Springer, New York.

Whitten, J. 2005. Systems Analysis and Design Methods, 7th edition. McGraw-Hill/Irwin, New York.

ONLINE RESOURCES

Public Health Information Network. http://www.cdc.gov/phin/

Health Alert Network. http://www.phppo.cdc.gov/HAN/Index.asp

COMCARE Data Standards. http://www.comcare.org/Data_Standards.html

National Incident Management System. http://www.fema.gov/pdf/emergency/nims

OASIS Emergency Management Technical Committee, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emergency

Office of the National Coordinator for Health Information Technology (ONC). http://www.hhs.gov/healthit/

Data & Technical Standards: Health Information Technology Standards Panel (HITSP). http://www.hitsp.org/

Chapter 19

SYSTEM EVALUATION AND USER TECHNOLOGY ADOPTION

A Case Study of BioPortal

PAUL JEN-HWA HU[1,*], DANIEL ZENG[2,3], and HSINCHUN CHEN[2]

CHAPTER OVERVIEW

The surveillance of infectious disease and epidemic outbreaks has become increasingly challenging for public health professionals. Monitoring infectious disease and epidemic outbreaks is an information-intensive task that can be supported by effective data gathering, integration, analysis, and visualization. The BioPortal system, a Web-based portal that provides convenient access to distributed, cross-jurisdictional data about various data sets related to infectious diseases, offers just such an informatics environment. This chapter attempts to raise awareness of the importance and difficulty of system evaluations and user studies of advanced infectious disease informatics or biodefense systems. To illustrate the fundamental system evaluation and user study aspects, including the research method, study design, measurement instruments, and task design, we report two empirical studies of BioPortal. This chapter provides a non-technical perspective on the ongoing dialogues between public health and IT researchers, including the recognition that technical merits and novelty alone do not suffice to guarantee or sustain system success; it instead demands appropriate system design/functionality

[1*] *Department of Operations and Information Systems, David Eccles School of Business, University of Utah, Salt Lake City, UT, USA, paul.hu@business.utah.edu*

[2] *Department of Management Information Systems, University of Arizona, Tucson, AZ, USA*

[3] *Institute of Automation, Chinese Academy of Sciences, Beijing, China*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_19,

and sound understanding of user behaviors. System evaluation and user study research thus can play a significant role in infectious disease information research.

Keywords: System evaluation; User studies; BioPortal; Infectious disease informatics; Biodefense systems

1. INTRODUCTION

The surveillance of infectious disease and epidemic outbreaks has become increasingly challenging for public health professionals (Ericsson and Steffen, 2000). Epidemic episodes of severe acute respiratory syndrome, foot-and-mouth disease (FMD), and West Nile virus (WNV), as well as potential outbreaks of avian influenza, have created enormous concerns both domestically and internationally (Thacker et al., 2001; Li et al., 2004). Bioterrorism threats also lurk on the horizon. The fallout of the September 11 and subsequent anthrax attacks in the United States made bioterrorism a prominent threat to national security (Lane et al., 2001; Richmond and McKinney, 2007), adding complexity to the already challenging surveillance of infectious disease and epidemic outbreaks. Terrorists with biomedical or biochemical competencies also might attack people living in target geographic areas by deliberately developing and disseminating an infectious disease using biological agents (Siegrist, 1999).

Monitoring an infectious disease or epidemic outbreak demands significant amounts of information and therefore could be supported by advanced systems that focus on infectious disease informatics or biodefense and support effective data gathering, integration, analysis, and visualization. Developing such complex systems requires addressing the issues surrounding the diverse, heterogeneous, and voluminous data stored in various organizations, systems, or repositories, which span jurisdictional constituencies and incumbent administrative structures, both horizontally and vertically (Pinner et al., 2003). Yasnoff et al. (2001) advocate fruitful collaborations among researchers and practitioners in public health and information systems. Yet though technological breakthroughs continue at a rapid rate, system evaluations have not advanced at the same pace. To realize the full potential of advanced infectious disease informatics and biodefense systems, we require effective forms of evaluation, which Rossi and Freeman (1989) define as the systematic application of social science research procedures to judge and improve the way information resources are designed and implemented. Yet as Davis et al. (1989, p. 982) note, “As technical barriers disappear, a pivotal factor in harnessing this expanding power of computer technology becomes our ability to create applications that people are willing to use." Contrary to the general perception, or misperception, evaluation is complex; in this case, it entails the intersection of infectious disease informatics, computer-based systems, and empirical methods (Friedman and Wyatt, 2006). Thus, as extant literature shows, the disparity between the benefits promised by an advanced system and the utilities actually received or perceived by users continues to widen to alarming proportions.

Both methodological rigor and theoretical premises also are crucial to system evaluation. Observations of early deployments of innovative technologies or advanced information systems seem to suggest that failures exceed successes; a case in point is the pioneer technology that enabled digital government before the turn of the century. As Heeks (1999) summarizes, approximately 20–25% of these initiatives were abandoned immediately after their implementation, and a further 33% fell short of their major objectives. In the remainder of cases, most experienced great difficulty in achieving sustainable success. In a similar way, the sustainable success of advanced infectious disease informatics or biodefense systems for the adopting institution or public health research community as a whole demands favorable outcomes and positive assessments with respect to the monitoring and analysis of infectious diseases or epidemic outbreaks by public health researchers and practitioners.

Information systems attempt to improve user task performance (measured in effectiveness or efficiency) and enhance organization competitiveness. System development is engineering-oriented, aimed at solving problems in the target domain. Most previous infectious disease informatics research emphasizes the construction of novel or better solutions for users' information searches or analysis tasks as means to create advanced artifacts, such as models, methods, techniques, algorithms, or systems (i.e., instantiations). Again, system evaluation receives far less attention and usually takes place in an ad hoc manner. Technical advancements and novelty cannot guarantee the success of an infectious disease informatics or biodefense system; users often define its success, because a system's utilities demand both system design/functionality and user behaviors and assessments. Any infectious disease informatics or biodefense system therefore must be thoroughly examined, and its effectiveness or utilities must be systematically evaluated. Our literature review reveals several essential evaluation dimensions, including user task performance effectiveness and efficiency, the system's usefulness and ease of use, and user satisfaction (e.g., DeLone and McLean, 1992; Hu, 2003).

This chapter highlights specifically the importance of system evaluation and attempts to foster awareness of the challenges and difficulties of system evaluation endeavors. According to Gediga et al. (2002), a system evaluation might be descriptive or predictive; the former describes the status and problems of a system, whereas the latter seeks to uncover recommendations for enhancements or future system developments. We focus on descriptive evaluations and examine the status of a working infectious disease informatics environment, called the BioPortal system, according to user task performance effectiveness and efficiency, perceptions of its usefulness and ease of use, and users' satisfaction with the information support, as well as their intention to use the system in the near future.

BioPortal is a Web-based system that provides public health researchers and practitioners with convenient access to distributed, cross-jurisdictional data about several infectious diseases, such as WNV, FMD, and botulism (Zeng et al., 2004). Using seamless data integrations across various data formats and system architectures/platforms, BioPortal offers advanced spatiotemporal data analysis functionalities and novel visualizations that display analysis results in an effective, intuitive, and easily comprehensible manner.

To illustrate the fundamental aspects of the system evaluation, we report on two evaluation studies. The first, a controlled experiment that involves 33 graduate students, uses a spreadsheet-based program for benchmark purposes. The second study involves a field evaluation that targets three highly experienced public health professionals affiliated with the State Health Services department in the United States. In addition to generating empirical evidence that suggests the practical utilities of BioPortal, our evaluation sheds light on specific areas for improvement to the current design and functionalities. We first provide an overview of BioPortal, then detail each evaluation study, and finally highlight some important results.

2. AN OVERVIEW OF BIOPORTAL

BioPortal (http://www.bioportal.org) is an integrated, cross-jurisdictional, Web-based portal that has been operational for testing and research purposes since 2004. BioPortal is loosely coupled with several state public health information systems in California and New York; these systems transmit WNV/botulism information through secure links to BioPortal using mutually agreed protocols. The integrated data reside in a data store internal to BioPortal, which automatically retrieves data items from sources (e.g., USGS) and saves them in the data store.

The architectural design of BioPortal consists of three main components: a Web portal, a data store, and a communication backbone. The Web portal component constitutes the user interface and provides several essential functionalities that allow users to search or query available infectious disease-related data sets, visualize the results with novel spatiotemporal visualization designs, access analysis or prediction functions, and explore an advanced alerting mechanism.

The data central to BioPortal come from different organizations and vary in structure and format. To ensure necessary data interoperability, it uses HL7 standards as the primary format. In general, contributing institutions transmit data to BioPortal as HL7-compliant XML messages through a secure network connection. After receiving these XML messages, BioPortal stores them directly in its data store. This HL7 XML-based design is advantageous compared with an alternative design that uses a consolidated database, which must consolidate and maintain the data fields for all data sets in the data store. To alleviate the computational requirements and probable performance bottlenecks associated with this HL7 XML-based approach, a core set of data fields supports frequent queries, then extracts them from all XML messages and stores them in a separate database table for fast retrieval. The system strikes a desirable balance between the need for maximal access to disease tracking and confidentiality-related concerns, as well as the risks of jeopardizing data reporting to the system.

The communication backbone component uses a collection of source-specific "connectors" to link to various sources. For example, data from New York's HIN system are transmitted to BioPortal in a "push" manner, such that HIN sends secure public health information network messaging system (PHIN MS) messages to BioPortal at prespecified time intervals. The connector on BioPortal runs a data receiver daemon to listen for incoming messages. When a message is received, the connector examines the data integrity syntactically and normalizes the data, then stores the verified message in the internal data store through its data ingest control module. Other data sources, including USGS, may use a "pull-type" connector that periodically downloads data from the source Web site and examines and stores the data in the data store. Overall, the communication backbone component contains data receiving/sending functions, performs source-specific data normalization, and provides data encryption capabilities when necessary.

3. AN EXPERIMENT-BASED EVALUATION STUDY AND KEY RESULTS

This evaluation study addresses two questions: Will users' system choice significantly affect their task performance, perceptions of the system's usefulness and ease of use, and satisfaction with the information support? Will the use of BioPortal generate improved task performance, more favorable perceptions of the system's usefulness and easy of use, and higher user satisfaction with the information support? We conducted a controlled experiment to examine individual task performance and assessments associated with BioPortal. For benchmark purposes, our study also includes a spreadsheet-based program that mimics existing systems commonly used by public health researchers or practitioners for their surveillance.

3.1 Hypotheses

The value or utilities of BioPortal must be reflected in the surveillance task performance achieved by researchers or practitioners, who often analyze vast amounts of rich, detailed, diverse data describing or related to particular infectious diseases, geographic or environmental contexts, or essential population characteristics. Specifically, the use of BioPortal should generate significant improvements in user task performance, as measured by effectiveness or efficiency. Effectiveness measures often rely on analysis accuracy, such as whether a user produces more accurate results when supported by BioPortal than by an existing system. Efficiency can be assessed by the amount of time a user needs to complete an analysis task. Public health professionals constantly compete against time and must provide analysis results or issue alerts in a timely manner. Practitioners and researchers thus should value BioPortal more if it enables them to identify and respond to potential threats, health hazards, or epidemic outbreaks with increased accuracy and timeliness.

