Success in Academic Surgery
Series Editors: Lillian Kao · Herbert Chen

Justin B. Dimick
Carrie C. Lubitz *Editors*

Health Services Research

Second Edition

Springer

Success in Academic Surgery

Series Editors

Lillian Kao
McGovern School
The University of Texas Health Science Centre
Houston, TX
USA

Herbert Chen
Department of Surgery
University of Alabama at Birmingham
Birmingham, AL
USA

More information about this series at http://www.springer.com/series/11216

Justin B. Dimick • Carrie C. Lubitz
Editors

Health Services Research

Second Edition

Springer

Editors
Justin B. Dimick
Department of Surgery
University of Michigan Health System
Ann Arbor, MI
USA

Carrie C. Lubitz
Department of Surgery
Massachusetts General Hospital
Boston, MA
USA

ISSN 2194-7481 ISSN 2194-749X (electronic)
Success in Academic Surgery
ISBN 978-3-030-28356-8 ISBN 978-3-030-28357-5 (eBook)
https://doi.org/10.1007/978-3-030-28357-5

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I Main Research Areas

Part II Emerging Areas of Research

Part III Tools of the Trade

Part IV Career Development

Part I

Main Research Areas

1 An Introduction to Health Services Research

Justin B. Dimick and Carrie C. Lubitz

Abstract

The scientific focus of academic surgery has changed dramatically over the past decade. Historically, surgeon-scientists engaged almost exclusively in basic science research. With the rise of health services and outcomes research, more trainees and junior faculty are pursuing research in these disciplines. Despite the increasing popularity of this field, there are very few resources for young surgeons interested in learning about these disciplines as applied to surgery.

We developed this book, and have revised our contents in this second edition, to help fill this gap. We start with a description of the main research areas in health services research followed by a look ahead into emerging areas of investigation. We then include several chapters that introduce the tools necessary to conduct this type of research. The final chapters provide practical advice on career development and program building for surgeon-scientists interested in pursuing this area of scholarly work.

1.1 What Is Health Services Research?

We often get asked how health services research is different from traditional "clinical research". Studying the end results of surgical care is clearly not new. As long as surgeons have been operating, we have been studying our patient's outcomes. Although there is clearly overlap between traditional clinically-focused scientific

J. B. Dimick (✉)
Department of Surgery, University of Michigan Health System, Ann Arbor, MI, USA
e-mail: jdimick@umich.edu

C. C. Lubitz
Department of Surgery, Massachusetts General Hospital, Boston, MA, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_1

work, health services research often takes different perspectives and looks at health care through a much broader lens. To illustrate these differences in perspective, it is useful to consider two popular definitions of health services research (HSR). AcademyHealth, the leading professional organization for health services researchers (their Annual Research Meeting is a great meeting to attend by the way), defines HSR as follows:

> AcademyHealth defines health services research as the multidisciplinary field of scientific investigation that studies how social factors, financing systems, organizational structures and processes, health technologies, and personal behaviors affect access to health care, the quality and cost of health care, and ultimately our health and well-being. Its research domains are individuals, families, organizations, institutions, communities, and populations.

The Agency for Healthcare Research and Quality (AHRQ), one of the leading funding agencies for HSR, uses the following definition:

> Health services research examines how people get access to health care, how much care costs, and what happens to patients as a result of this care. The main goals of health services research are to identify the most effective ways to organize, manage, finance, and deliver high quality care; reduce medical errors; and improve patient safety.

1.2 What Is Outcomes Research?

Outcomes research is sometimes used interchangeably with health services research but is probably best considered one chief discipline within HSR. The Agency for Healthcare Research and Quality (AHRQ) defines outcomes research as follows:

> Outcomes research seeks to understand the end results of particular health care practices and interventions. End results include effects that people experience and care about, such as change in the ability to function. In particular, for individuals with chronic conditions—where cure is not always possible—end results include quality of life as well as mortality. By linking the care people get to the outcomes they experience, outcomes research has become the key to developing better ways to monitor and improve the quality of care.

While this formal definition of outcomes research is not as broad as the definition of health services research above, it is still different from traditional clinical research in a few important ways. Most importantly, there is a focus on a broader set of outcomes beyond clinical endpoints (e.g., mortality and morbidity), including quality of life and patient-centered outcomes. With the increasing popularity of Patient Centered Outcomes Research (PCOR), and the creation and funding of the Patient Centered Outcomes Research Institute (PCORI), researchers who primarily focus on this area would probably label themselves as “outcomes researchers”. Whereas investigators who focus on health care policy evaluation may refer to themselves as “health services researchers”. However, for the purposes of this overview, we view the two as comprising a single area of scientific endeavor which we will refer to as HSR.

Table 1.1 Key differences between health services research and traditional clinical research

The questions	HSR asks broader questions. Rather than assessing clinical treatments, HSR questions often address the organization, delivery, financing, and regulation of the health care system
The setting	HSR studies health care in “real world” settings as opposed to the carefully constructed environment of a clinical trial. This difference is often described as studying “effectiveness” (real world) vs. “efficacy” (randomized clinical trial)
The outcomes	HSR often uses different end-points. Rather than focusing on clinical endpoints (morbidity and mortality), HSR often uses patient centered outcomes, such as quality of life and symptom bother
The data	Rather than directly collecting data from the medical record, HSR often uses large datasets to conduct observational research; or, at the other end of the spectrum, surveys or interviews with patients are used to gather very detailed information
The tools	The research tools necessary to perform sophisticated HSR vary with the nature of the question and span from large database analysis and econometrics to qualitative research and psychometrics

As evident in these definitions above, there are several key distinctions between HSR and traditional clinical research, including important differences in the questions, the settings, the outcomes, the data, and the tools (Table 1.1).

1.3 Part I. Main Research Areas

The book begins with an introduction to the main research themes that investigators are currently pursuing. Drs. Massarweh and Chang (Chap. 2) provide an overview of comparative effectiveness research and describes how this field goes beyond randomized clinical trials. They describe the spectrum of study designs (e.g., pragmatic trials, observational studies) available for assessing which treatments are most effective, and how effectiveness may vary across different patient and provider subgroups and varying perspectives.

As described above, HSR often asks questions much more broadly than traditional clinical research, including investigating differences in practice style and treatment across large areas and understanding how these are shaped by healthcare policy. Drs. Columbo Goodney (Chap. 3) provide an overview of the seminal work done on variations across geographic areas by the Dartmouth Atlas group that opened our eyes to wide, unwarranted practice variations in the United States. Drs. Chaabra and Dimick (Chap. 4) then discuss the importance of taking a broad perspective in evaluating how health care policy research can help improve the context in which we work by critically evaluating the incentives and structure that are largely invisible, but shape our daily work. Dr. Chang (Chap. 5) then considers the inequities in our health care system that lead to disparities in use and outcomes of surgery. He emphasizes that the field of disparities research needs to move beyond documenting quality gaps, and, instead, begin identifying strategies to fix them.

Another important focus of HSR is on measuring and improving quality. Drs. Merkow and Bilimoria, experts with the American College of Surgeons National

Surgical Quality Improvement Program (ACS-NSQIP), describe the field of quality measurement, including the pros and cons of the structure, process, and outcomes that are used as performance indicators (Chap. 6). Drs. Reinke and Kelz, who serve as leaders locally within their institutions and nationally, then discuss how these quality metrics can be leveraged to improve patient-care locally (Chap. 7).

1.4 Part II. Emerging Areas of Research

We next consider several emerging areas within HSR that are likely to become integral to our field within the next 10 years. Many of these emerging areas will bring forth new knowledge that has translational value for our organizations, and others, to move the quality, safety and appropriateness mission forward. While many of these fields are well-established outside surgery, only a small number of investigators are pursuing these within our profession, creating a large opportunity for young surgeon-scientists.

Drs. Dossett and Telem (Chap. 8) describe the field of implementation and dissemination research. Implementation science explicitly recognizes the gap in translating evidence into practice, providing a rich set of theoretical frameworks and research tools to rigorously study barriers and facilitators of the adoption of evidence in real-world settings. Building on this knowledge of the importance of "context" in optimizing healthcare, Dr. Haynes (Chap. 9) examines the important role organizational culture plays in creating well-functioning environments that are safe and favorable to successful adoption of best practices.

There is also a growing emphasis on assessing outcomes from the patient perspective. Traditional clinical endpoints are clearly important, but there is often divergence between clinician and patient perspectives. For example, after inguinal hernia repair, surgeons usually measure the recurrence rate, which is quite rare. Patients, however, are much more bothered by chronic inguinal pain, which has a much higher rate than recurrence. Drs. Temple and Morris (Chap. 10) provide an overview of Patient-Reported Outcome measures, which are an increasingly important part of health services research. Funds are available for CER focused on the patient perspective through PCORI and other institutes, it is an exciting area within HSR for young surgeons to get involved in.

Perhaps the most cutting edge research in our field is aiming to get inside the "black box" of what happens in the operating room. Most existing quality improvement work focuses on optimizing perioperative care (e.g., antibiotics for prevention of SSI) and completely ignores how the operation itself is conducted. Drs. Hu and Greenberg (Chap. 11) describe multidisciplinary efforts to understand and improve the performance of systems, teams and individuals in the operating room environment. New to this second edition, Drs. Yule and Smink describe the leading framework for assessing the capability of surgeons to create an optimal environment among the OR team, the Non-Technical Skills for Surgeons (NOTSS). This instrument, which assesses the following domains: situation awareness, decision making, communication/teamwork, and leadership, has emerged as a powerful practical tool for research and clinical improvement.

Once we have creative solutions for improving quality and performance it is essential to have an infrastructure to disseminate and test them in the community. Dr. Finks, who was integral to the success of the statewide bariatric surgery collaborative in Michigan, describes the power of using regional collaboratives as a potential laboratory for evaluating these interventions. He describes the power of such collaboratives for implementing best practices across large geographic areas and in diverse practice settings.

Finally, we have added a new chapter in this edition by Dr. Schneider (Chap. 14) to cover perhaps one of the newest areas of inquiry for surgeon-scientists: Machine learning and artificial intelligience. Machine learning has recently arrived as the next generation of statistical modeling and, in some use cases, has better predictive ability than traditional regression approaches. Artificial intelligience is the science of creating intelligent machines that have the ability to achieve goals like humans via a combination of algorithms combined with predictive models. As the evidence supporting high-value use cases grow (e.g., reading EKGs, diagnostic imaging), there are no doubt numerous areas in medicine—and surgery—where artificial intelligience applications will change how we deliver care in the next few decades.

1.5 Part III. Tools of the Trade

The tools necessary to conduct HSR are diverse and in flux. There is constant innovation in HSR bringing in expertise from additional fields. However, there are certain tools that are trademarks of HSR and we will cover those in this section of the book.

Many young surgeons begin their research careers working with large datasets. These are relatively inexpensive and can help fellows and junior faculty get over "bibliopenia"—a necessary first step towards establishing yourself as an investigator. Drs. Sheetz and Dimick (Chap. 5) provide an overview of large datasets available for conducting health services research. Because we often try to make causal inference from these large datasets, tools are need to address confounding and selection bias. Methods for addressing these problems and thereby enhancing causal inference are central to the HSR toolbox. Drs. Sheetz and Nathan (Chap. 16) introduces commonly used methods, including multivariable regression, propensity score matching, and instrumental variable analysis. None of these methods are perfect, and every surgeon-scientist conducting HSR should be familiar with the strengths and weakness of each of these methods as well as the epidemiological theory underlying various biases.

Dr. Maggard (Chap. 17) describes how individual studies can be brought together and synthesized in a meta-analysis. Besides giving a single summary "best estimate" of available studies, these techniques also allow us to systematically study how the treatment effect varies across patient and provider subgroups (i.e., establish treatment-effect heterogeneity). Drs. Lee and Pitt (Chap. 18) describe the scientific study of understanding how patients make decisions. New in this second edition we have added another chapter by Dr. Poulouse (Chap. 19) on decision analysis techniques and how they can be used to synthesize evidence to explicity model tradeoffs in surgery, and how the methods can be expanded to create estimates of cost-effectiveness.

Large datasets are usually a great starting point for young surgeon-scientists but they lack the detail required to answer many important questions. Dr. Brasel (Chap. 20) discusses the use of survey research to generate primary data about knowledge, attitudes, and beliefs. Dr. Schwarze (Chap. 21) then provides an introduction to qualitative research, a rich field of inquiry that uses focuses groups, interviews, and ethnographic methods to gather information. Qualitative research uses words rather than numbers as data and is an entire area of science unto itself. These methods are absolutely crucial for helping us understand "why" things do and do not work in healthcare. In particular, qualitative methods are central to many areas of research discussed above, especially implementation bibliopenia science.

1.6 Part IV. Career Development

The final section provides practical advice for young surgeons interested in building a career focused in health services research. The first few chapters, new to this second edition, focus on writing and disseminating scientific results. The first chapter in this section (Chap. 22), by Drs. Ibrahim and Ghaferi, provides a template and structure for writing a scientific manuscript. The next chapter (Chap. 23) describes the use of Visual Abstracts for disseminating science through social media and other channels. Written by Dr. Ibrahim, who pioneered the development of Visual Abstracts as the creative director at Annals of Surgery, provides the basic steps of creating these visual elements and also discusses the evidence that their use improves dissemintation. In the final chapter in this section (Chap. 24), also new to this second edition, Dr. Locke discusses how to engage policymakers to translate research into practice. She brings the perspective of a transplant surgeon-scientist who has made many translational policy contributions.

The next section focuses on mentorship and career development for HSR surgeon-scientists. Drs. Fry and Sheetz (Chap. 25) describe the keys to engaging medical students in outcomes research so it is a win/win proposition. Dr. Johnston (Chap. 26) offers advice on finding a mentor to conduct health services research and to guide career development more broadly. Dr. Cauley (Chap. 27), who obtained mentored-research training and her M.P.H. during residency, describes what the goals should be for a resident during their research fellowship.

The book concludes with several chapters on external funding and program building. The first chapter on funding, by Dr. Segev (Chap. 22), focuses on the major sources of funding for conducting HSR and includes tips and tricks for preparing proposals. New to this second edition, Dr. Patzer (Chap. 29) provides high-level strategies for grantsmanship in HSR, covering both mentored (K-type) and independent (R-type) awards. We conclude with and a chapter by Drs. Hawkins and

Regenbogen on what to consider when choosing your first job as an HSR surgeon-scientist (Chap. 30), and then close with a chapter by Drs. Brooke and Finlayson designed for more senior investigators considering the programmatic building of an outcome research center (Chap. 31).

Further Reading

Birkmeyer JD. Outcomes research and surgeons. Surgery. 1998;124:477–83.

Brook RH, Ware JE, Rogers WH, et al. Does free care improve adult's health? Results from a randomized controlled trial. N Engl J Med. 1983;309:1426–34.

Cabana MD, Rand CS, Pose NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282(15):1458–65.

Donabedian A. Evaluating the quality of medical care. Milbank Meml Fund Q. 1966;44:166–203.

Lohr KN, Steinwachs DM. Health services research: an evolving definition of the field. Health Serv Res. 2002;37:7–9.

Wennberg JE, Gittelsohn A. Small area variations in health care delivery. Science. 1973;192:1102–8.

Comparative Effectiveness Research 2

Nader N. Massarweh and George J. Chang

Abstract

Comparative effectiveness research (CER) involves human subjects or the use of secondary data to compare the effectiveness of one preventive, diagnostic, therapeutic, or care delivery modality to another using real-world approaches and under real-world conditions. CER utilizes a variety of data sources and methods to generate pragmatic data that can be used to improve health outcomes and value for health care systems by developing and disseminating evidence-based information for patients, clinicians, and other decision-makers. The intent of this type of research is to help stakeholders better delineate which type of treatments are best for which specific patients and under what conditions.

2.1 Introduction

New research findings are used to drive practice, to better understand the role for novel technologies and therapeutics in the care of patients, and to provide health care practitioners with information at the point of care regarding the management of

N. N. Massarweh (✉)
Center for Innovations in Quality, Effectiveness and Safety, Michael E DeBakey VA Medical Center, Houston, TX, USA

Michael E DeBakey Department of Surgery, Baylor College of Medicine, Houston, TX, USA
e-mail: massarwe@bcm.edu

G. J. Chang
Department of Surgical Oncology, University of Texas, MD Anderson Cancer Center, Houston, TX, USA

Department of Health Services Research, University of Texas, MD Anderson Cancer Center, Houston, TX, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_2

patients with rare or uncommon conditions. While the US healthcare system supports a robust research enterprise, an important shortcoming of many contemporaneously published studies is that they don't address a fundamental question relevant to patients and providers—what is the best treatment for this specific patient in this specific clinical context? Given an ever-increasing emphasis on shared decision-making and value in US healthcare, the importance and timeliness of Comparative Effectiveness Research (CER), which is intended to address this specific question, cannot be understated.

Fundamentally, clinical research is intended to compare the safety, benefits, and/or limitations of two or more treatments. In this regard, one might assume that any study in which two interventions are compared is CER. However, CER is specifically intended to provide data comparing the effectiveness of two interventions when applied under *real-world conditions*. Furthermore, CER encompasses research derived from different data sources, and utilizes a variety of study designs and analytic methodologies. These varied sources are assimilated to ascertain which specific intervention(s) will work best for which specific patient (or population of patients) while at the same time balancing the relative benefits and harms in order to inform patient care or health policy decisions. This is one of the main reasons for the proliferation of CER over the past decade, why new opportunities for funding CER have emerged, and why there is an ongoing need for CER to inform contemporary health care reform efforts and the transition from volume-based to value-based care models.

2.2 Efficacy vs Effectiveness

2.2.1 Efficacy

Randomized clinical trials (RCTs) remain the benchmark for data used to either change clinical practice or to drive evidence-based care. However, most RCTs typically address a very specific question when comparing interventions—what is the efficacy of one intervention over another? Simply defined, *efficacy* is a measure of the effect of a particular intervention under idealized circumstances. In addition to the cost and time needed for study initiation and completion (which are acknowledged limitations of RCTs), the external validity (i.e.: generalizability) of findings from RCTs frequently create important challenges to the integration of data into practice. More specifically, because the patient populations included in RCTs are typically strictly defined with numerous inclusion and exclusion criteria (which may not reflect the population of patients providers see in actual practice) and because of the intensity of the care enrolled patients receive (which may not reflect the level of care or the type of practice where a patient is being treated), it is not clear that the findings from a given RCT will directly translate into real-world conditions.

2.2.2 Effectiveness

By comparison, *effectiveness* is defined as the effect of an intervention under real-world conditions and includes an evaluation of not only the benefits, but also the harms. CER is important because what may demonstrate efficacy in the strictly controlled context of a clinical trial may not yield the same outcomes in everyday practice. In many respects, this type of data is much closer to what health care providers and patients need at the point of care when choosing between two different interventions. CER studies attempt to make comparisons between two or more treatment strategies within populations reflective of the types of patients a provider might see in his or her practice and, as importantly, to ensure the conditions under which the comparison is made reflect the varied practice environments in which care is delivered in the general community.

2.3 The Evolution of CER

Distilled to its primary goals, CER compares two or more treatment modalities or approaches to the care of patients or populations. Thus, CER is research comparing the effectiveness of two or more preventive, diagnostic, therapeutic, or care delivery strategies using real-world approaches and under real-world conditions. The interventions compared in CER studies can be anything used in the care of patients or populations of patients including, health care interventions, treatment protocols, care delivery models, invasive procedures, medical devices, diagnostic tools, pharmaceuticals therapeutics, and any other strategies used for treatment, diagnosis, or prevention of illness or injury.

While the principles underlying CER have been around for a number of years, it is the recent emphasis on value in US healthcare and the transition from volume-based to value-based care that has brought the attention and support of policy makers for this type of research. Over the past decade, two important pieces of legislation have contributed to the growth of CER. The American Recovery and Reinvestment Act of 2009 allocated $1.1 billion to the Department of Health and Human Services, the National Institutes of Health, and the Agency for Healthcare Research and Quality stipulating that this funding should be used for the dual purpose of supporting research intended to compare the outcomes, effectiveness, and appropriateness of interventions for the prevention, diagnosis, or treatment of patients *AND* to encourage the development and use of more robust clinical data sources. This legislation also established the Federal Coordinating Council for Comparative Effectiveness Research whose charge was “to foster optimum coordination of CER conducted or supported by Federal departments and agencies”.

The second piece of legislation was The Patient Protection and Affordable Care Act passed by Congress and signed into law by President Obama in 2010 which established and funded the Patient-Centered Outcomes Research Institute (PCORI). Prior to the establishment of PCORI, there had been numerous efforts in both the private and public sectors to conduct CER studies and to generate comparative effectiveness data, but these efforts were limited by the lack of a unified definition for CER, variable funding priorities, and they lacked a robust means of tracking the types of studies being performed and on which topics. To fill these gaps, PCORI was created to become the primary funding agency for investigators performing CER. Since its inception, PCORI has:

- Provided $2.3 billion to help fund a broad portfolio of CER studies, develop research infrastructure, and disseminate and promote findings into actual practice.

- Established a policy for funded researchers to share their data, documentation, and statistical programming to encourage data sharing through open science.
- Developed methodologic standards (through the Methodology Committee) for performing CER and patient-centered outcomes research.
- Created a national data platform to support and improve the efficiency of conducting CER (i.e.: PCORnet).

PCORI has established National Priorities for Research in the following domains:

- Comparing the effectiveness and safety of alternative prevention, diagnosis, and treatment options to see which one works best for different people with a particular problem.
- Comparing health system–level approaches to improving access, supporting patient self-care, innovative use of health information technology, coordinating care for complex conditions, and deploying workforce effectively.
- Comparing approaches to providing comparative effectiveness research information, empowering people to ask for and use the information, and supporting shared decision-making between patients and their providers.
- Identifying potential differences in prevention, diagnosis, or treatment effectiveness, or preferred clinical outcomes across patient populations and the healthcare required to achieve best outcomes in each population.
- Improving the nation's capacity to conduct patient-centered outcomes research by building data infrastructure, improving analytic methods, and training researchers, patients, and other stakeholders to participate in this research.