Users' assessments are also essential; individual perceptions of BioPortal's usefulness and ease of use might affect their system use. Perceived usefulness refers to the extent to which a user considers BioPortal useful in his or her work role (Davis, 1989), whereas perceived ease of use denotes the degree to which the user considers his or her use of BioPortal to be free of effort (Davis, 1989). All else being equal, people are more likely to use a system that is easy to use. We expect the use of BioPortal to be voluntary; therefore, perceived usefulness and ease of use represent crucial precursors.

User satisfaction represents another critical aspect of system evaluation. We focus on user information satisfaction, which here refers to the user's satisfaction with BioPortal with respect to information requirements (Ives et al., 1983). Our decision to focus on user information satisfaction rather than general end-user satisfaction is primarily based on the distinct importance of information support in the targeted public health context. If it is effective, users should exhibit greater information satisfaction with BioPortal. Accordingly, we test the following hypotheses:

H1: The choice of system has a significant influence on the user's analysis accuracy.

H2: The analysis accuracy associated with BioPortal is significantly greater than that of the benchmark spreadsheet program.

H3: The choice of system has a significant influence on the amount of time the user needs to complete an analysis task.

H4: The amount of time a user needs to complete an analysis task is significantly less when supported by BioPortal than by the benchmark spreadsheet program.

H5: The choice of system has a significant influence on the user's perception of system usefulness.

H6: Users perceive BioPortal as more useful than the benchmark spreadsheet program.

H7: The choice of system has a significant influence on the user's perception of the system's ease of use.

H8: Users perceive BioPortal as easier to use than the benchmark spreadsheet program.

H9: The choice of system has a significant influence on the user's satisfaction with the information support.

H10: Users exhibit greater satisfaction with the information support from BioPortal than with that from the benchmark spreadsheet program.

3.2 Experimental Design

We adopt a randomized between-groups factors design, in which we define system at two levels: BioPortal versus a benchmark spreadsheet-based program. The subject's general knowledge about public health represents the other factor, also defined at two levels – high versus low. Our design supports tests of our hypotheses and allows for direct comparisons of BioPortal and the benchmark program in terms of user task performance and assessment. In our experiment, we randomly assign subjects to BioPortal or the benchmark system, but not both, and maintain a balance in the system assignment to attain a comparable number of subjects in each experimental condition.

3.3 Measurements

We take a "gold-standard" approach to examine the accuracy of the subject's analysis of each task. Three experts individually examined each analysis task carefully and generated a recommended result. We consolidate these recommendations to produce a gold-standard result for each task. In turn, we can measure the accuracy of a subject's analysis of a particular task according to the corresponding gold-standard result and on the basis of a ten-point scale, such that 1 indicates "completely incorrect" and 10 is "completely correct." All partially completed tasks receive a score of 1. We measure task performance efficiency using the amount of time a subject took to complete an analysis task. Our study design includes a 50-min time constraint, which is appropriate according to the results of our pilot study. We explicitly informed each subject of this time constraint before he or she started the experiment tasks. We use items adapted from previously validated scales (Davis, 1989) to measure perceived usefulness and perceived ease of use. Furthermore, to measure user information satisfaction, we use items from Ives et al. (1983). All question items employ a seven-point Likert scale, with 1 being "strongly disagree" and 7 being "strongly agree." In Appendix 1, we list the items used in the reported evaluation studies.

3.4 Subjects

Our subjects are graduate students from the business school or public health school at a major research university located in the United States. They participated in the study voluntarily and differ notably in their general knowledge about public health, which is high among public health students and low among business school students. All subjects received compensation for their time and effort and rendered their consent before taking part in the experiment.

3.5 Experimental Tasks

Several highly experienced public health researchers and professionals assessed the design of the particular analysis tasks to be completed by subjects in the experiment. Specifically, we created six analysis scenarios and designed a total of 11 tasks, ranging from simplistic frequency counts to fairly complex trend detection or pattern identification. In Appendix 2, we list the analysis scenarios and tasks used in the experiment.

3.6 Data Collection

We administered the experiment to subjects individually or in small groups of two or three persons. We used a scripted document to inform all subjects explicitly of the study's purpose, experimental procedure, and how we would analyze and manage the data collected in the experiment. We specifically addressed concerns about information privacy and ensured subjects that we would perform data analyses at an aggregate level, not in any personally identifiable manner.

3.7 Evaluation Results

Our experiment involves 33 subjects; among them, 17 used BioPortal and the remaining were supported by the spreadsheet system. In the BioPortal group, nine subjects are public health students and the remaining eight subjects are from the business school. In the spreadsheet system group, seven subjects are public health students, and nine subjects are from the business school. We have 20 male and 13 female subjects; our subjects are highly comparable in their key demographic characteristics (including age and education) and self-report similar general computer efficacy and Internet usage.

We use the corresponding gold-standard result to evaluate the accuracy of each analysis task performed by a subject. We then aggregate each subject's analysis accuracy across all the tasks and use the resulting overall accuracy to test H1 and H2. According to our results, system choice (BioPortal versus the spreadsheet program) has a significant effect on the subject's analysis accuracy (F-value = 8.46; *p*-value < 0.01). The accuracy recorded by BioPortal (mean = 81.94, SD = 21.23) is significantly higher than that of the benchmark program (mean = 61.19, SD = 17.92; *p*-value < 0.01). Thus, our data support H1 and H2. System choice also has a significant effect on the amount of time required to complete an analysis task (F-value = 16.19; *p*-value < 0.01). Subjects supported by BioPortal complete a task significantly faster (mean = 36.28 min, SD = 11.33 min) than their counterparts using the spreadsheet program (mean = 48.23 min, SD = 5.07 min; *p*-value < 0.01). Hence, our data support H3 and H4.

According to our results, system choice has a significant impact on subjects' perceptions of the system's usefulness (F-value = 6.45; *p*-value < 0.05). Subjects perceive BioPortal as easier to use (mean = 2.31, SD = 1.06) than the spreadsheet program (mean = 3.24, SD = 0.88), and the difference is significant at the 0.01 level. Hence, our data support H5 and H6. Similarly, the effect of system choice on subjects' perceptions of the system's ease of use is significant statistically (F-value = 7.01; *p*-value < 0.05); that is, subjects consider BioPortal easier to use (mean = 2.31, SD = 1.06) than the spreadsheet program (mean = 3.24, SD = 0.88), significant at the 0.01 level. Hence, our data support H7 and H8. Finally, system choice has a significant influence on subjects' satisfaction with the information support (F-value = 12.01; *p*-value < 0.01). Subjects using BioPortal exhibit higher information satisfaction (mean = 2.34, SD = 1.02) than those using the spreadsheet program (mean = 3.68, SD = 1.23), with a statistically significant difference (*p*-value < 0.01). Therefore, our data support H9 and H10.

Overall, the data support all our hypotheses, suggesting that system choice in general has important influence on users' task performance and assessments. In particular, the use of BioPortal can improve the surveillance task performance, as measured by analysis accuracy and time requirements. Our subjects consider BioPortal more useful and easier to use than the benchmark spreadsheet program, and they exhibit greater satisfaction with its information support.

4. A FIELD USER STUDY AND KEY RESULTS

We performed a field user study that targets three experienced public health professionals. Our objective was to gather and analyze their task performance and evaluative responses in the work context. In the following sections, we describe our research questions, measurements, subjects, tasks, and evaluation results.

4.1 Research Questions

The objective of this field user study is to examine the value and utilities of BioPortal in its target public health settings. Specifically, we address the following questions:

1. Does BioPortal provide sufficient query criteria or support (e.g., different query modes)? If not, what additional query criteria or support should be included?
2. How useful are BioPortal's aggregate views? Specifically, how do these views help public health professionals perform their analyses or problem-solving tasks?
3. How useful is the GIS tool in BioPortal? How can this tool be enhanced and better support public health professionals' analyses or problem-solving tasks?
4. How useful is the Timeslider in BioPortal? Does this tool enhance health professionals' analyses or problem solving?
5. How do public health professionals assess BioPortal's usefulness and ease of use? What is their satisfaction level with its information support and their intention to use it in the near future?

4.2 Measurements

Our study focuses on analysis accuracy, task completion time requirements, and users' satisfaction with the information support provided by BioPortal, as well as their intentions to use it in the near future. Similar to the experiment-based evaluation study, we measure analysis accuracy using the gold-standard results established by a panel of domain experts who also assisted in our task designs. The task completion time requirement equals the amount of time elapsed between the beginning and the completion of a task. We continue using the items adapted from Ives et al. (1983) to measure user information satisfaction and employ items from Davis (1989) to assess practitioners' intentions to use BioPortal in their work context. The specific analysis scenarios and tasks used in the field study are in Appendix 3.

4.3 Subjects

Three public health professionals took part in our study voluntarily, one woman and two men. Our subjects are between 31 and 36 years of age and have doctoral degrees in public health or related disciplines. According to analyses of their self-assessments, these practitioners have fairly good general computer efficacy and use the Internet frequently and routinely. Each subject is highly knowledgeable about epidemiological practices and shows great confidence in analyzing and interpreting data about different infectious diseases or epidemic outbreaks.

4.4 Tasks

With the assistance of several domain experts, we designed a set of analysis scenarios and tasks that closely resemble some of the surveillance tasks common in public health. In Appendix 3, we list the specific scenarios and tasks used in this study. We emphasize user task performance (measured by analysis accuracy and task completion time) and satisfaction with the information support offered by BioPortal. We also collect from each practitioner his or her qualitative assessment of BioPortal, using several semi-structured questions.

4.5 Evaluation Results

As a group, our subjects accomplish satisfactory analysis accuracy, as suggested by an average score of 1.91 on a two-point scale, on which 2 indicates completely correct, 1 means partially correct, and 0 is incorrect. On average, a subject took 1 h and 12 min (SD = 11.67) to complete all the tasks, which the subjects consider notably shorter than the time commonly required for their analyses using existing systems. The analysis of their evaluative responses shows high user information satisfaction; i.e., mean = 5.78, SD = 1.12, seven-point Likert scale, on which 7 is "strongly agree." The subjects also exhibit high intentions to use BioPortal in the near future; i.e., mean = 6.0, SD = 1.24, seven-point Likert scale on which 1 equals “strongly disagree” and 7 indicates “strongly agree.” Judging by the respective mean values, all of which are considerably higher than 4 (i.e., neutral), the three practitioners exhibit positive assessments of BioPortal’s usefulness and ease of use and strong intentions to use it in the near future.

Furthermore, most of the qualitative assessments of BioPortal are positive. One practitioner noted, “Capability of BioPortal is huge – could link to state data and would have a great foundation. Serotypes also useful for linking cases epidemiologically.” In their evaluations, they indicate that BioPortal has an adequate design and is easy to use. A sample comment provides an illustration: “design is one of its strengths – very intuitive and user friendly.” Another subject made a similar remark: “BioPortal is a little easier to use than existing syndromic surveillance systems.” The practitioners appear particularly fond of the spatial temporal visualizer in BioPortal. According to one subject: “The GIS thing is big! Especially West Nile virus is big in [our] County – [we] would like to be able to look at the geographic spread. This could influence mosquito intervention, and see movement over time, when cases stop and/or pop up somewhere else. Would also be good for rabies.” The other subjects provide similar favorable assessments; one noted that “Just being able to pick a time period (for example, 1 week) and see how it unfolds. Also seeing the faded out cases very helpful.” The other commented further that “Aggregated views are good for overall picture of data and for answering specific questions by choosing 2 × 2 table variables.” Furthermore, the subjects expressed great appreciation for the built-in hot spot analysis in BioPortal; for example, “Hotspot analysis is instrumental to what this user does everyday – the job function is to detect any health event in the community before diagnoses” and “The hotspot analysis tool embedded in STV is very useful. When user clicked a mock data set, it went straight to STV, then used the tool to pick SatScan and parameters. Would be easier to use that way to change baseline.”