A major criticism of contemporary clinical research is that the findings from very few studies actually fill a practical knowledge gap that can impact everyday clinical practice. Because a principal goal of CER is to improve individuals' ability to make informed healthcare decisions through the generation of data that can help patients, providers, and policy makers understand what treatment(s) will work best and for whom, a unique aspect of PCORI is the engagement of stakeholders such as patients, providers, and other decision-makers throughout the CER process. By involving stakeholders in the research process, the hope is that the most relevant questions and priorities can be identified, knowledge gaps can be better addressed, and approaches for dissemination and implementation of study findings can be optimized. It is this engagement that has led CER to be referred to at times as 'patient-centered outcomes research' and is believed to be a previously under-appreciated avenue for enhancing dissemination of data and translation into practice.

2.4 Conducting CER

CER is intended to impact the care of either individual patients or patient populations and can be conducted from various stakeholder perspectives. It can also affect health policy decisions as well as how or why care is delivered, organized, and paid for by health care entities. As such, a key component of CER is the external validity of the data or the ability to generalize the results to patients and clinical settings outside of the study population. Given the breadth of topics that can be addressed by CER, a variety of study designs and analytic methods are employed. However, prior to initiating a CER study, an understanding of the limitations of a given research question and specific study design are equally critical to the successful execution of a CER study with internal validity. In this regard, several important questions must be addressed during the study conception and design phase to ensure the right data source is selected, an appropriate study design is chosen, and appropriate statistical methods are employed.

- *Is the intent of the study to compare the effect of an intervention at the individual patient-level or at the population-level?*
  - Certain data allow for the analysis of patients clustered within hospitals, health systems, or geographic regions while others do not.
- *Is the research question appropriate for CER methods?*
  - The available data (or data that can be readily generated) must be able to answer the research question through the application of appropriate statistical methods.
- *Is the data source appropriate to address the chosen research question?*
  - Observational data sources used for CER often have important, unique, and inherent limitations that can create relevant sources of bias that must be considered and addressed either through the study design, the selection of the study population, and/or the methodology employed. In addition, for studies that truly seek to address a CER question, the data source should support the external validity of the findings.
- *Will the chosen study design and/or analytic methods minimize bias and enhance the internal validity of the findings?*
  - Investigators must have a working knowledge of available statistical tools and analytic approaches and understand the extent to which conclusions may (or may not) be supported by the data.

The EQUATOR (Enhancing the QUAlity and Transparency Of health Research) network is an organization that has developed a standardized set of reporting guidelines for many of the typical types of CER studies like RCTs and cohort studies. These guidelines were developed in a collaborative and multi-disciplinary fashion with input from funding agencies, journal editors, peer reviewers, and researchers with the primary goal being to elevate and improve the overall quality of contemporary published research. These guidelines can be helpful to ensure the rigor and clarity of presentation for CER studies.

2.5 Types of CER Study Designs

There are four principal, broad categories of study design used to conduct CER, each with their own advantages and limitations. These can be applied to generate new data to fill knowledge gaps in current clinical practice or to evaluate the existing evidence regarding benefits and harms of specific treatments when applied to different patients or populations. CER studies can either be prospective or retrospective and can be based on primary data collection or secondary analysis of existing data.

2.5.1 Randomized Clinical Trial

Data derived from RCTs remain the benchmark against which all other sources of data driving changes in clinical practice are compared. RCTs can span a spectrum from **explanatory** to **pragmatic** trials (Fig. 2.1). The majority of trials conducted and published are explanatory in nature and designed to address the issue of efficacy. As such, most explanatory trials will have study protocols with stringent inclusion and exclusion criteria. Not only are enrolled patients frequently extremely healthy, which may not reflect the real-world population of patients with a given condition, but the trial protocols also generally involve rigorous patient follow-up and monitoring, which also may not be indicative of typical day-to-day practice for

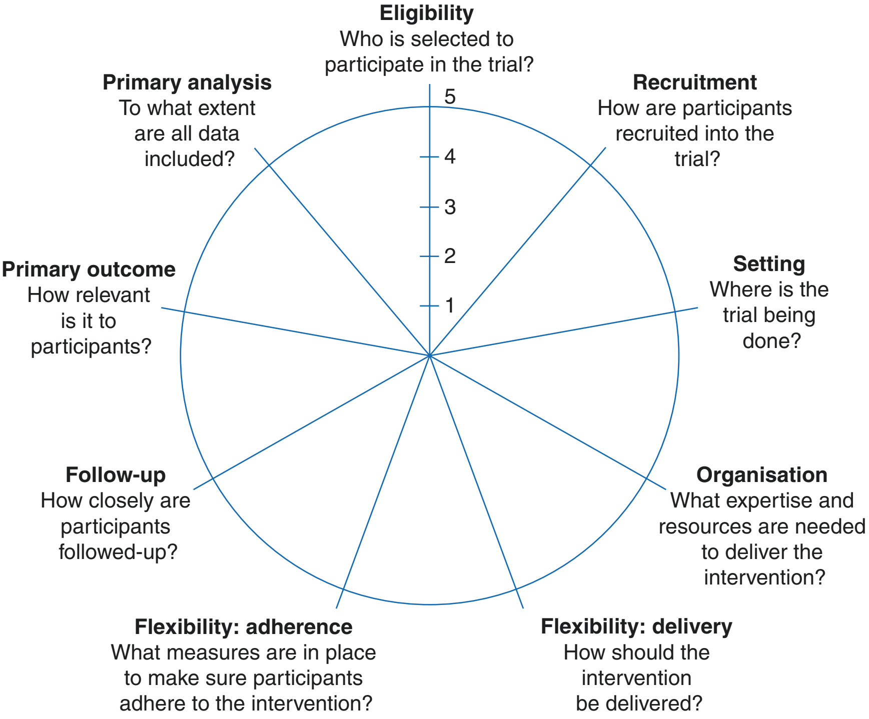

Eligibility
Who is selected to participate in the trial?
Recruitment
How are participants recruited into the trial?
Setting
Where is the trial being done?
Organisation
What expertise and resources are needed to deliver the intervention?
Flexibility: delivery
How should the intervention be delivered?
Flexibility: adherence
What measures are in place to make sure participants adhere to the intervention?
Follow-up
How closely are participants followed-up?
Primary outcome
How relevant is it to participants?
Primary analysis
To what extent are all data included?
5
4
3
2
1

Fig. 2.1 Tool for determining where a given RCT protocol falls in the explanatory to pragmatic continuum. Each of 9 domains for a given trail are scored from 1 to 5 (1 = very explanatory and 5 = very pragmatic) and then used to gauge where on that continuum it falls (taken from Loudon K, et al. BMJ. 2015)

providers in most practice settings. These are important drawbacks that can have important ramifications for the external validity of these types of studies.

By comparison, and as the name would suggest, pragmatic trials are intended to define the effectiveness of a given intervention and are more in line with the goals of CER. Whereas a strict study protocol and numerous inclusion and exclusion criteria are important for evaluating efficacy in an explanatory trial (as these features help to minimize any possible impact of confounding on study findings) this creates a critical blind spot for patients and practitioners—namely, how will this therapy work in routine clinical practice? In line with the goals of CER, pragmatic trials are intended to compare the effectiveness of varying treatments or management strategies with findings that can readily be generalized to most patients being treated in most clinical contexts or settings. To this end, the inclusion and exclusion criteria for such trials are typically more inclusive with study protocols that may even be flexible. In addition, study outcomes frequently only represent the most pertinent information required to address the research question and/or are the most easily assessed or adjudicated. A limitation of these studies is that the more parsimonious approach to data collection can limit the ability to conduct subgroup analyses or to perform post-hoc secondary data analyses addressing related questions.

Cluster RCTs are an example of a pragmatic trial design. In cluster trials, randomization is not performed at the individual level, but rather as a group, a practice, a clinic, or any other specified population. Within each cluster, patients receive usual care in addition to the experimental intervention and may not be aware they are participating in an RCT. This approach can markedly improve the external validity of study findings. However, a drawback to cluster trials is that because the unit of analysis is the cluster rather than the individual patient, the required sample size to ensure adequate statistical power may be larger and statistical methods, such as hierarchical models, must be used to address the within cluster correlation of the data (i.e.: patients treated within a given cluster are likely receiving similar care and thus are likely to have similar outcomes).

Stepped-wedge RCTs are another unique type of trial that can be considered a subtype of the cluster design. Whereas in cluster RCTs, each cluster is assigned to either the control or the intervention, in a stepped-wedge design all clusters initially start not being exposed to the intervention and will eventually receive the study intervention by the end of the trial, but the timing with which the intervention is administered is random. One of the benefits of these types of trials is that all study participants will receive the intervention. So, in cases where the intervention seems likely to be beneficial, this could enhance willingness for trial participation. Another benefit is the efficiency of the this design because the nature of the randomization process allows for each cluster to act as its own control. This also provides data that allows for both between and within cluster comparisons.

Adaptive RCTs are designed to allow changes to the protocol or the statistical analysis after trial initiation. Such changes are based on Bayesian analytic approaches as compared to the frequentist approaches typically employed in more traditional RCTs. This provides adaptive RCTs with a number of advantages. For example, protocol and/or procedural changes have already been approved as part of the trial design and, as such, can be implemented more efficiently. Total accrual and even study outcomes can change during the conduct of the trial as data accumulate. In this regard, adaptive RCTs can actually allow for more rapid study completion. However, by their nature adaptive RCT designs are more complex and as the trial protocol changes Bayesian analytic approaches become compulsory. As such, investigators should be well-versed in Bayesian statistics and ensure they have biostatistical support to ensure the integrity of trail results.

2.5.2 Observational Studies

Observational studies constitute the majority of contemporary HSR and outcomes research. The availability of numerous data sources, the efficiency with which data can be obtained and analyzed, and the relatively low associated costs for conducting this type of research are all reasons why these also represent a very common form of CER. In comparison to the rigorous protocols often used in controlled trials, an important feature of observational studies, in particular those based on the secondary use of local, regional, or national data sources (e.g.: administrative claims, registry data, or electronic health record data), is that they frequently reflect the actual management patients received during a given episode of care. Whereas, the emphasis in RCTs is frequently on internal validity sometimes at the expense of external validity, observational studies often implicitly emphasize external validity at the expense of internal validity. Specifically, although the data may reflect the type of care patients actually receive in real-world clinical practice settings and contexts, because of the non-controlled nature of observational studies, numerous sources of bias and confounding must be considered and either addressed through the study design, the selection of the study population, or through the application of various analytic and statistical approaches. Issues such as selection bias, confounding by indication, and missing data are all potential barriers to the internal validity of the findings from observational CER studies that must be considered and addressed. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for reporting of observational studies can provide investigators with a useful list of considerations when preparing a manuscript of their study results.

2.5.3 Research Synthesis

In certain situations, there may be a body of literature regarding a given intervention or treatment approach, but the data from individual studies could either be discordant or the sample size may not be large enough to clearly and definitively support a specific conclusion. In this context, **systematic reviews**, **meta-analyses**, and **health technology assessments** are valuable tools that can be used to synthesize the existing data. The goals of each of these three types of studies are distinct. In a systematic review, the goal is to provide an unbiased, comprehensive, clear summary of the body of data on a given topic. In meta-analysis, the goal is to combine the results of available studies on a given topic through quantitative techniques in order to create a collective data set that is better powered than its component data sources. In a health technology assessment, the goal is to generate data (specifically regarding a health technology) that can be used to inform clinical and policy-level decision making directed at the introduction and diffusion into practice of a given innovation. In all three types of research syntheses, the data used can be based on RCTs or on observational studies.