5. CONCLUSION

In this chapter, we discuss system evaluation and user study issues in the context of infectious disease informatics. Using two empirical evaluation studies of the BioPortal system, we present in detail key elements of system evaluation and user study research, covering the overall research methodology and critical considerations, together with study design, measurement instruments, and task design, as well as applicable system performance measures such as task performance efficiency, user information satisfaction, and usability.

To the best of our knowledge, systematic studies of system evaluation and user study issues in infectious disease informatics have been sparse. We argue, based on our long and well-documented experience with IT adoption and evaluation studies in domains outside of infectious disease informatics, that this lack of attention could seriously undermine both research and practical efforts to attain sustainable success in the development and adoption of infectious disease informatics systems. We hope this chapter helps raise greater awareness of the importance of non-technical evaluation and adoption considerations in the public health community.

From a research perspective, we believe that ongoing dialogues between public health officials and information systems evaluation researchers could lead to new and exciting results. The studies presented in this chapter might be conceptualized as applications of known evaluation frameworks in the domain of infectious disease informatics. Further collaborative research with potentially high impact could deliver new evaluation frameworks tailored for other infectious disease informatics, depending on the specific characteristics of user behavior, task environment, organizational structure, and technology needs in this domain.

QUESTIONS FOR DISCUSSION

1. What are the key factors facilitating/hindering wide adoption of advanced IDI tools and systems in practice?
2. How can one effectively evaluate surveillance analytics functions provided by biosurveillance systems, with users in the loop?
3. What are the key factors one needs to consider when designing biosurveillance tasks to evaluate the efficacy of the system?
4. What are the key measurement instruments relevant to biosurveillance system evaluation?
5. What are the unique challenges to technology adoption in biosurveillance, relative to technology adoption in the commercial world such as e-commerce and enterprise systems?

Appendix 1: Listing of Question Items Used in the Evaluation Studies (Illustrated Using BioPortal)

Perceived Usefulness

1. Using BioPortal allows me to complete an analysis task more quickly.
2. Using BioPortal can improve my analysis of public health problems or trends.

3. Using BioPortal can make me more productive in analyzing public health problems or trends.
4. Using BioPortal can make me more effective in analyzing public health problems or trends.
5. Using BioPortal can make my analysis of public health problems or trends easier.
6. I would use BioPortal in my analysis of public health problems or trends.
7. Overall, I find BioPortal useful for supporting my analysis tasks.

Perceived Ease of Use:

1. Learning to operate BioPortal would not be difficult for me.
2. I find it easy to use BioPortal to analyze what I need to when examining public health problems or trends.
3. I find it easy to learn the functions of BioPortal.
4. My interaction with BioPortal is understandable.
5. I find BioPortal flexible to interact with in analyzing public health problems or trends.
6. It would be easy for me to become skillful in using BioPortal.
7. Overall, I find BioPortal easy to use.

User Information Satisfaction (UIS):

1. BioPortal offers valuable utility in my analysis of public health problems or trends.
2. I can understand the functions of BioPortal.
3. Using BioPortal can quickly generate the analysis results that I need.
4. The analysis results by BioPortal are reliable.
5. The visualization designs of BioPortal are good.
6. In general, I am satisfied with the response time of BioPortal.
7. Overall, I find the results generated by BioPortal relevant to my analysis of public health problems or trends.
8. The analysis results by BioPortal are accurate.
9. Overall, I have good control over using BioPortal to complete an analysis task.
10. BioPortal is flexible for supporting different analysis tasks in public health.

Intention to Use BioPortal:

1. When I have access to BioPortal, I would use it as often as needed.
2. To the extent possible, I intend to use BioPortal in my job.
3. Whenever possible, I would use BioPortal for my tasks.

Appendix 2: Listing of Analysis Scenarios and Tasks Used in the Experiment-Based Evaluation Study

Scenario 1: Examine data related to WNV.

Task 1: In 2002, which county in New York had the highest dead bird count?

Task 2: Of the three listed bird species, Bluejay, Crow, and House Sparrow, which had the highest number of positive cases of West Nile virus?

Scenario 2: Examine a correlation between botulism and gender.

Task 3: In California, for year 2001, did more men or more women suffer from botulism?

Task 4: In California, for year 2002, did more men or more women suffer from botulism?

Scenario 3: Determine the occurrence of foot-and-mouth disease in 2001 for three countries.

Task 5: In 2001, in which week(s) do the highest number of foot-and-mouth disease cases occur in Iran?

Task 6: In 2001, in which week(s) do the highest number of foot-and-mouth disease cases occur in Turkey?

Task 7: In 2001, in which week(s) do the highest number of foot-and-mouth disease cases occur in Argentina?

Scenario 4: Determine the location of the most intensive outbreak of WNV during 1999 in New York.

Task 8: During 1999, where (in which county) and when did the most intensive occurrence (i.e., highest number of cases) of West Nile virus happen in New York State?

Scenario 5: Describe the spread (geographically and over time) of dead crow sightings for New York in 2002.

Task 9: Please describe the spread (geographically and over time) of dead crow sightings in New York in 2002.

Scenario 6: Determine correlations between the incidence of WNV and dead bird occurrences and mosquito pool counts.

Task 10: Using the BioPortal system or the spreadsheets, as assigned, to investigate West Nile virus disease, can you determine if, during 2002, there is a correlation between the dead bird occurrences and mosquito pool counts?

Task 11: (Continued from Task 10) If so, what correlation do you observe?

Appendix 3: Listing of Analysis Scenarios and Tasks Used in the Field Evaluation Study

Scenario 1: BioPortal Website Functionalities

This scenario will focus on the use of the BioPortal Website. In this scenario, the user is asked to provide characteristics of the target data set. These characteristics include:

– The number of cases with certain syndromes within a time period.
– The date of the first case of a certain syndrome.
– Detailed case information.
– In this scenario, the user will make use of the following BioPortal functionalities: query, case detail display, aggregate view and advanced query.

Data Set: Scottsdale Health Center Chief Complaints

Task 1: Describe the number of positive cases in the current data set:

Task 2: Find the time coverage (first and last case dates) in this data set

Date of first case: _________; Date of last case: _________

Task 3: Identify the week with the highest number of cases.

Task 4: Identify how many female patients with GI syndrome can be found within this data set.

Task 5: Identify the patient ID of the first patient with botulism syndrome within the given data set.

Task 6: For the female patients in the age group 30–39, identify the top three syndromes besides “unknown”:

Scenario 2: Spatial-Temporal Visualizer (STV)

This scenario, presented in two parts, will focus on the use of the STV tool to visually inspect the data distribution in both space and time. The user will be asked to identify information such as the peak number of cases, the area with the highest number of cases, and temporal and spatial distribution trends. The user will make use of the following tools provided in STV: Time Slider, GeoMap, Periodic Pattern Tool, Histogram and Timeline Tool.

Scenario 2-A

Data Set: User Study Test Data set 1

Task 1: Start STV with the User Study Test Data set 1 and zip code boundary, isolate botulism cases (by removing other syndromes from the map), and then identify the day of week with the most botulism cases during the time period.

Task 2: View the case distribution in the Histogram tool and describe the temporal distribution.

Task 3: Examine the spatial distribution using moving time window and expanding time window techniques and describe the spatial movement trend of botulism cases.

Scenario 2-B

Data Set: Mesa Fire Department EMS data

Task 1: Start STV with Mesa EMS data (between September 1, 2006 and September 30, 2006) and zip code boundary, change the color of categories with similar colors to avoid ambiguity, and then identify the hours of the most trauma cases.

Task 2: Identify the zip codes with the most toxin/poison cases.

Task 3: Identify the address of the youngest cardio-respiratory case in the Apache Junction area (the easternmost zip code).

Task 4: Isolate the Apache Junction area and identify the day of week with the most general medicine cases:

Scenario 3: Hotspot Analysis

This scenario will target the use of the Hotspot Analysis tool embedded in the STV. Users will have two or three simulated data sets to evaluate. For each data set, the user will be asked to identify outbreaks and investigate case details using STV.

Data Sets: Under User Study Page [in our simulation process, the first 15–22 days are baseline data].

Task 1: Regarding 167 Hemo syndrome cases with one injected outbreak, describe the outbreak you discovered.

Task 2: Regarding 300 GI syndrome cases with one short-term outbreak, describe the outbreak you discovered.

REFERENCES

Davis, F.D. (1989) "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, 13, 319–339.

Davis, F.D., Bagozzi, P.R., and Warshaw, R. (1989) "User acceptance of computer technology: A comparison of two theoretical models," *Management Science*, 35, 982–1002.

DeLone, W.H., and McLean, E.R. (1992) "Information systems success: The quest for the dependent variable," *Information Systems Research*, 3(1), 60–95.

Ericsson, C.D., and Steffen, R. (2000). "Population mobility and infectious disease: The diminishing impact of classical infectious diseases and new approaches for the 21st century," *Clinical Infectious Diseases*, 31, 776–780.

Friedman, C.P., and Wyatt, J.C. (2006) *Evaluation Methods in Biomedical Informatics*. New York: Springer.

Gediga, G., Hamborg, K.-C., and Duntsch, I. (2002) "Evaluation of software systems," In *Encyclopedia of Library and Information Science*, 72 (Kent, A., and Williams, J. G. Eds.) Boca Raton: CRC Press, 166–192.

Heeks, R. (1999) *Information and Communication Technologies, Poverty and Development*. Manchester: Institute for Development Policy and Management, University of Manchester.

Hu, P.J. (2003) "Evaluating telemedicine systems success: A revised model," *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, Hawaii, January 3–6.

Ives, B., Olson, M., and Baroudi, J.J. (1983) "The measurement of user information satisfaction," *Communications of the ACM*, 26(10), 785–793.

Lane, H.C., LaMontagne, J., and Fauci, A.S. (2001). "Bioterrorism: A clear and present danger," *Nature Medicine*, 7(12), 1271–1273.

Li, Y., Yu, L.T., Xu, P., Lee, J.H., Wong, T.W., Ooi, P.L., and Sleigh, A.C. (2004) "Predicting super spreading events during the 2003 Severe Acute Respiratory Syndrome epidemics in Hong Kong and Singapore," *American Journal of Epidemiology*, 160, 719–728.

Pinner, R., Rebmann, C., Schuchat, A., and Hughes, J. (2003) "Disease surveillance and the academic, clinical, and public health communities," *Emerging Infectious Diseases*, 9, 781–787.

Richmond, J.Y., and McKinney, R.W. (2007). *Biosafety in Microbiological and Biomedical Laboratories*. Washington: U.S. Government Printing Office.

Rossi, P.H., and Freeman, H.E. (1989) *Evaluation: A Systematic Approach*, California: Sage.

Siegrist, D. (1999) "The threat of biological attack: Why concern now?" *Emerging Infectious Diseases*, 5, 505–508.

Thacker, S.B., Dannenberg, A.L., and Hamilton, D.H. (2001) "Epidemic intelligence service of the Centers for Disease Control and Prevention: 50 years of training and service in applied epidemiology," *American Journal of Epidemiology*, 154, 985–992.