2.5.4 Decision Analysis

Decision analyses are informed by two types of data. The first is the probability of an outcome given a particular treatment or management pathway. The second is the patient's current and future health status, which inherently considers both the benefits and harms attributable to that treatment or pathway. These two components are applied and used to perform model-based quantitative evaluations of the outcomes associated with specific management strategies in specific situations. These are central study designs for CER because the underlying goal is to help patients and providers derive the best treatment decision for a specific patient in a specified clinical context or from a specific health-care perspective. Cost-effectiveness analyses also integrate aspects of decision analytic techniques to incorporate cost and quality of life inputs to assess the comparative value attributable to a given intervention or treatment approach. Through simulation modeling using best available data and assessing which parameters impact the outcomes most, future areas of needed research (i.e. RCTs or other prospective designs) can be prioritized.

2.6 Commonly Used Statistical Methodology

For observational CER, the appropriate selection and use of statistical methodology is critical for ensuring the internal validity of the study and for addressing sources of bias and/or confounding. While different statistical approaches might be appropriate for a given study, often times the 'correct' choice is predicated on the data source, the nature of the research question, and the experience or expertise of the investigative team. Additionally, using a combination of these statistical approaches can be helpful to evaluate the robustness of study findings in the context of varying assumptions about the data. Similarly, carefully planned subgroup and sensitivity analyses can also help to bolster the robustness of study results to varying assumptions.

2.6.1 Methods to Address Confounding

One of the most common approaches for addressing confounding (Fig. 2.2) is the use of a **multivariable model**. Models are used to estimate the effect of a given

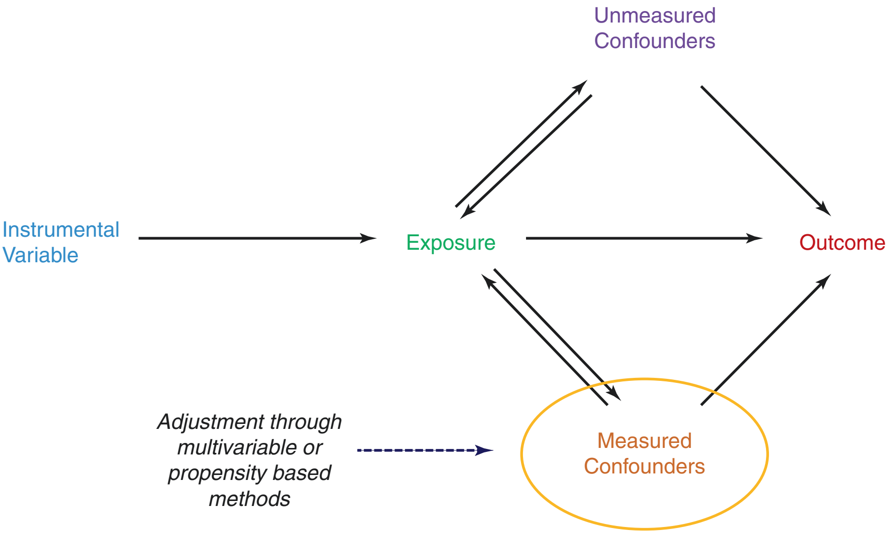


Fig. 2.2 Conceptual diagram of the association between exposure, outcome, confounders and potential instrumental variable in observational studies. Multivariable modeling and propensity score based adjustment can adjust for measured confounder, but neither approach is able to adjust for the effect of unmeasured confounders. Instrumental variables adjust for the effect of both measured and unmeasured confounders because they are only related to the outcome indirectly through the exposure and thus are not subject to confounders that may affect the association between exposure and outcome

exposure (e.g.: treatment) on a specified outcome while adjusting this estimate for the effect of factors that can potentially confound (i.e.: obscure) this relationship. The type of model used in a given study depends largely on the nature of the outcome of interest. For continuous outcomes (e.g.: post-operative length of stay), linear regression is most commonly applied. For binary outcomes (e.g.: perioperative mortality), logistic regression is frequently used. For time-to-event outcomes (e.g.: time from diagnosis to death), Cox proportional hazard regression is used. The benefits of multivariable models are that they are efficient and familiar for most investigators. In addition, there are hierarchical versions of these models that can be used to evaluate correlated data (e.g.: clustering of patients within a provider or hospital), to explore between and within cluster variation, and to address potentially less reliable estimates due to small sample size. In some instances, model performance can be improved by the inclusion of one or more interaction terms between covariates. An interaction occurs when the effect of one variable on the outcome of interest is predicated on the value of a second variable—also known as effect modification. The value of including interaction terms can be assessed by evaluating the model fit both with and without the inclusion of the interaction. An important limitation and consideration when using a model is that the completeness of adjustment is entirely predicated on the availability of data regarding measured confounders as the model cannot adjust for factors which are not measured or observed in the dataset.

The use of **propensity score analyses** has increased dramatically in recent years. In this type of analysis, the estimated probability (i.e.: propensity) for a patient to receive a given exposure relative to another is calculated. Propensity scores are most frequently used to estimate this probability when comparing the effect of two treatments on a given outcome. The score is derived by using a multivariable logistic regression model with the binary outcome of interest being the two possible treatments of interest. Other available factors that can potentially be associated with the receipt of a given treatment, the outcome, or are believed to be confounders of the relationship between the exposure and outcome are included as covariates in the propensity model. This estimated probability of treatment assignment can then be used in several ways to address potential confounding when comparing the effect of the two treatments on the outcome(s) of interest. For example, the propensity scores can be included as a covariate in the model estimating the association between the exposure and outcome which can be an efficient way to address issues related to statistical power for infrequently occurring outcomes. A popular approach is to perform propensity matching (Fig. 2.3). In this case, the

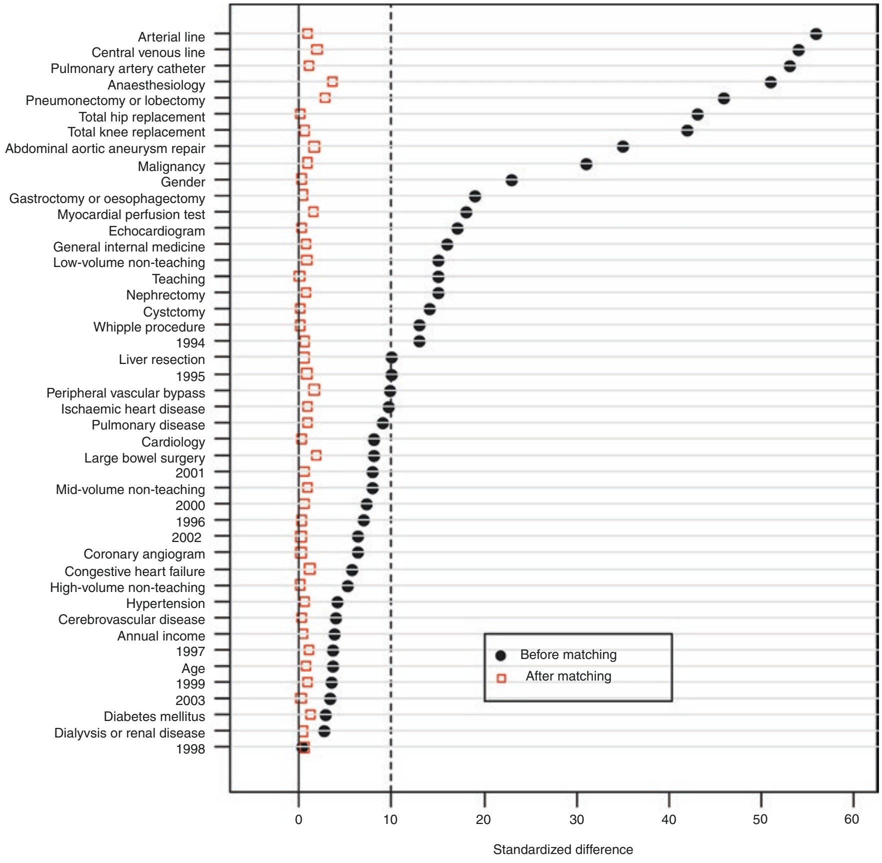


Fig. 2.3 Graphical representation of standardized differences before and after propensity matching of covariates (taken from Gayat E, et al. Intensive Care Med. 2010)

propensity score is used to identify patients with an identical or very similar propensity for having received a given treatment. After matching, observed covariates are often well-balanced and can appear to simulate what might be observed in the context of randomization. However, there are important limitations of propensity methods. A major limitation is that propensity score methods can only account for measured factors, while there may be a number of unmeasured factors that are important to treatment assignment (a.k.a. hidden bias) and can influence outcomes. While matching on propensity scores can result in a forced balance between observed covariates, it does not address confounding related to unmeasured factors and may actually exacerbate imbalance in such factors. In addition, matching can significantly reduce sample size and statistical power. As such, matching is often best applied in large datasets with numerous potential covariates where sample size is less of a consideration. Finally, it is unclear that propensity scores achieve significantly different estimates as compared to multivariable modeling.

In **instrumental variable (IV) analyses**, a specific variable is chosen to serve as the "instrument" for comparing two interventions. An instrument is chosen such that it is the external cause of the intervention or exposure of interest, but by itself is unrelated to the outcome except through the causal pathway (Fig. 2.2). Randomization of patients in RCTs is an example of an instrument—the treatment a patient receives is entirely predicated on the randomization, but randomization has no effect on the outcome except through the treatment the patient receives. As an example, a CER study might seek to utilize an existing data source to compare the effect of a minimally invasive surgical approach relative to an open approach on a given outcome. However, because a variety of factors play a role in a clinician's decision regarding whether or not to recommend a minimally invasive approach (e.g.: prior surgery in the chest or abdomen; the patient's body habitus; other anatomic considerations at the surgical site; concurrent co-morbidities), a simple comparison of patients treated with these two approaches could be biased because of confounding by indication (based on unmeasured factors). An IV for this comparison might be hospital-level or regional rates of minimally invasive surgery. When patients are categorized into groups based on the value of the instrument, the rates of treatment will differ, the probability of treatment is no longer affected by potential confounding characteristics of an individual patient, and the comparison of interest becomes analogous to comparing randomized groups. Relative to multivariable models and propensity scores, an important benefit of IV analyses is that they not only address imbalance in measured confounders, but they are also believed to address imbalance in unmeasured variables as well. In this respect, estimates from IV analyses are believed to be better for addressing residual confounding (i.e.: confounding from unmeasured or unadjusted factors) and more accurately reflect the true association between a given exposure and outcome in observational studies.

2.6.2 Addressing Sources of Bias

While confounding is related to the effect of one or more measured or unmeasured factors that can obfuscate the association between an intervention and an outcome of interest, bias is a form of error within the design or analysis of a study that can also distort the estimate of the exposure-outcome relationship. Whereas confounders are typically addressed through model-based adjustment, bias is more effectively dealt with through either study design, selection of the study population, or the use of specific statistical approaches.

Missing data is frequently an issue when using observational data sources. There are two main consequences of not adequately addressing missing values. The first, is that the sample size (and thus the power) of a study can be significantly decreased if case-complete (i.e.: analysis of only patients with non-missing data) approaches are selected. Methods such as imputation can be useful to address this issue. However, prior to doing so, it is important to consider the second issue which is the introduction of bias. It is important to consider which variables have missing data and why, how patients with missing data differ from those without missing data, and whether missing values can be predicted based on observed data. Multiple imputation methods are frequently used to address missing values and are believed to provide better powered, unbiased estimates in cases when data are missing completely at random or missing at random. In cases where data are missing not at random (i.e.: missing values are related to unmeasured, non-random, patient-level factors), any methods of addressing missing data will likely result in the introduction of bias.