Yasnoff, W.A., Overhage, J.M., Humphreys, B.L., and LaVenture, M.L. (2001) "A national agenda for public health informatics," *Journal of American Medical Informatics Association*, 8, 535–545.

Zeng, D., Chen, H., Tseng, L., Larson, C., Eidson, M., Gotham, I., Lynch, C., and Ascher, M. (2004) "West Nile virus and botulism portal: A case study in infectious disease informatics," in *Lecture Notes in Computer Science*, 3073 (Chen, H., Moore, R., Zeng, D., and Leavitt, J. Eds.) New York: Springer, 28–41.

SUGGESTED READING

Shneiderman, B., Plaisant, C., Cohen, M., and Jacobs, S. (2009). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 5th edition. Reading, MA: Addison-Wesley.

Friedman, C.P., and Wyatt, J.C. (2006). *Evaluation Methods in Biomedical Informatics*. New York: Springer.

DeLone, W.H., and McLean, E.R. (1992) "Information systems success: The quest for the dependent variable," *Information Systems Research*, 3(1), 60–95.

Bettencourt, L., Cintron-Arias, A., Kaiser, D., and Castillo-Chavez, C. "The Power of a good idea: quantitative modeling of the spread of ideas from epidemiological models," Physica A, Vol. 364, pp. 513–536, 2006.

ONLINE RESOURCES

http://www.bioportal.org: BioPortal system demonstrations.

Chapter 20

SYNDROMIC SURVEILLANCE FOR THE G8 HOKKAIDO TOYAKO SUMMIT MEETING

YASUSHI OHKUSA[1,*], TAMIE SUGAWARA[1], HIROAKI SUGIURA[2], KAZUO KODAMA[3], TAKUSHI HORIE[4], KIYOSHI KIKUCHI[5], KIYOSU TANIGUCHI[1], and NOBUHIKO OKABE[1]

CHAPTER OVERVIEW

We conducted syndromic surveillance during the G8 summit meeting held in Toyako, Hokkaido, July 7–9, 2008, as a counter-measure to bioterrorism attacks or other health emergencies. Surveillance actually started on June 23, 2 weeks prior to the G8 summit, and ended on July 23, 2 weeks after the closing of the meeting. Part of the syndromic surveillance for prescription drugs was fully automated, while the remainder was done manually through the Internet. Similarly, data on ambulance utilization was collected and included in the syndromic surveillance system. We also purchased data on OTC sales from two private research firms in Japan. In an effort to share the surveillance information and discuss whether further investigation was needed, virtual conferences were held and Hokkaido local government, local health departments and laboratory, National Institute of Infectious Diseases, and Ministry of Health, Labor and Welfare personnel were among the attendees. Information was collected automatically from 23 pharmacies on prescription drugs and manually entered for 71 pharmaceutical companies on drug sales. One fire department that covered the Toyako area and was in

[1] *National Institute of Infectious Diseases, Tokyo, Japan*
[2] *Sugiura Clinic, Japan*
[3] *Kodama Clinic, Japan*
[4] *Chiimiya-Horie Clinic, Japan*
[5] *Department of Pediatrics, Shimane Prefectural Central Hospital, Japan*

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*,
Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0_20,

charge of highlevel officers participated in the fully automated surveillance system, and seven other departments in the surrounding area conducted the manually-entered surveillance. OTC sales information was reported for 79 drugs with a delay of 1 day, and thus had to be processed manually. Health conditions were reported by 472 households that agreed to participate in the web-based survey; this data was analyzed automatically. Fortunately, we did not observe any suspected outbreaks during G8. However, local health departments investigated seven cases based on abberrances in ambulance utilization detected by the syndromic surveillance. Undoubtedly, a fully automated surveillance system is the best method for detecting an early signs of outbreak. Nevertheless, we had to use a semi-automated surveillance system during the G8 summit due to a limitation on our data collection. Our attempt at syndromic surveillance showed that it was useful and suggested that a routine and fully automated surveillance system, without manual data entry, would be needed for closer monitoring to catch signs of any suspected outbreak in the community. A routinized and fully automatic system without manual input is the next step for syndromic surveillance in Japan.

Keywords: Syndromic surveillance system; G8 summit meetings; Ambulance transfer; Prescription drug; School absenteeism; Full automatic

1. INTRODUCTION

Currently, when high profile events such as the G8 summit meetings, Olympic games, or other mass gatherings or important political events are held, syndromic surveillance is routinely performed as well [1, 2]. The first attempt at syndromic surveillance in Japan was conducted in 2000 when the Kyushu Okinawa G8 summit-related meeting was held in Fukuoka and Miyazaki [3]. However, it only reclassified notifiable or sentinel weekly reporting diseases enforced by Infectious Control Law into five syndromic categories; it did not monitor symptoms. The second attempt was conducted during the FIFA World Cup 2002 in Korea and Japan. The syndromic surveillance at that time required the reporting of symptoms of emergency hospitalized patients in five categories through a web site. However, its burden on the hospitals was too heavy and thus it was stopped 2 weeks after the final game in each stadium. Therefore, automated routine syndromic surveillance, which has been used in the U.S. or Taiwan, has not been performed officially to date.

In order to build up routine syndromic surveillance in Japan, research was started in 2004, and a prototype system was made. This system was officially used during the Hokkaido Toyako G8 summit meeting held July 7–9, 2008, in Japan as a counter-measure to bioterrorism attacks or other health emergencies. This chapter discusses the workable systems in Japan as well as some of our relevant experiments.

We performed three types of syndromic surveillance, i.e., for prescriptions, ambulance transfers, and OTC (non-prescribed, over-the-counter) drug sales, during the Hokkaido Toyako G8 summit meeting in July 2008. We also performed some experiments using data from Electronic Medical Records (EMR), orders for medical examinations, absenteeism at school, and nosocomial outbreaks. In the following sections, we explain these surveillances in detail. Before that, we will explain the background and situation concerning syndromic surveillance in Japan.

2. BACKGROUND

Infection Control Law in Japan has asked doctors to cooperate in syndromic surveillance for pandemic flu and smallpox since 2007. However, doctors have to report by typing the number of patients on the web site, as in the system used during the FIFA World Cup, or by sending a fax to local public health centers. It imposes a heavy burden of reporting, and thus it has not worked yet. Therefore, we need an automated system for routine syndromic surveillance.

On the other hand, medical services for outpatients are well developed due to universal public health insurance. Even patients who have mild symptoms can visit a clinic freely in Japan. Thus the monitoring of outpatients provides very timely information to detect unusual events. Conversely, EMRs haven’t had much penetration, less than 10% at clinics and 20% at hospitals. Moreover, neither HL7 nor other standards are used by EMRs. Therefore, it is very difficult to develop a syndromic surveillance system using EMRs such as those in the U.S. We would have to develop a system for each EMR and this has a heavy cost.

Privacy is a much bigger concern in Japan, in comparison with the U.S. Even zip codes, which are more specific than just the city, are not permitted to be used. Moreover, almost all hospitals and clinics have not connected EMRs to the outside even for public health purposes, so as to protect privacy.

On the other hand, systems surrounding EMRs, such as medical claim data at clinics or hospitals, or prescription data at pharmacies, are recorded electronically at a much higher rate than EMRs.

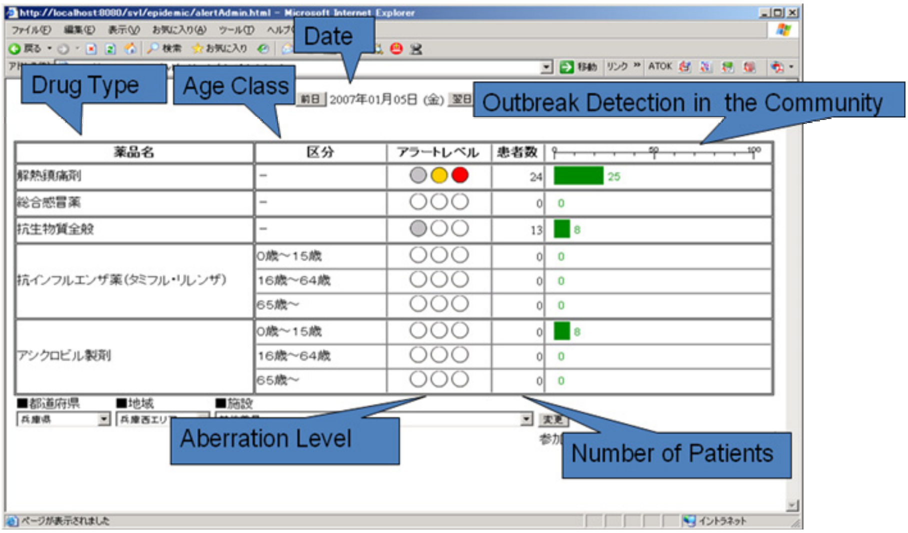
3. METHODS

3.1. Syndromic Surveillance for Prescriptions

In Japan, there are about 40,000pharmacies and almost half of the drugs prescribed are delivered through pharmacies. Almost all pharmacies record prescriptions electronically. We collaborated with EM SYSTEMS Co. Ltd., which is the leading company of the systems for pharmacies and, especially, provides the Application Server Provider (ASP) system to more than 3,000 pharmacies. The ASP system is very useful for syndromic surveillance because data transfer is unnecessary, and thus it can reduce costs dramatically and maintain confidentially.

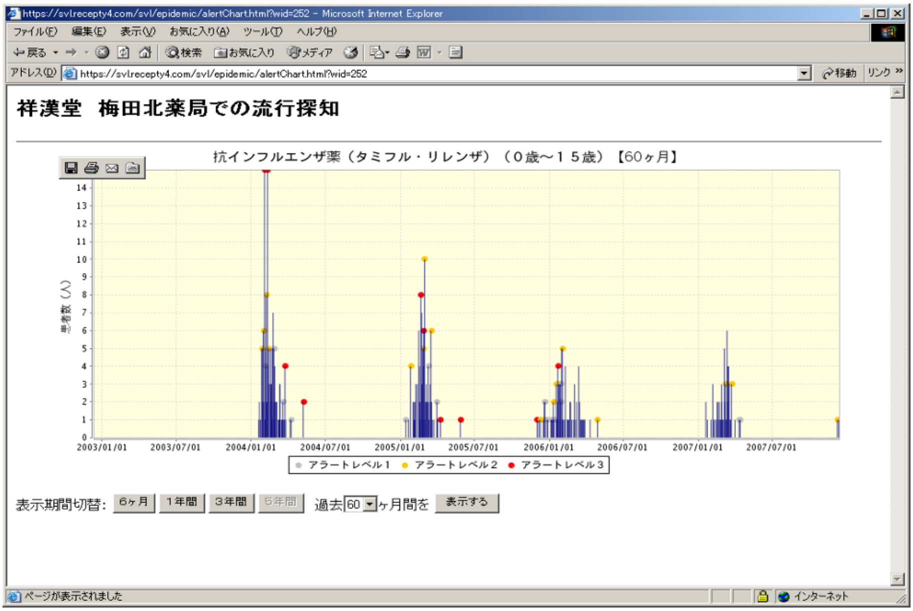
Because the system uses only prescription information and the symptoms of patients are not recorded, the syndrome categories used are the types of drugs. Currently, it monitors drugs for relief of fever and pain, drugs for common colds, antiviral drugs, anti-influenza virus drugs (except for Amantazine), and anti-Varicella-Zoster virus drugs. The last two are classified by age, with the categories including: less than 15, 16–64, older than 65 years old.