Selection bias occurs when allocation of study subjects to a given intervention does not accurately reflect what happens in actual practice. For example, an observational study might demonstrate a clear benefit associated with the use of adjuvant chemotherapy in patients with colon cancer. However, an important factor that could introduce selection bias into this analysis is whether the data source provides information on postoperative complications (like surgical site infections which are common after colorectal surgery). If patients were simply categorized based on whether or not they received adjuvant therapy without accounting for patients who may not have received adjuvant therapy because they had a postoperative complication, the observed benefit in patients who received adjuvant therapy could be explained by the better postoperative outcome that would be expected when a complication does not occur rather than any effect attributable to the adjuvant therapy itself. Careful selection of the data source and the patients included in this type of study as well as well-selected sensitivity analyses are useful approaches to mitigate, to the extent possible, the effect of selection bias.

Survivor treatment bias is a particularly important consideration in oncologic studies evaluating the survival benefit of adjuvant interventions occurring after surgery. In order for a patient to receive a treatment after an operation, they must survive through the post-operative period. Put simply, patients who live longer after an operation have more of an opportunity to receive additional treatment. A l*andmark analysis* can be a useful approach to address this issue. In landmark analyses, survival is estimated for groups of patients conditional on the fact that they have survived to at least a specified time point—the landmark (e.g.: all patients in the analysis survived at least 90-days beyond surgery).

Lead-time and **length-time bias** are both relevant to studies evaluating screening interventions. *Lead-time bias* occurs when the survival benefit associated with a given intervention is due entirely to the earlier detection of a disease (as opposed to the patient presenting after it has become clinically apparent) rather than any actual effect of the intervention itself. Put differently, it is the time interval between when a disease is detected by an intervention relative to when it would be typically diagnosed. This type of bias can make cancer screening interventions appear to make patients live longer. By comparison, *length-time bias* occurs when slow progressing cases are detected more often and thus patients live longer.

2.7 Limitations of CER Studies

Conducting research studies that can provide meaningful, generalizable data while at the same time ensuring internal validity by anticipating relevant sources of bias and confounding can be a real challenge even for experienced investigators. This is one reason why RCTs remain the benchmark against which all other types of studies attempting to inform evidence-based practice are judged. However, RCTs can be prohibitively costly and time consuming. When done properly, the estimated effect sizes obtained from well-performed observational CER studies can be quite similar to those obtained from RCTs. There are considerations that both consumers of the peer reviewed literature and investigators should keep in mind when interpreting the findings from a CER study.

It is important to ask if the direction and size of the observed association is believable and consistent with what may already be known. In observational studies, there may be a tendency to believe that estimates derived from the use of advanced statistical techniques by themselves address all sources of bias and thus provide valid estimates of the association—this is simply not the case. Assuming a given statistical approach has been applied correctly and the analysis is sound, frequently the data source used to conduct an observational study may have specific nuances or limitations that are not fully considered during the conduct of the study and can result in biased estimates. In cases where the size of the observed effect is too large relative to what is known from existing RCT data, the results should be viewed with a circumspect eye and consideration should be given to the manner in which relevant sources of bias may have affected the findings. On the other hand, if the findings are corroborated across a variety of data sources, patient populations, and/or using different statistical methodologies, this can lend credence to the study findings.

In observational studies, an association should not be immediately interpreted as causality. However, there are established criteria that can support the conclusion of causal inference for an observed association. Of the nine historically described criteria, the following six are the most relevant to observational CER studies: strength; consistency; specificity; temporality; presence of a biological gradient; plausibility. The more of these criteria that are present for the findings of an observational study, the stronger the case that a true association exists.

2.8 Barriers to the Conduct and Implementation of Findings from CER Studies

Although the mission and value of CER are well-established, important and existing barriers within the US healthcare system remain to the generation and implementation of new data. A wealth of data clearly demonstrates that there are national disparities in care for certain populations, that there is ongoing variation in the quality and costs of care, and that health care in the US costs more and is of lower quality than in other comparable industrialized nations. Despite general agreement on the reasons that change is needed and, perhaps to a lesser extent, the manner in which this change should occur, there are numerous legislative obstacles to implementation of findings from CER studies that could inform the transition toward more value-based health care models.

Research supported by PCORI is intended to improve care quality, increase transparency, and increase access to better health care for all patients. However, PCORI is explicitly prohibited from funding research studies that evaluate or apply cost-effectiveness thresholds or utilize quality adjusted life years. There is also specific language in the act that the reports and research findings may not be construed as practice guidelines or policy recommendations and that the Secretary of Health and Human Services may not use the findings to deny coverage. These stipulations are likely based on societal and/or political fears that research findings could lead to the potential for health care rationing.

Current spending on prescription drugs is estimated at between $400–550 billion. Although many new and novel therapeutic drugs are brought to market, many provide only added costs with minimal clinical benefit. While CER studies (in particular cost-effectiveness analyses) of various types of drugs would be of great value, at present there are statutory limitations in how these data could be used. The Medicare Modernization Act in 2003 established Medicare Part D to provide beneficiaries with prescription coverage. But, it also stipulated that the Secretary of Health and Human Services could neither establish a formulary or negotiate drug prices. The research community will need to work with policy makers and legislators to overcome hurdles such as these to ensure that data from CER studies can fulfill their intended mission and better inform the care of patients within the US healthcare system.

2.9 Conclusion

For policy makers, CER has become an important priority in an effort to identify ways to address the rising cost of healthcare and the shift toward more value-based care models. Acknowledged variation in quality and outcomes as well as an ever-increasing number of new therapeutic options creates a need for a steady stream of data that can better inform patients, providers, and other stakeholders as to the incremental value of a given treatment in a real-world context and to identify and promote the most effective interventions. Because the goal of CER is ensure that individual patients, providers, and the US healthcare system as a whole make the best healthcare decisions, it will be imperative for the health care and research communities to work in tandem to conduct impactful CER studies on relevant topics and, even more importantly, break down barriers to the dissemination and implementation of data from these types of studies.

Acknowledgements The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, Baylor College of Medicine, or MD Anderson Cancer Center.

Conflict of Interest Disclosures: None of the authors have any conflicts of interest to report.

Funding/Support: This work was supported by the United States Department of Veterans Affairs HSR&D Service of the VA Office of Research and Development Merit Review award I01 HX002447 (NNM) and the Center for Innovations in Quality, Effectiveness and Safety award CIN 13–413 (NNM).

Role of the Sponsor: The funding body played no part in the design and/or general conduct of this study, had no access to the data or role in data collection, management, analysis, or interpretation, and had no role in preparation, review, or approval of the manuscript.

Further Reading

1. Conway PH, Clancy C. Comparative-effectiveness research—implications of the Federal Coordinating Council's report. N Engl J Med. 2009;361(4):328–30.
2. Patient-Centered Outcomes Research Institute. National priorities and Research Agenda. Available at https://www.pcori.org/research-results/about-our-research/research-we-support/national-priorities-and-research-agenda. Accessed 4 September 2018.
3. EQUATOR network. Reporting guidelines for main study types. Available at https://www.equator-network.org/. Accessed on 10 September 2018.
4. Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA. 2007;297(3):278–85.
5. Newgard CD, Lewis RJ. Missing data: how to best account for what is not known. JAMA. 2015;314(9):940–1.
6. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. JAMA. 2015;314(18):1966–7.
7. Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes. 2011;4(3):363–71.
8. Sox HC, Goodman SN. The methods of comparative effectiveness research. Annu Rev Public Health. 2012;33:425–45.
9. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014; https://doi.org/10.1002/14651858.MR000034.pub2.

3 Understanding Variations in the Use of Surgery

Jesse A. Columbo and Philip P. Goodney

Abstract

This chapter details the origins of the study of variations in health care, with special attention to variations in the use, indications, and outcomes in surgery. We review the initial studies that demonstrated the value of this methodology, describe how the study of variations allows insight into surgical practice, and how limiting variation can improve patient outcomes across a variety of surgical specialties. Finally, we note how the study of variation can be used to inform care for patients and be used for early outcome detection by device and drug regulators.

3.1 Introduction

Any surgeon "worth their salt" will attest that surgery can be technically difficult at certain times. Retro-hepatic vena cava exposure, penetrating trauma to Zone III of the neck, and drainage of a pancreatic pseudocyst are three examples of technically difficult surgical exercises that can test even the most experienced technicians.

But more often than not, the most challenging aspect of surgery is deciding when – and when not – to operate. Surgeons of all specialties face these difficult scenarios every day. A 60 year old man with an elevated prostate-specific antigen (PSA) test. A 75 year old woman with critical but asymptomatic carotid artery

J. A. Columbo
Vascular Surgery, Section of Vascular Surgery, Dartmouth-Hitchcock Medical Center,
Lebanon, NH, USA

P. P. Goodney (✉)
Section of Vascular Surgery, Dartmouth Medical School, Department of Surgery,
The Dartmouth Institute for Health Policy and Clinical Practice, Dartmouth-Hitchcock
Medical Center, Lebanon, NH, USA
e-mail: Philip.P.Goodney@hitchcock.org

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_3

stenosis. A 50 year old man with back pain. In each of these clinical settings, there are a variety of approaches that could be the "right" answer.

Some argue that each of these scenarios allows the practicing surgeon to emphasize the "art" of surgery as much as the science of healthcare in deciding whom should be offered surgical treatment. A careful, informed conversation in the office, outlining the risks and benefits of a surgical approach, is a cornerstone of an effective clinic visit. Discussing the options with the patient and their family, and formulating a plan that leaves both patients and physicians feeling like they are "making the best choice" can be both satisfying and rewarding.

But many of us approach different scenarios in different ways, and different approaches lead to different practice patterns. The study of the variation in these practice patterns is the focus of this chapter. This variation can be helpful – by introducing "natural experiments" wherein different approaches can be studied. But this variation can also be harmful, resulting in overuse or underuse of surgical treatments, with loss of their potential benefits or exacerbations of their potential harms.

This chapter will introduce the study of variation and its potential implications in surgery. In our current era of patient safety, some study surgical practice as they would other complex systematic processes, such as manufacturing an automobile. In automobile manufacturing, variations can produce remarkable results – such as an elegant, beautiful hand-made sports car. However, while beautiful and elegant, a Ferrari (for many reasons) would not serve as an ideal mode of transportation for large populations. Rather, many argue Henry Ford's approach – standardization and eliminating variation – may be much better. By limiting variation and ensuring quality, Ford delivered a better car, at a lower price, to more drivers than any of his competitors around the world.

Surgery is certainly not an assembly line. However, there are shared characteristics between complex processes in manufacturing, and complex processes in patient selection and process measure performance in surgery. Limiting variation in both settings leads to better results. In this chapter, we explore the beginnings of efforts to study variations in health care, examine the progression towards variations in surgery and subspecialty surgery, and finally outline how attempts to limit variation have – and will – affect health policy in the United States.

3.2 The First Classic Papers: In the Beginning

The study of variation in health care, and surgery in particular, began in the early 1970's, when J.P. Bunker examined differences in the numbers of surgeons and the number of operations between the United States [1]. This overview studied relationships between the supply of surgeons, the number of procedures, and health policy in a nationalized health care setting (England) and a fee-for-service environment (the United States). These insights prompted early interest. But subsequent work by John Wennberg, an internist and nephrologist with public health training, prompted the first real interest and excitement surrounding the variation in patient care.

Interestingly, this work was not published in JAMA or the New England Journal, but in a journal with an even broader impact – Science [2].

This landmark paper was not a broad, sweeping analysis comparing entire health systems, as the Bunker analysis had done a few years earlier. Rather, Wennberg's approach was exactly the opposite. He chose to approach the problem from a different, and entirely novel, perspective – by studying small area variation. Instead of examining practice patterns across the country, he examined them across an entire state, and a small one at that – Vermont, which was (and still is) a small state, with a population of around 600,000 residents. As shown in Fig. 3.1, Dr. Wennberg studied patterns of care by creating a new "unit of analysis" - the hospital service area. Wennberg and his colleague, Alan Gittleson, painstakingly examined each community in Vermont, and studied where patients sought surgical care. He categorized the

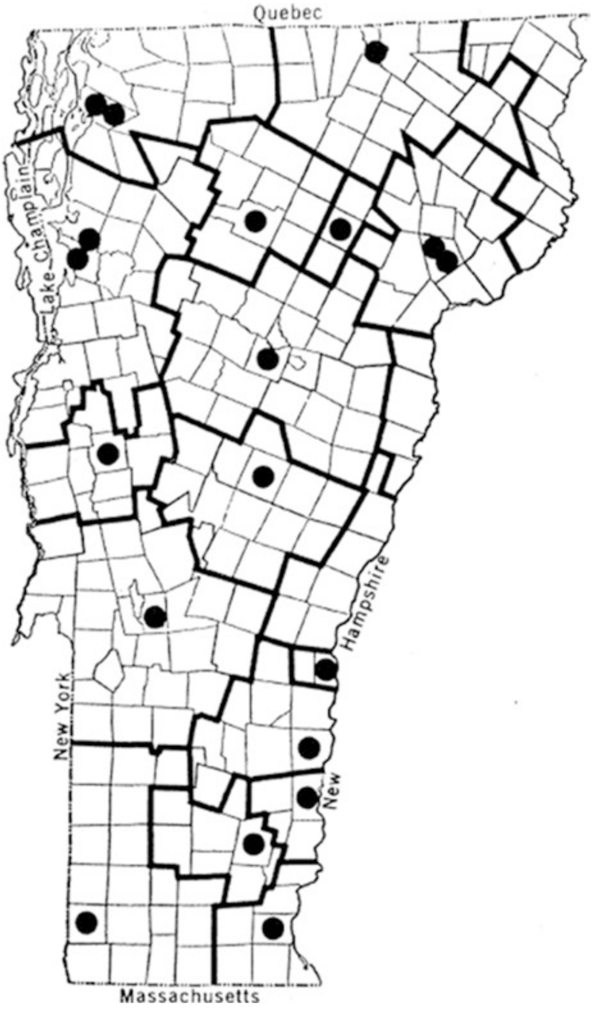


Fig. 3.1 Map of Vermont demonstrating hospital services areas. (Reproduced with permission according to JSTOR, Science , Vol 182, No 4117 (Dec 14, 1973, pp 1102-108.)

state into thirteen distinct regions – terms hospital service areas – where each service area represented the population of patients served by a community or academic hospital or group of hospitals. Using these methods, Wennberg was able to study variation in patients, utilization, and outcomes across these service areas.

Wennberg's findings were striking. Tonsillectomy rates per 10,000 persons, adjusted for age, varied from 13 in some regions to 151 in others. Similar extent of variation was seen in appendectomy (10–32 per 10,000 population), prostatectomy (11–38 per 10,000 population), and hysterectomy (20–60 per 10,000 population). And when Wennberg looked for explanations for these striking variations, he found a simple but elegant explanation. The more physicians and hospitals in a service area, the more services they provided. These relationships held fast across a broad variety of measures – number of procedures, population size, and number and type of specialists. What Wennberg did not find was large differences in patients across the communities in Vermont. Patients, overall, were similar – but the amount of care they received was not.

Wennberg concluded, in this early work, that there are wide variations in resource input, utilization of services, and expenditures – even in neighboring communities. Further, these variations in utilization seemed to be directly related to considerable uncertainty about the effectiveness of specific health services. His "prescription" for these uncertainties was to spend the next 40 years attempting to use informed choice to leverage patient decision-making towards trying to limit variations in care.

3.3 Gaining Momentum: Bigger Is Better

Building on these initial analyses, Wennberg and colleagues sought to broaden their work from a small state in New England to more representative – and generalizable – insights about the extent of variation occurring across the United States. Accordingly, Wennberg, John Birkmeyer, and colleagues used an aggregate of the hospital service area – called the hospital referral region – to study variation in common surgical procedures [3]. While the hospital service area studied care at the level of neighborhoods and communities, the hospital referral region (n = 306 across the United States) studied care at the level of a regional referral center. And, instead of using data from one state, Medicare claims were selected to provide a national, generalizable view of variations in surgical care.

Birkmeyer's findings centered around two important principles. First, just as Wennberg found dramatic variation across different – and sometimes neighboring communities in Vermont, Birkmeyer found dramatic variation across different hospital referral regions across the United States. For example, as shown in Fig. 3.2, rates of carotid endarterectomy varied nearly three-fold across different regions of the United States. Maps demonstrating region variation, inspired by the work of investigators in the Dartmouth Atlas, became universal in terms of a way to demonstrate regional differences in utilization. Darker areas represented areas where procedures were performed more commonly, and lighter areas represented the areas where procedures were performed less commonly. These representations brought

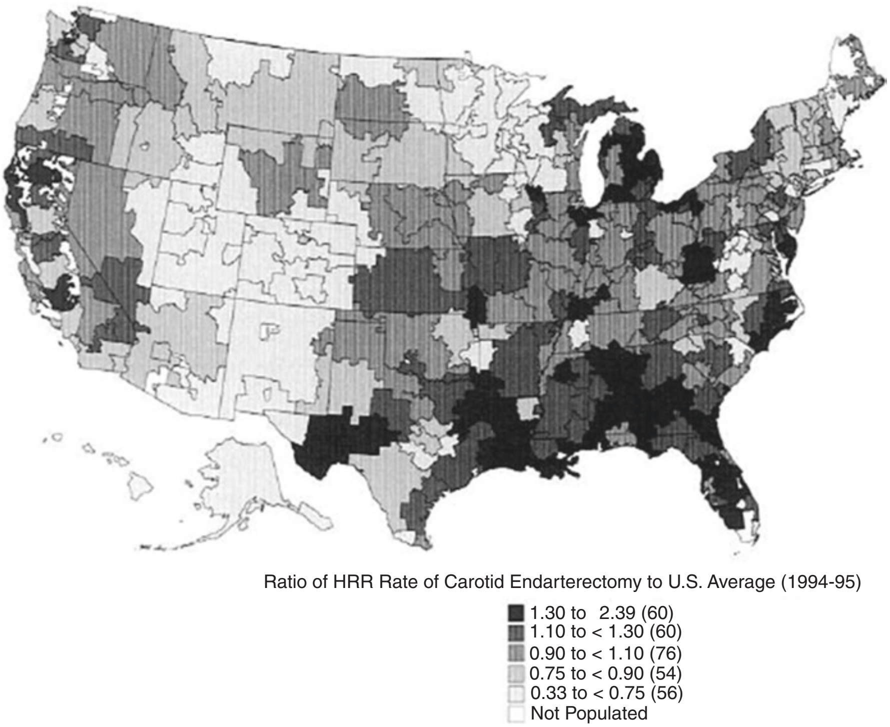


Fig. 3.2 Map demonstrating variation in rates of carotid endarterectomy across the 306 hospital referral regions of the United States. (Reproduced with permission from Elseivier, Surgery 1998 124:917-923.)

these differences to stark contrast, and one cannot help looking at the map and seeing what color – and utilization rate – is reflected in the region you call home.

The second important finding this work demonstrated was that the extent of variation was different across different types of operations. As shown in Fig. 3.3, there were certain operations where consensus existed, in terms of when to proceed with surgery. Hip fracture demonstrated this axiom quite nicely, and unsurprisingly so. The indication for surgery is clear in this setting, as a hip fracture is easy to diagnose. The benefits are easily seen as well, as all but the most moribund patients do better with surgery than with non-operative care. Therefore, there is little variation across the United States in terms of the utilization of hip fracture surgery. Figure 3.3 demonstrates this concept by showing each hospital referral region (HRR) as a dot, and listing the procedures across the x-axis. All HRRs cluster closely together for procedures like hip fracture.

However, for procedures like carotid endarterectomy, back surgery, and radical prostatectomy, the HRRs spread over a much wider range. These procedures, unlike hip fracture, are much more discretionary in their utilization. In general, it is evident that procedures with the highest degree of variation reflect areas of substantial disagreement about both diagnosis (what does an elevated PSA really mean) and

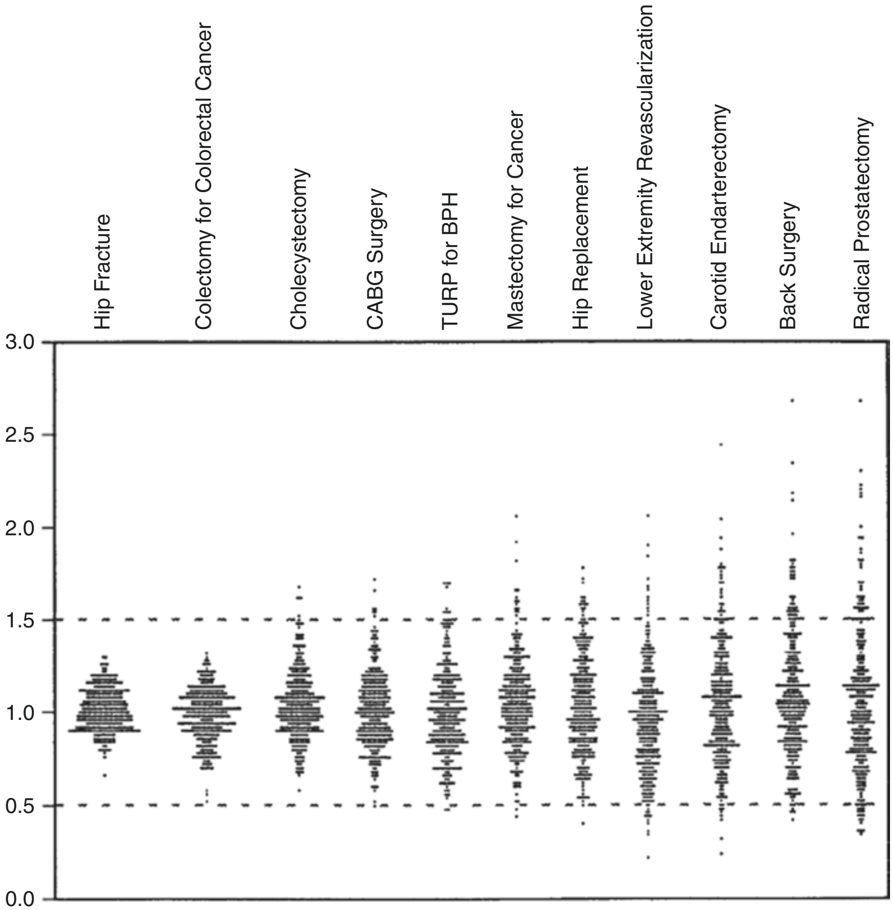


Fig. 3.3 Variation profiles of 11 surgical procedures. (Reproduced with permission from Elseivier, Birkmeyer et al, Surgery 1998 124:917–923.)

treatment (is back surgery really better than conservative treatment)? Dealing with this variation will require, Birkmeyer argues, better understanding of surgical effectiveness, patient-specific outcome assessment, and a more thorough understanding of patient preferences. Patients, clinicians, payers, and policymakers all will need to work together, he argues, to determine "which rate is right."