The data collection and analysis are operated automatically at night and results are shown on the home page on the secure Internet in the early morning. Figure 20-1 shows feedback on the home page for corporate pharmacies in the upper panel, and the lower panel is for public health centers or local government which has the responsibility to control health risk events. The left hand side shows the situation in each pharmacy. The first column shows the types of drugs. The second column shows alerts in each type of drug in this pharmacy at level three. The red circle means the highest level of alert was found. The third column summarizes the number of patients with each type of drug in this pharmacy. On the right hand side, it shows the proportion of pharmacies in this area which found alerts in each type of drug. If all pharmacies which are cooperating in the syndromic surveillance in this area find alerts in a type of drug on the same day, this green bar becomes longest and its score is 100. The buttons which are labeled as “Graph” provide figures of patients or outbreak detection in the community. Figure 20-2 provides an example of daily prescription of Tamiflu or Rirenza in a pharmacy for 5 years. If we find an aberration, the colored circles are marked. Figure 20-3 shows an example of daily outbreak detection in the community. Public health centers or local governments restrict access to the information at the community level and thus the feedback home page must be as shown in Figure 20-4.


http://localhost:8080/svl/epidemic/alertAdmin.html - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ
Date
Drug Type
Age Class
Outbreak Detection in the Community
Aberration Level
Number of Patients
前日 2007年01月05日 (金) 翌日
■都道府県 兵庫県
■地域 兵庫西エリア
■施設
変更
ページが表示されました
イントラネット

薬品名	区分	アラートレベル	患者数	0 50 100
解熱鎮痛剤	–	●●●	24	25
総合感冒薬	–	○○○	0	0
抗生物質全般	–	●○○	13	8
抗インフルエンザ薬(タミフル・リレンザ)	0歳～15歳	○○○	0	0
	16歳～64歳	○○○	0	0
	65歳～	○○○	0	0
アシクロビル製剤	0歳～15歳	○○○	0	8
	16歳～64歳	○○○	0	0
	65歳～	○○○	0	0

Figure 20-1. Feedback HP for each corporate pharmacy.


https://svl.recepty4.com/svl/epidemic/alertChart.html?wid=252 - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り メディア
アドレス(D) https://svl.recepty4.com/svl/epidemic/alertChart.html?wid=252
移動 リンク
抗インフルエンザ薬（タミフル・リレンザ）（０歳～１５歳）【60ヶ月】
患者数（人）
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
2003/01/01
2003/07/01
2004/01/01
2004/07/01
2005/01/01
2005/07/01
2006/01/01
2006/07/01
2007/01/01
2007/07/01
● アラートレベル１ ● アラートレベル２ ● アラートレベル３
6ヶ月
1年間
3年間
5年間
表示する
ページが表示されました
インターネット

祥漢堂　梅田北薬局での流行探知

表示期間切替: 6ヶ月 1年間 3年間 5年間 過去60ヶ月間を 表示する

Figure 20-2. Patients who were prescribed Tamiflu or Rirenza at each corporate pharmacy for 5 years.

Aberration is defined through the multiple regression model. Namely, we regress the number of patients on each type of drug on dummies for the week number (1–52, 53), the day of the week (Sunday–Saturday), post-holiday, and time trends such that:

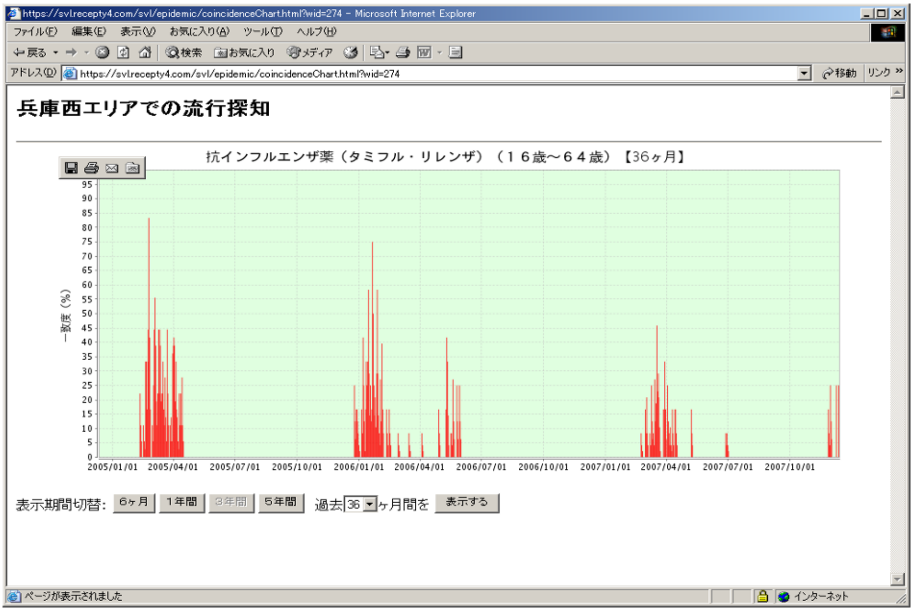
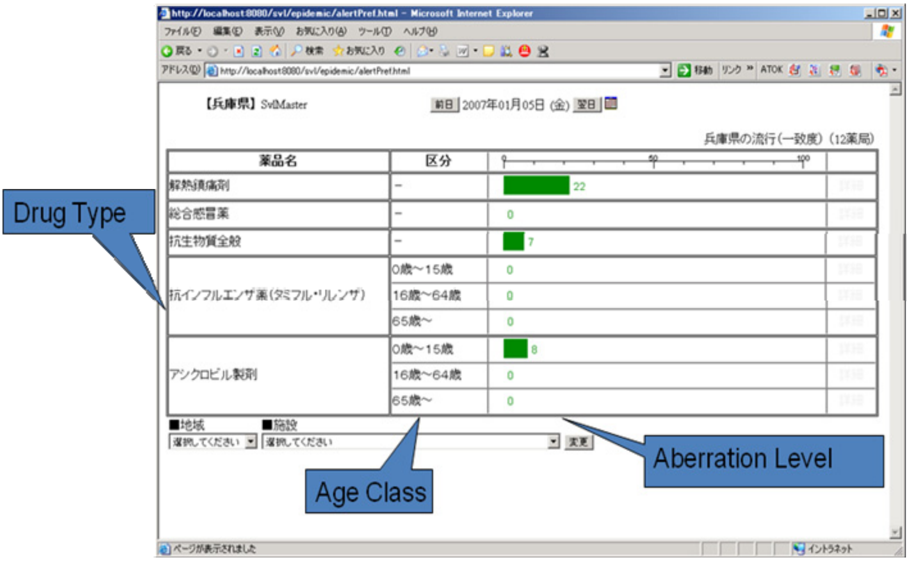

https://svl.recepty4.com/svl/epidemic/coincidenceChart.html?wid=274 - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) https://svl.recepty4.com/svl/epidemic/coincidenceChart.html?wid=274
兵庫西エリアでの流行探知
抗インフルエンザ薬（タミフル・リレンザ）（１６歳〜６４歳）【36ヶ月】
一致度（%）
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
2005/01/01
2005/04/01
2005/07/01
2005/10/01
2006/01/01
2006/04/01
2006/07/01
2006/10/01
2007/01/01
2007/04/01
2007/07/01
2007/10/01
表示期間切替: 6ヶ月 1年間 3年間 5年間 過去36ヶ月間を 表示する
ページが表示されました
インターネット

Figure 20-3. Outbreak detection in the community for Tamiflu or Rirenza in adults.


http://localhost:8080/svl/epidemic/alertPref.html - Microsoft Internet Explorer
アドレス(D) http://localhost8080/svl/epidemic/alertPref.html
Drug Type
Age Class
Aberration Level
ページが表示されました
イントラネット

【兵庫県】SvlMaster 前日 2007年01月05日 (金) 翌日

兵庫県の流行(一致度) (12薬局)

薬品名	区分	0 50 100	
解熱鎮痛剤	−	22	詳細
総合感冒薬	−	0	詳細
抗生物質全般	−	7	詳細
抗インフルエンザ薬(タミフル・リレンザ)	0歳〜15歳	0	詳細
	16歳〜64歳	0	詳細
	65歳〜	0	詳細
アシクロビル製剤	0歳〜15歳	8	詳細
	16歳〜64歳	0	詳細
	65歳〜	0	詳細

■地域 選択してください
■施設 選択してください 変更

Figure 20-4. Feedback HP for local government or public health center.

$$\text{Number of cases}_t = \alpha + \Sigma_i \beta_i (\text{Week No})_i + \Sigma_j \gamma_j (\text{Day-of-the-Week})_j + \eta(\text{the Day after Holiday}) + \theta t + \delta t^2 + \varepsilon_t \quad (20\text{-}1)$$

by using Poisson regression. Three criteria are used for aberration: low level, if the probability of the number of observed cases that occur is less than 2.5%, medium level, 1%, and high level, 0.1%.

When the Hokkaido Toyako G8 summit meeting was held, 32 pharmacies (approximately 5% of pharmacies) in the surrounding area of Toyako cooperated in this surveillance.

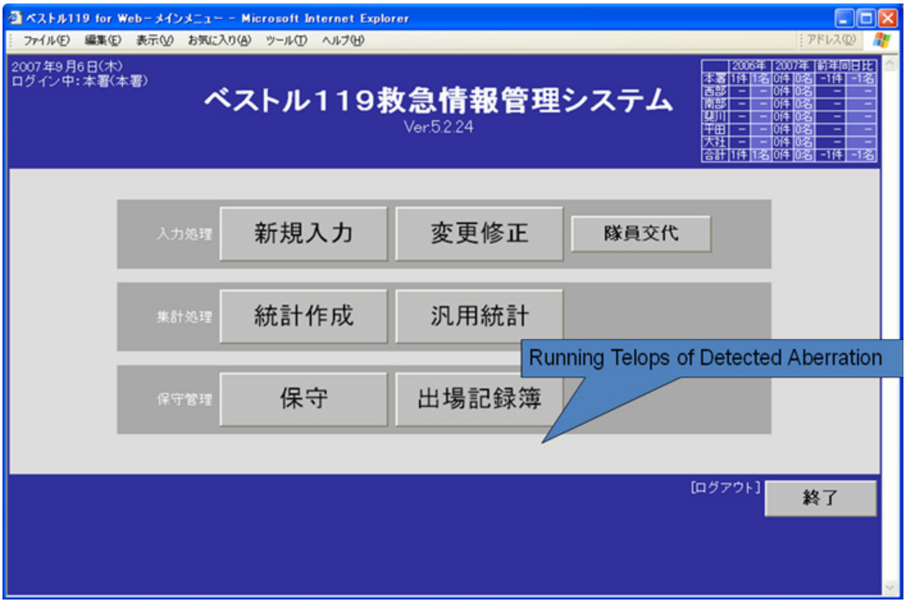
3.2. Syndromic Surveillance for Ambulance Transfer

In Japan, there are about 800 fire department headquarters and almost all of them record their activity or information about patients, but very few of them record patients' symptoms. We have studied Tokyo Fire Department's role in syndromic surveillance since 2005. Based on our findings, the Tokyo Local Government will adopt this system as their own policy as a countermeasure to a bioterrorism attack, outbreak of infectious diseases, or other health risk incidents, and will complete construction and begin operating in 2009. However, cities other than Tokyo are relatively small, and thus they cannot create their own systems themselves. Therefore, we have been constructing the reporting system with syndromic surveillance since 2007, collaborating with WAKO SHOUJI Co. Ltd. which develops and sells Bestoru, a reporting system for ambulance transfer at small fire departments. Bestoru has two types of systems: one a standalone system with the server in the headquarters of the fire department, and the other type is ASP.