3.4 Innovating Approaches, and Integrating Ideas – From Medicine to Surgery

After these publications in the early 1990s, Wennberg and his colleagues spent the next decade refining analytic methods, and incorporating what seemed to be a recurrent theme in their work: that there was significant variation in the provision of medical care, and more care was not necessarily associated with better outcomes. But critics wondered if this work, limited in clinical detail, actually reflected different care on similar patients – because clinical variables for risk adjustment were commonly unavailable. To deal with these limitations, researchers began to use clinical events – such as death to create cohorts similar in risk strata.

In the most prominent of these approaches, Wennberg and Fisher created cohorts of patients who were undergoing care – medical, surgical and otherwise – at the end of life [4, 5]. By studying care provided in the last year of life, they argued, all patients in the cohort had similar 1-year mortality – 100%—therefore limiting the effect of any un-measurable confounders. This research, published in 2003 and widely referenced, concluded that nearly 30% of spending on end of life care offers little benefit, and may in fact be harmful.

Surgeons were quick to translate these innovative approaches, and integrate these ideas into surgical analyses. In a manuscript published in the Lancet in 2011, Gawande, Jha and colleagues adopted this technique and studied surgical care in the last year of life [6]. They had two basic questions. First, they asked if regional "intensity" of surgical care varied by the number of hospital beds, or by the number of surgeons in a region. And second, they examined relationships between regional surgical intensity and its mortality and spending rate.

Their team found that nearly one in three Medicare patients underwent a surgical procedure in the last year of life, and that this proportion was related to patient age (Fig. 3.4). Regions with the highest number of beds were mostly likely to operate on patients in the last year of life (R = 0.37), as were regions where overall spending in the last year of life was highest (R = 0.50). These findings reinforced earlier considerations about the need for patient-specific outcomes, and patient preferences in the provision of care at the end of life.

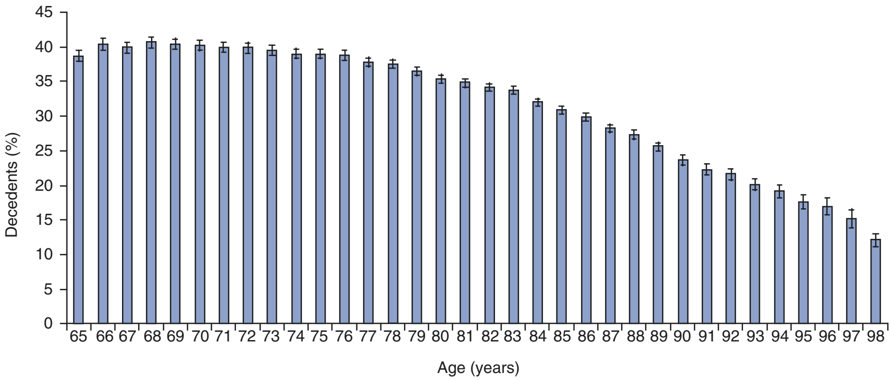


Fig. 3.4 Percentage of 2008 elderly Medicare decedents who underwent at least one surgical procedure in the last year of life. (Reproduced with permission from Elseivier, Kwok et al, Lancet 2011; 378:1408–1413.)

3.5 Specialty Surgeons and their Efforts in Describing and Limiting Variation

Many of the previously described investigations approached the subject of surgical variation using broad strokes – studying procedures as diverse as hip fracture, lower extremity bypass, and hernia repair, all within in the same cohorts. These approaches garnered effective, "big-picture" results, and surgeons grew interested in studying variation. Just as Wennberg sought to establish precise detail in the level of variation, surgeons now grew interested in exploring the different extent and drivers of variation across different specialties. In this section, we discuss three areas of subspecialty variation spine surgery and vascular surgery.

3.5.1 Variation in Spine Surgery

Patients presenting with back pain are a diverse cohort, and treatment with surgery is used at different rates in different parts of the country. As interest in studying the extent of variation and its causes began to build momentum, Weinstein and colleagues explored variation in the use of spine surgery for lumbar fusion [7]. These interests were brought to the fore with the development of devices such as prosthetic vertebral implants and biologics such as bone morphogenetic protein, all placed into everyday practice with a dearth of high quality evidence from randomized trials.

Weinstein and colleagues saw these changes occurring in "real-time", in the context of their clinical interests as spine surgery specialists. They found that rates of spine surgery rose dramatically over between 1993 and 2003. By 2003, Medicare spent more than 1 billion dollars on spine surgery. In 1992, lumbar fusion accounted for 14% of this spending, and by 2004, fusion accounted for almost half of total spending on spine surgery (Fig. 3.5). These observations led them to investigate the extent of this variation. What they found was truly remarkable. As shown in Figure x, there was nearly a 20-fold range in the rates of lumbar fusion across different hospital referral regions – the largest coefficient of variation reported with any surgical procedure to that date, a value five-fold greater than any variation seen in patients undergoing hip fracture. These data served to motivate extensive funding for the SPORT (Spine Patient Outcomes Research Trial), one of the largest continually funded randomized trials funded by the National Institutes of Health [8].

3.5.2 Variation in Vascular Surgery

Clinical changes have motivated research into variation in other areas as well, especially in patients with lower extremity vascular disease. Much like cages and bone proteins revolutionized spine surgery, the development of endovascular techniques revolutionized the treatment of lower extremity and cerebrovascular occlusive disease. Before the mid-1990s, patients with carotid stenosis or lower extremity vascular disease, for the most part, had only surgical options for revascularization.

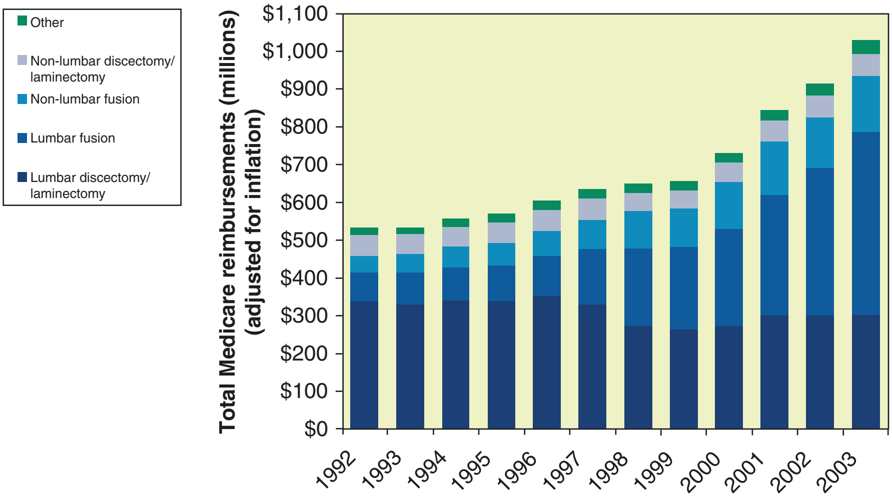


Fig. 3.5 Total Medicare reimbursements, for lumbar spine surgery, by year. (Reproduced with permission from Elseivier, Weinstein et al, Spine 2008; 33:2789–2800.)

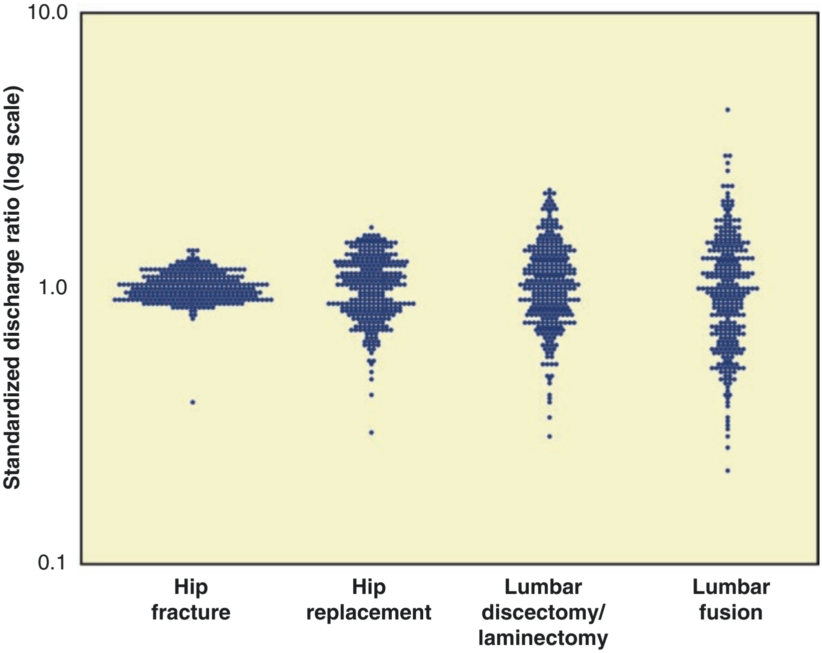
However, with the endovascular revolution, dramatic changes occurred in two important ways.

Given a less invasive endovascular option, many patients who were not candidates for open surgery could now undergo less-invasive endovascular surgery. And, because these approaches no longer required a surgical approach, the pool of potential practitioners now grew instantly – from surgeons alone, to surgeons, radiologists, interventional cardiologists, and a variety of catheter-trained specialists.

Motivated by these changes, Goodney and colleagues explored trends in lower extremity revascularization [9], aortic aneurysm repair [10], and carotid revascularization [11]. They found that temporal changes occurred in the national utilization of open surgical repair and endovascular interventions for all of these procedures (Fig. 3.6). And moreover, changes in specialty profile often were linked directly to changes in the utilization of these procedures. For example, cardiologists and vascular surgeons, over the last decade, came to supplant interventional radiologists as the principal providers of lower extremity revascularization (Fig. 3.7). Therefore, changes in the types of providers, as well as the types of procedures, often contribute to variation in utilization.

3.6 Informing Health Policy – Limiting Variation in Surgical Care

Over the last decade, the evidence became irrefutable that unwarranted variation was present in many aspects of surgery, especially in settings where new technology and uncertainty in indications cross paths. Upon these foundations, surgeons began to study the effect of these variations on outcomes. In this section, we will review the



	Hip fracture	Hip replacement	Lumbar discectomy/ laminectomy	Lumbar fusion
Mean	7.7	3.2	2.3	1.1
Extremal ratio	3.62	5.59	7.96	21.0
Interquartile ratio	1.19	1.49	1.62	2.01
Coefficient of Variation	11.9	25.4	34.6	49.5

Fig. 3.6 Variation profiles of orthopedic procedures. (Reproduced with permission from Elseivier, Weinstein et al, Spine 2008; 33:2789-2800.)

manner in which variation in care can create “natural experiments” – settings where patients receive different care because of variation, and where these differences can be used to examine the effect of different exposures on outcomes (Fig. 3.8).