Recorded symptoms include fever, difficulty in breathing, diarrhea, vomiting, and convulsion. Several symptoms can be selected simultaneously and it analyzes every hour for 24 h, thus this system has high timeliness in comparison with other syndromic surveillance systems in Japan. This system is fully automated and there is no burden to ambulance teams except for their routine tasks. The aberration detection algorithm is the same as the syndromic surveillance for prescriptions as mentioned above.

Figure 20-5 shows the top page of Bestoru, and running telop inform the ambulance transfer staff if some aberrations in the last 24 h were detected. Of course, it also provides such information to public health centers or local governments through the home page as in Figure 20-6 which is very similar to Figure 20-1, but classification is changed to symptoms instead of types of drugs. It also provides some graphs and aberration detection as in Figure 20-7.

When the Hokkaido Toyako G8 summit meeting was held, this syndromic surveillance system had operated routinely at the Nishiiburi Fire Department which covers the Toyako area where the meeting was held. Moreover, the simpler version had been operating at the delegation team location for a week when the meeting was held. Other fire departments around the summit meeting also cooperated in reporting the number of transfers by symptom for 1 month.


ベストル119 for Web－メインメニュー － Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D)

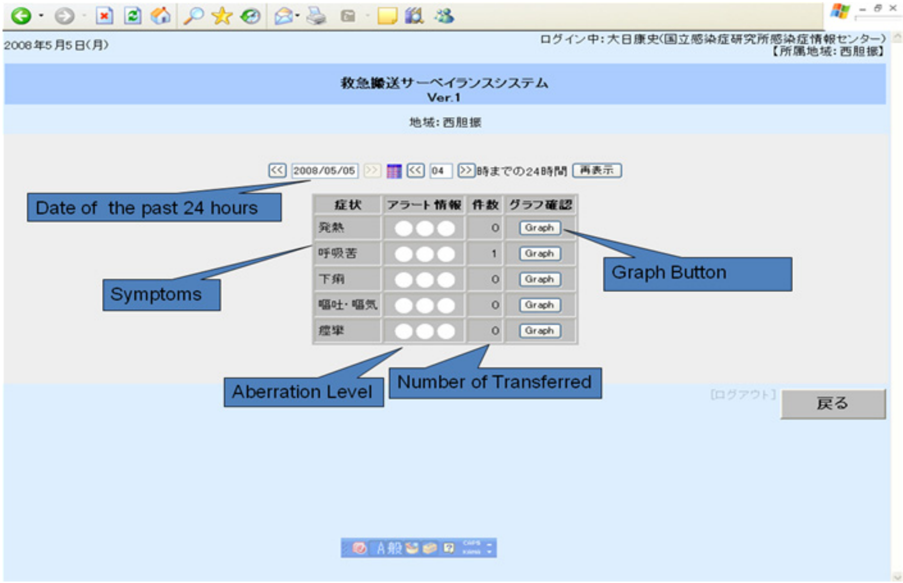
2007年9月6日(木)
ログイン中：本署(本署)

ベストル119救急情報管理システム
Ver.5.2.24

	2006年		2007年		前年同日比	
本署	1件	1名	0件	0名	-1件	-1名
西部	－	－	0件	0名	－	－
南部	－	－	0件	0名	－	－
奥川	－	－	0件	0名	－	－
平田	－	－	0件	0名	－	－
大社	－	－	0件	0名	－	－
合計	1件	1名	0件	0名	-1件	-1名

入力処理
新規入力
変更修正
隊員交代
集計処理
統計作成
汎用統計
保守管理
保守
出場記録簿
Running Telops of Detected Aberration
[ログアウト]
終了

Figure 20-5. Top page of Bestoru.


2008年5月5日(月)

ログイン中：大日康史(国立感染症研究所感染症情報センター)
【所属地域：西胆振】

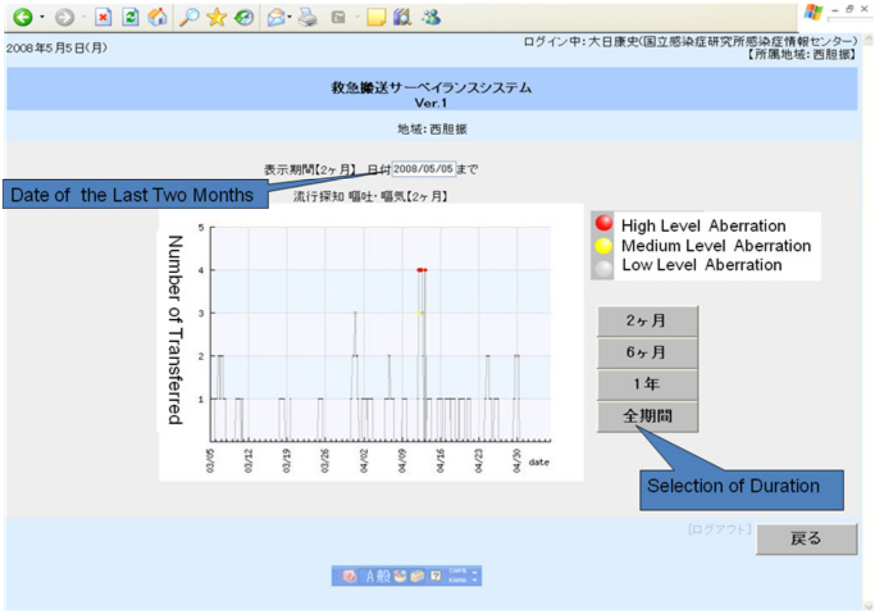
救急搬送サーベイランスシステム
Ver.1

地域：西胆振

<< 2008/05/05 >> << 04 >>時までの24時間 再表示

症状	アラート情報	件数	グラフ確認
発熱	○○○	0	Graph
呼吸苦	○○○	1	Graph
下痢	○○○	0	Graph
嘔吐・嘔気	○○○	0	Graph
痙攣	○○○	0	Graph

Figure 20-6. Feedback HP for local government or public health center.



2008年5月5日(月)

ログイン中：大日康史(国立感染症研究所感染症情報センター)
【所属地域：西胆振】

救急搬送サーベイランスシステム
Ver.1

地域：西胆振

表示期間【2ヶ月】 日付 2008/05/05 まで

流行探知 嘔吐・嘔気【2ヶ月】

Date of the Last Two Months

Selection of Duration

Figure 20-7. Number of transferred due to vomiting for the last 2 months.

Currently, Bestoru with syndromic surveillance has been operating routinely at two other departments besides Nishiiburi. The number of operating departments increased to 20 in 2007. However the share of Bestoru is just 2.5%. It is important that syndromic surveillance is adjusted to systems in addition to Bestoru, as well as Bestoru increasing its share so as to cover a wider area in Japan.

3.3. Syndromic Surveillance for OTC Drug Sales

We can buy OTC sales data from private research firms in Japan, though it is provided without any charge from national chain pharmacies in the U.S. Thus we cannot operate syndromic surveillance using OTC sales routinely; it must be used in surveillance for high profile events such as G8 summit meetings or the World Cup. However, its sensitivity and timeliness, especially in the case of OTC drugs for the common cold, have been confirmed in Japan [3].

For the Hokkaido Toyako G8 summit meeting, we monitored OTC sales of common cold drugs, drugs for relief of fever and pain, antidiarrhea drugs, drugs for eyes, and drugs for skin. In total, 71 pharmacies around the meeting location joined for 1 month, i.e., from 2 weeks before the meeting until 2 weeks after the meeting. However, so as to obtain more credible detection of aberrations, we used the past data of each pharmacy for 2 years. The aberration detection algorithm is the same as the syndromic surveillance for prescriptions and ambulance transfer as mentioned above.

Unfortunately, data transfer from pharmacies is not automatic, so the system requires manual input, which costs time and money. Usually, data transfer is completed by 2 o'clock p.m. After that it is analyzed and the results are shared by 5 o'clock in the evening. Therefore this system lacks timeliness in comparison with the two other syndromic surveillance systems.

3.4. Joint Conference for Evaluation of Aberration Signals from the Syndromic Surveillance System

Though aberration signals are reported from statistical analysis, we have to check manually and decide whether we require local public health centers to investigate so as to obtain more detailed information. Thus we held a joint meeting with the Hokkaido local government, local public health center, Hokkaido local public laboratory, National Institute of Infectious Diseases, and the Ministry of Health, Labor and Welfare everyday even on the weekend.

4. RESULTS

The G8 summit meeting closed without any health threat event or emergency. We used the system for 1 month. However, the syndromic surveillance system found some aberrations. Table 20-1 summarizes the results of aberration detection by surveillance. The surveillance for ambulance transfer found 40 aberrations, among which we found the high aberration seven times on the second, third, fourth, seventh, eighth, tenth and 17th of July. Since the timing is just before, during, or just after the meeting, and the location of aberration detection was very close to the meeting place, the local public health center investigated and collected information about characteristics of patients and their symptoms. Such information was reported to Hokkaido local government and the National Institute of Infectious Diseases; however, their evaluations determined that there was no public health threat. Two weeks later, the official disease surveillance reported an increase in the number of patients with Herpangina, and thus we conjecture that those aberrations were due to severe Herpangina cases.

Table 20-1. Number of aberration detections and their level by surveillance.

	Ambulance Transfer	**Prescriptions**	**OTC**
Low	23	8	1
Middle	10	0	0
High	7	0	0

As mentioned before, the system was semi-automated, with part being automated and another part operated manually. Though all participants understood the importance of cooperating in this syndromic surveillance for the G8 summit meeting, it was difficult to maintain the focus needed for manual operation even for just 1 month. For example, Figure 20-8 shows the reporting rate for prescriptions from pharmacies. A sharp trough means Sunday, when most pharmacies are closed. About 30% of pharmacies reported automatically, so the minimum reporting rate is higher than 30%. It indicates clearly that the reporting rate declined gradually for 1 month. Other than manual data input, there were also some mistakes in manual calculations, and the system was down once. Though the system downtime was scheduled in advance, it affected the syndromic surveillance system heavily and the system showed its vulnerability. It is critical to have a fully automated syndromic surveillance approach with fault-tolerance measures.

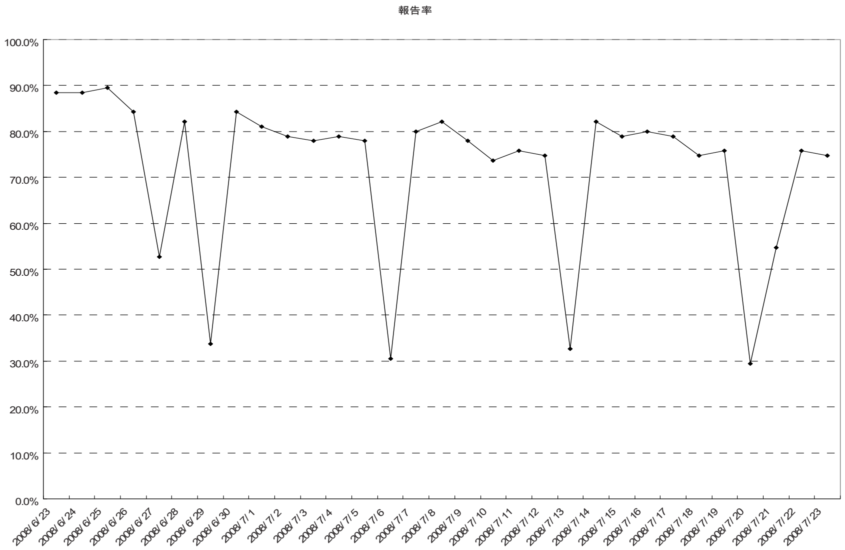


Figure 20-8. Reporting rate in the surveillance for prescriptions.

5. CONCLUSIONS AND DISCUSSION

The value of syndromic surveillance is obvious during large events. Routine syndromic surveillance practice is important to detect and prepare responses to bioterror attacks and emerging diseases such as the A/H1N1 influenza. After the G8 summit meeting, we started to construct a routine syndromic surveillance system based on our experience with syndromic surveillance during the G8 summit meeting. In the three surveillance systems presented, we constructed the system of prescriptions for practical use covering all of Japan until the end of April 2009. In total, about 2,100 corporate pharmacies, 5% of all pharmacies in Japan, are included. Because of a novel influenza outbreak, the system will expand to include 3,000 pharmacies, about 7% of all pharmacies, until the end of June 2009. Other research [4] showed that a system with 1,200 pharmacies, if distributed uniformly in Japan, can function almost equivalently with the current sentinel surveillance for (seasonal) influenza. However, the corporate pharmacies do not distribute uniformly. In Figure 20-9 a smaller number of the population covered by one corporate pharmacy is marked with a dark color, whereas white-color regions show prefectures where we cannot operate this surveillance method. This figure shows there are some differences in covered population per corporate pharmacy, and the system has not yet been operated in two prefectures. The next step is to start this system in the two prefectures. Moreover, the purpose of the syndromic surveillance is not to

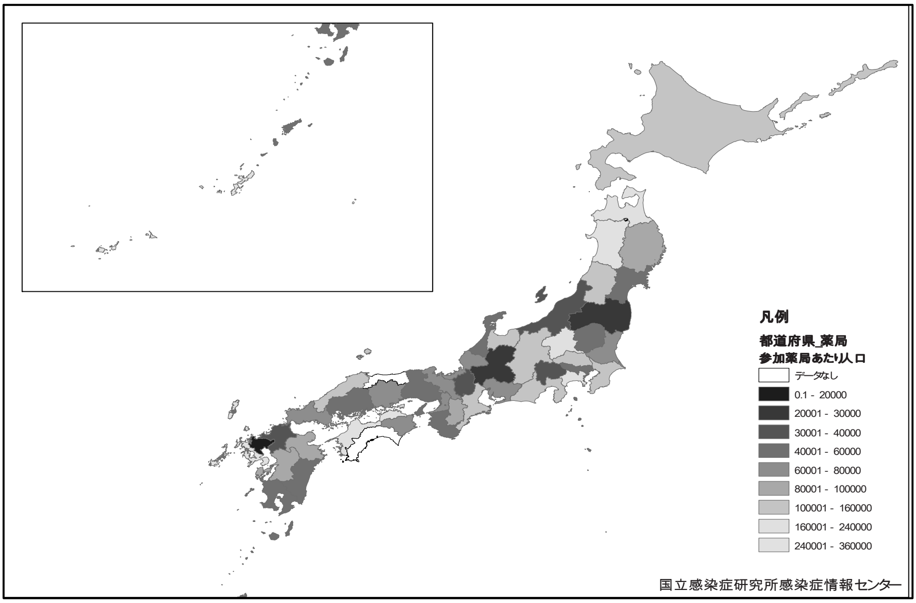


Figure 20-9. Population per corporate pharmacy.

trace the trend of the number of patients, but early detection. In order to maximize early detection, more pharmacies must join the effort. We believe it will be important to extend the system with more pharmacies before the second wave of this A/H1N1 pandemic develops this winter.

6. OTHER SYNDROMIC SURVEILLANCE SYSTEMS AT THE EXPERIMENTAL LEVEL IN JAPAN

Finally, we are also developing and constructing other syndromic surveillance systems as experiments, other than the systems mentioned above. These remain at an experimental level, however, and will not operate practically in the near future. The systems summarized in this section include syndromic surveillance from EMRs, orders for medical examinations, absenteeism at school, and nosocomial outbreaks; we describe each system briefly.

6.1. Syndromic Surveillance from EMRs

Syndromic surveillance from EMRs is the most widely adopted syndromic surveillance used in the U.S. and Taiwan. However, as mentioned before, low penetration and no standardization in EMRs and high privacy concerns are a barrier to developing syndromic surveillance from EMRs in Japan. Nonetheless, we are developing syndromic surveillance from EMRs for outpatients and then applying it to inpatients so as to detect nosocomial outbreaks as early as possible.

We have completed the development of a syndromic surveillance system for outpatients for three types of EMRs. One is the most widely used EMR at clinics in Japan and it has more than 2,000 users out of 0.1 million clinics in total. However the number of users of this type of EMR with syndromic surveillance is limited to just two clinics. We expect the number of cooperating clinics with this type of EMR will grow rapidly. The second type of EMR which has adopted syndromic surveillance has small prevalence, less than 100. Though only 12 clinics which use this type of EMR cooperate in syndromic surveillance, half of them are located in the same city, Izumo City, Shimane Prefecture, which has a population of about 200,000. The other half of them are covered by the same public health center, which covers a population of 22,000. Thus intensity of this syndromic surveillance is quite high. The third EMR is for one hospital with more than 600 beds as well as for detection of nosocomial outbreak, which it performs by monitoring inpatients. It is located in Izumo City, and has joined the EMR syndromic surveillance network there.

The system was started in 2006 and is presented in Figure 20-10. First, it searches for keywords such as fever or diarrhea in complaints in EMRs in a clinic or hospital. Then it counts the number of patients and sends the information to a web server which is outside the clinic or hospital, everyday at midnight. The web server calculates the proportion of clinics or hospitals which find some alerts of outbreak detection in the community. It also provides this information to the local government and local public health center, as well as to the cooperating clinics and hospital.

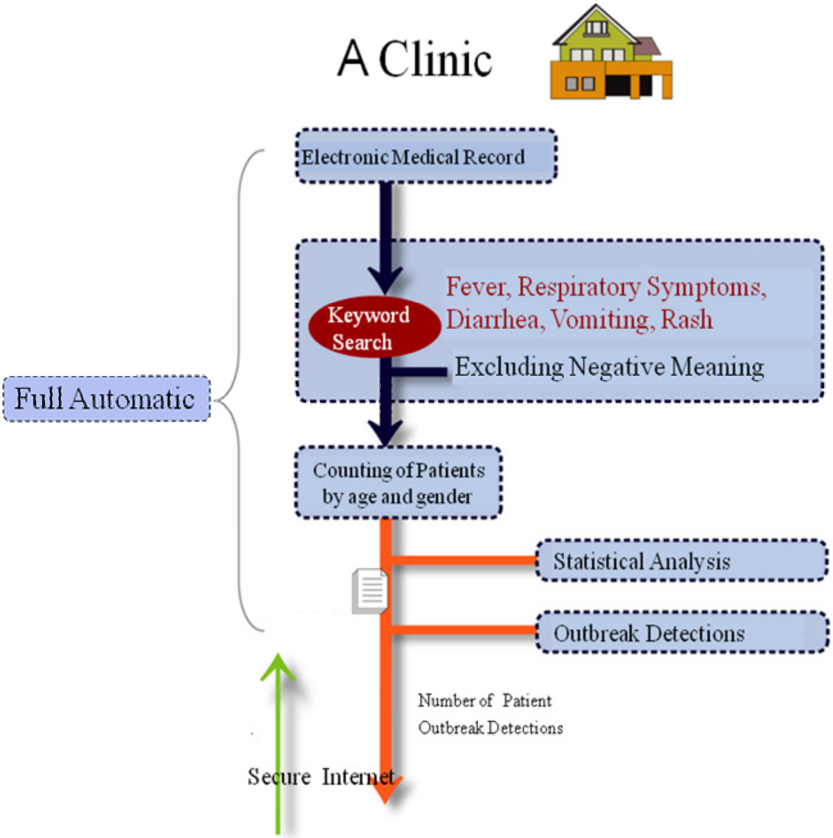
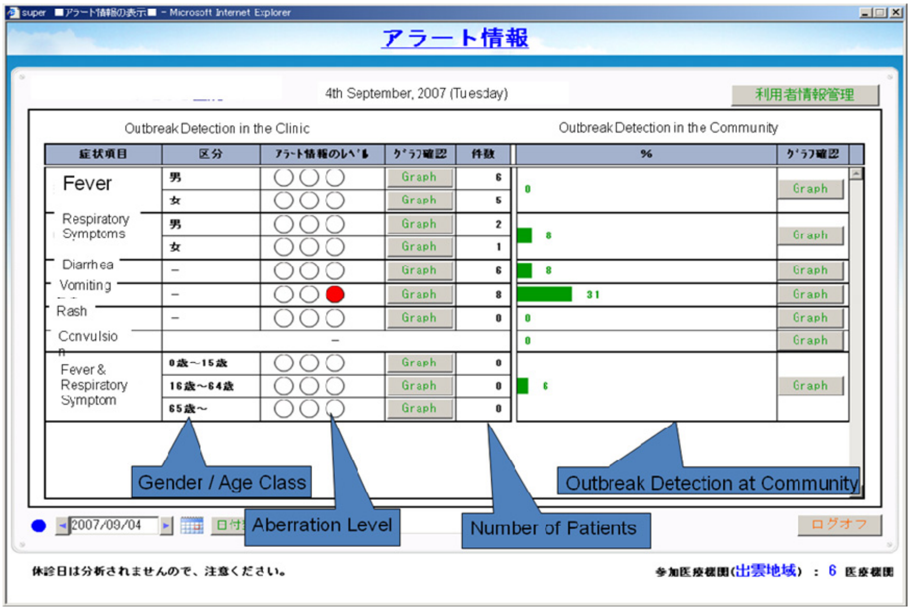

A Clinic
Electronic Medical Record
Keyword Search
Fever, Respiratory Symptoms, Diarrhea, Vomiting, Rash
Excluding Negative Meaning
Full Automatic
Counting of Patients by age and gender
Statistical Analysis
Outbreak Detections
Number of Patient
Outbreak Detections
Secure Internet

Figure 20-10. System for electronic medical record.

Figure 20-11 shows the home page on the web server. The left hand side shows the situation in each clinic and hospital. The first column shows the symptoms such as fever, respiratory symptoms, diarrhea, vomiting, convulsion, and fever with respiratory symptoms. Some symptoms are classified by gender or age class, under 15, under 65, and older than 65. The second column shows alerts in each symptom in this clinic or hospital at level three. The red circle means the highest level of alert was found. The buttons labeled "graph" show the number of patients and alerts. The last column on the left hand side shows the number of patients with each symptom in this clinic or hospital.


super ■アラート情報の表示■ - Microsoft Internet Explorer

アラート情報

4th September, 2007 (Tuesday)

利用者情報管理

Outbreak Detection in the Clinic | Outbreak Detection in the Community

症状項目	区分	ｱﾗｰﾄ情報のﾚﾍﾞﾙ	ｸﾞﾗﾌ確認	件数	%	ｸﾞﾗﾌ確認
Fever	男	○○○	Graph	6	0	Graph
	女	○○○	Graph	5		
Respiratory Symptoms	男	○○○	Graph	2	8	Graph
	女	○○○	Graph	1		
Diarrhea	–	○○○	Graph	6	8	Graph
Vomiting	–	○○●	Graph	8	31	Graph
Rash	–	○○○	Graph	0	0	Graph
Ccnvulsion	–				0	Graph
Fever & Respiratory Symptom	0歳～15歳	○○○	Graph	0	6	Graph
	16歳～64歳	○○○	Graph	0		
	65歳～	○○○	Graph	0		

● ◂ 2007/09/04 ▸ 日付 ログオフ

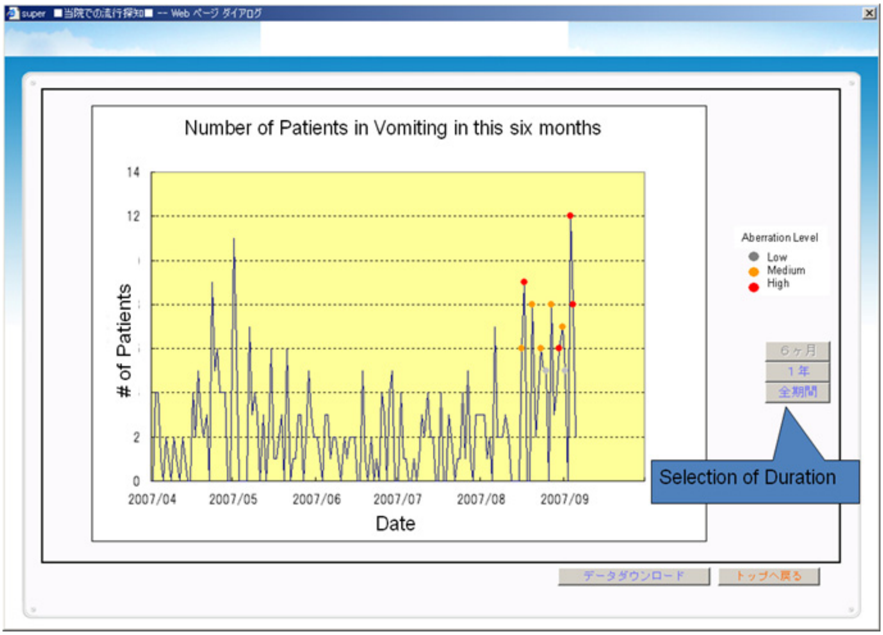
休診日は分析されませんので、注意ください。 参加医療機関(出雲地域) : 6 医療機関

Figure 20-11. Feedback HP for each corporate clinic or hospital.

On the right hand side, it shows the proportion of clinics or hospitals which found alerts in each symptom. If all clinics and hospitals which are cooperating in this project find alerts in a symptom in the same day, this green bar becomes longest and its score is 100. The buttons labeled "Graph" also show the number of outbreaks detected in the community.

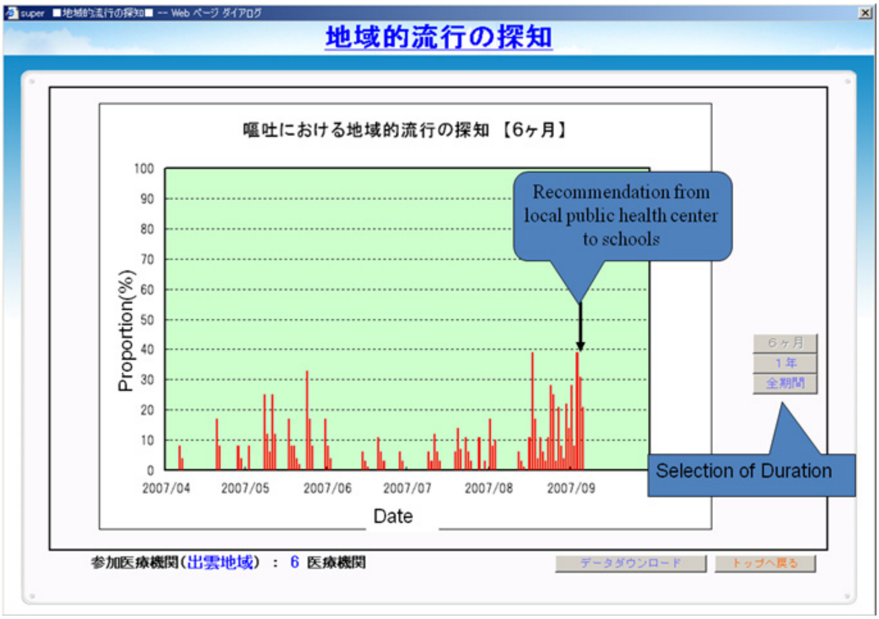
Figure 20-12 is provided by clicking the graph button on the left hand side. It shows the number of patients with each symptom in this clinic or hospital for 6 months. Red or yellow circles indicate high and medium level alerts in vomiting recently. The graph shown in Figure 20-13 is brought up by clicking the button on the right hand side; it shows the proportion of clinics or hospitals which found alerts in the last 6 months. The outbreak of vomiting seems to be significant even at the community level and the local public health center recommended intensified hand-washing to schools on September 4th.

If the system finds the highest alert in a clinic or hospital, it sends an e-mail automatically to the clinic or hospital telling them to check the home page. Moreover, if the system receives the highest alert from multiple clinics and hospitals, it sends an e-mail to all clinics and hospitals as well as the local government and local public health center.



Selection of Duration

Figure 20-12. Number of patients due to vomiting and aberration detection at each clinic or hospital.



Recommendation from local public health center to schools

Selection of Duration

Figure 20-13. Outbreak detection in the community.

This system confirmed the Norovirus outbreak in 2007 and detected the late influenza outbreak in the middle of March. Moreover, it detected the outbreak of meningitis in September. A doctor used the information about the early detection of influenza to announce to patients and neighborhood pharmacies to cancel leave plans and to purchase masks or rapid test kits.

6.2. Syndromic Surveillance from Orders for Medical Examinations

Even though EMRs have low penetration, claim data is gradually changing over to electronic format because public health insurance will limit the acceptance of claims from the web only starting in 2011. The Japan Medical Association has developed their own original claim data system, ORCA, and distributed it freely. ORCA will be used to record all treatment or medical examinations which are performed and which require patients to pay out-of-pocket. Thus, if we can draw some information from ORCA, especially for ordered or performed medical examinations, we can guess patients' symptom even though ORCA does not record the results of such examinations. For example, the number of rapid tests for influenza will reflect the number of patients with influenza-like-illness, but not confirmed influenza.

This system is currently under construction. However, ORCA users number more than 5,000 and it will increase to more than 10,000, which is 10% of all clinics in Japan, by 2011. Therefore syndromic surveillance with ORCA also will cover a substantial number of clinics. Moreover, the Japan Medical Association already declared the use of ORCA for monitoring for outbreak of infectious diseases. Therefore, we hope that the Japan Medical Association will perform syndromic surveillance for medical examinations using ORCA.

6.3. Syndromic Surveillance from Absenteeism at School

School absenteeism is one source of syndromic surveillance in the U.S. [5] or Taiwan. Moreover, since school closure may mitigate an influenza epidemic [5], the monitoring of school absenteeism is important for control not only of seasonal influenza, but also pandemic flu. Actually, elementary and junior high schools in Japan sometimes close in the winter when the rate of absenteeism is higher than a certain level, about 30%. However such a criterion seems to be very high to control influenza activity. Therefore, we focus on school absenteeism below the closure level, so as to detect influenza or other infectious disease outbreaks as early as possible.

Unfortunately, school absenteeism is not recorded electronically to date. Thus, as a start, we are developing an absenteeism recording system on the web. It records the number of absentees by class and by symptoms, which are fever, headache, diarrhea or stomachache, vomiting or nausea, influenza, and others. Then we apply C1 of EARS to detect aberrations. The aberration information is shared with the education committee, local public health center, and doctors surrounding the school.

Since September 2008, we are performing the number of cooperating schools to 20 elementary and junior high schools. We hope that the system can detect outbreaks early and trace their geographical spread among schools.

6.4. Syndromic Surveillance for Nosocomial Outbreak

As mentioned before, the syndromic surveillance for EMRs is also applied to detect nosocomial outbreaks [6]. Beginning in March 2007, two other university hospitals adopted this system.

Moreover, we applied a similar syndromic surveillance system to long term care facilities for the elderly. The elderly population is increasing in Japan, and about 4% of the elderly live in long term care facilities. Sometimes, such facilities are attacked by influenza or Norovirus outbreaks and this becomes a public concern. However, long term care facilities do not have EMR systems and thus we are developing a recording system on the web initially, the same as for school absenteeism. We will then apply EARS C1 to detect aberrations. Currently, six facilities are cooperating in this experiment. We hope to extend the number of cooperating facilities and exchange aberration information among close facilities and local public health centers.

ACKNOWLEDGEMENT

We acknowledge all participants of the syndromic surveillance for the G8 summit meeting.

QUESTIONS FOR DISCUSSION

1. How should we design a syndromic surveillance system for high profile events such as Olympic Games or G8 summit meetings?
2. How should we design a routine syndromic surveillance system?
3. What is common and different in these surveillance designs?

REFERENCES

1. Dafni UG, et al. Algorithm for statistical detection of peaks – Syndromic Surveillance System for the Athens 2004 Olympic games. Morbidity and Mortality Weekly Report 2004; 53(Suppl.): 86–94.
2. Jorm LR, et al. Watching the games: public health surveillance for the Sydney 2000 Olympic games. Journal of Epidemiology and Community Health 2003; 57: 102–108.
3. Osaka K, Takahashi H, Ohyama T. Testing a symptom-based surveillance system at high-profile gatherings as a preparatory measure for bioterrorism. Epidemiology and Infection 2002; 129: 429–434.
4. Yoshida M, et al. Seasonal influenza surveillance using prescription data for anti-influenza medications. Japanese Journal of Infectious Diseases 2009; 62: 233–235.
5. Cauchemez S, Valleron AJ, Boelle PY, Flahault A, Ferguson NM. Estimating the impact of school closure on influenza transmission from Sentinel data. Nature 2008; 452(7188): 750–754.
6. Kikuchi K, Ohkusa Y, et al. Syndromic surveillance for early detection of nosocomial outbreaks, in Daniel Zeng et al., eds. Intelligence and Security Informatics: Biosurveillance, Springer, New York, 2007.

SUGGESTED READING

Wagener MM, et al. Syndrome and outbreak detection using chief-complaint data – experience of the real-time outbreak and disease surveillance project. MMWR Morbidity and Mortality Weekly Report 2004; 53(Suppl.): 25–27 .

Henning KJ. What is syndromic surveillance? MMWR Morbidity and Mortality Weekly Report 2004; 53(Suppl.): 7–11.

Siegist DW and Tennyson SL. Technologically-Based Biodefense, Potomac Institute for Policy Studies, Virginia, 2003.

Buehler JW, Berkelman RL, Hartley DM, Peters CJ. Syndromic surveillance and bioterrorism-related epidemics. Emerging Infectious Diseases 2003; 9: 1197–1204.

INDEX

D. Zeng et al. (eds.), *Infectious Disease Informatics and Biosurveillance*, Integrated Series in Information Systems 27, DOI 10.1007/978-1-4419-6892-0,