In our first example, Ghaferi and colleagues studied variation in surgical mortality after several common general and vascular surgery procedures [12]. Using data from the National Surgical Quality Improvement Program (NSQIP), they categorized hospitals according to their mortality rates, and found significant variation in mortality across hospitals. While this was interesting in and of itself, they also noted key differences between hospitals when they studied surgical complication rates, in hospitals with low mortality rates, as well as in hospitals with high mortality rates. Their study noted that while complication rates were similar across mortality risk, those patients treated at hospitals with the highest mortality were most likely to die following complications – unlike patients in low mortality hospitals, who were likely to be “rescued’ following a surgical complication and were unlikely to die from it (Fig. 3.9). These data provided powerful evidence, and a guideline for national quality improvement efforts aimed at limiting mortality with inpatient surgery. To address hospitals with high mortality rates, the most effective strategies

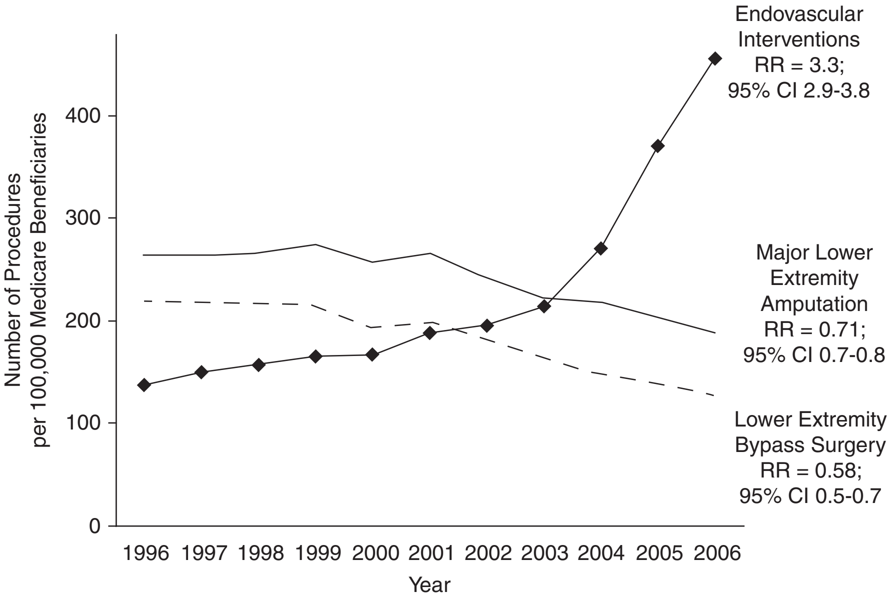


Fig. 3.7 Rates of endovascular interventions, lower extremity bypass surgery, and major lower extremity amputation in Medicare patients 1996–2006. (Reproduced with permission from Elseivier, Goodney et al, Journal of Vascular Surgery 2009; 50:54–60.)

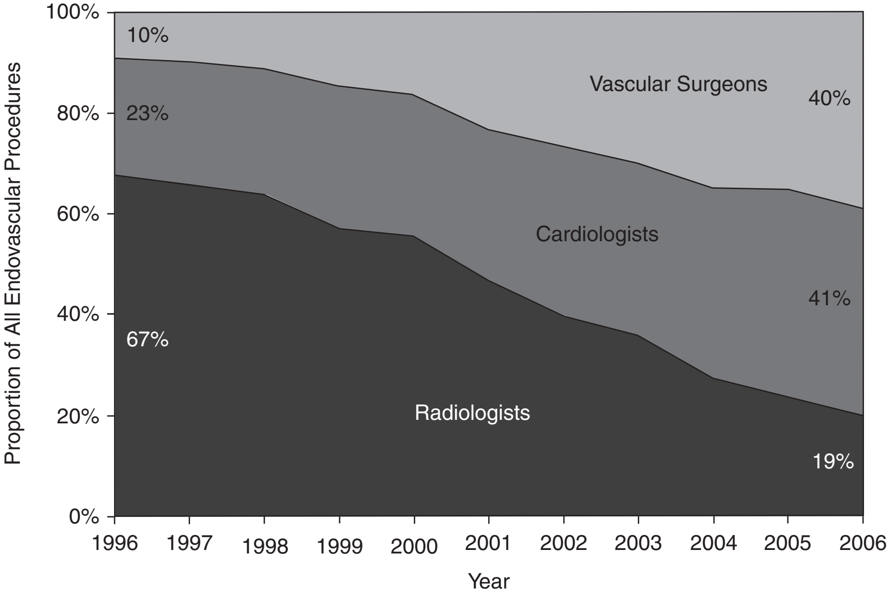


Fig. 3.8 Proportion of endovascular interventions, by specialty. (Reproduced with permission from Elseivier, Goodney et al, Journal of Vascular Surgery 2009; 50:54–60.)

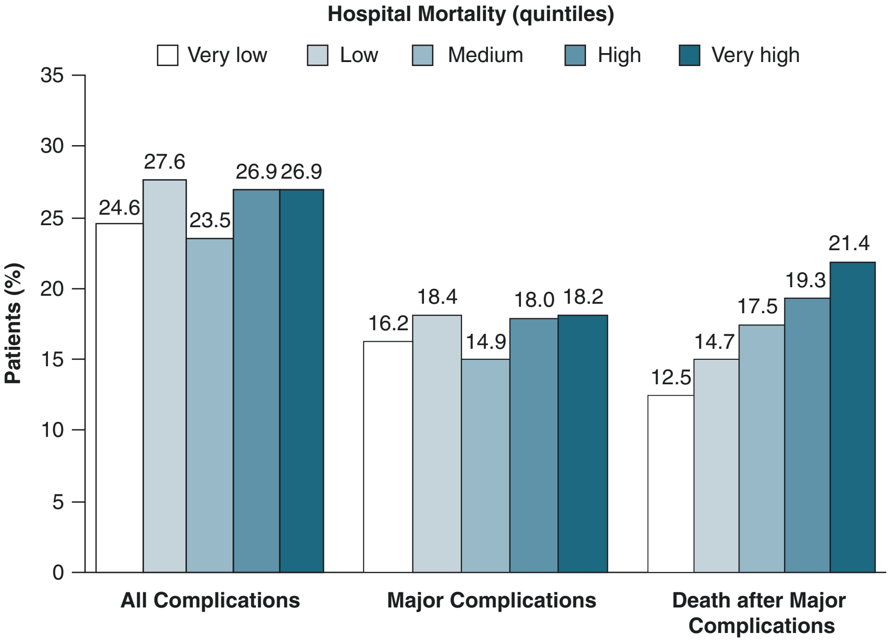


Fig. 3.9 Rates of all complications, major complications, and death after major complications, by hospital quintile of mortality. (Reproduced with permission, personal communications Dr. Dimick, New England Journal of Medicine 2009; 361: 1368–1375.)

may not be to simply try to limit complications, but instead may try to teach physicians and hospitals how to better deal with complications after they occur.

In our second example, Goodney et al. examined the effect of different levels of intensity of vascular care on outcomes in patients with severe lower extremity peripheral arterial disease [13]. Using the regional rates of intensity of vascular care as an exposure variable, their study noted that those regions most likely to perform invasive vascular care tended to have lower population-based rates of amputation. In other words, no matter how you measured it – diagnostic vascular care, invasive endovascular procedures, or open surgical reconstructions, more vascular care was closely correlated with lower rates of amputation for population in those regions (Fig. 3.10). These data were vital in arguing that patients at risk for amputation require attention and identification for more care, unlike other areas in surgery where variation has primarily focused on under-treatment.

3.7 Informing Health Policy – Understanding Variation and Implementing Change

Simply identifying variation in surgical care is not enough. For improvements in health care to be made, patients, policymakers, and stakeholders, must understand what variation means, and how this information can be used to inform changes that will benefit patients.

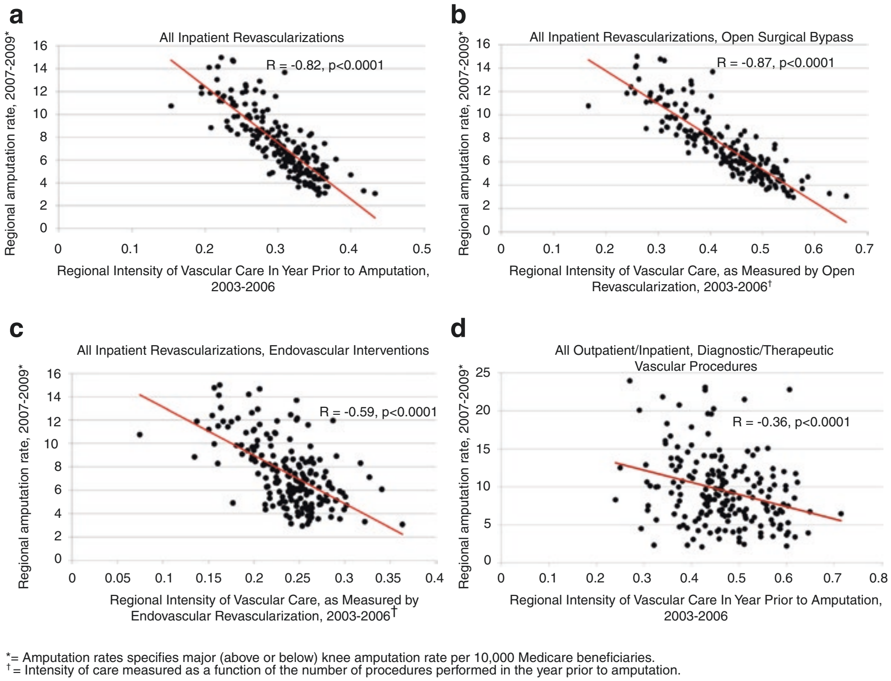


Fig. 3.10 Relationships between regional intensity of vascular care and amputation rate, for (**a**) all inpatient revascularizations, (**b**) open surgical bypass, (**c**) endovascular interventions, and (**d**) all outpatient/inpatient procedures. (Reproduced with permission from Elseivier, Goodney et al, Journal of Vascular Surgery 2013, in press.)

The United States Food and Drug Administration has recently made an effort to do this. Realizing the value of "Real World Evidence" – data which is collected about procedures and devices outside of clinical trials, therefore representing results that can be expected in clinical practice – the FDA has made studies using Real World Evidence a priority. This type of data is precisely the kind of data used in the studies described above by Wennberg, Birkmeyer, and Goodney, highlighting the importance of this type of work.

Such "Real World Evidence" can be used in important ways. Ibrahim and Dimick in 2018 demonstrated the failing safety of gastric bands used in bariatric surgery for weight loss [14]. They noted a transition point in annual payments for gastric bands, a point where more than 50% of annual payments for gastric bands were allocated to reoperations for failing devices instead of new placement procedures (Fig. 3.11). Interestingly, they found that payment data clearly documented failing devices years prior to gastric bands falling out of favor in clinical practice. Their findings highlight the importance of using available evidence to monitor device performance.

A more recent advance in the collection of such data has been combining two different – but complementary – types of data. The studies mentioned above have primarily focused on administrative billing claims data. These powerful data sources, most commonly Medicare, allow for long-term outcome assessment

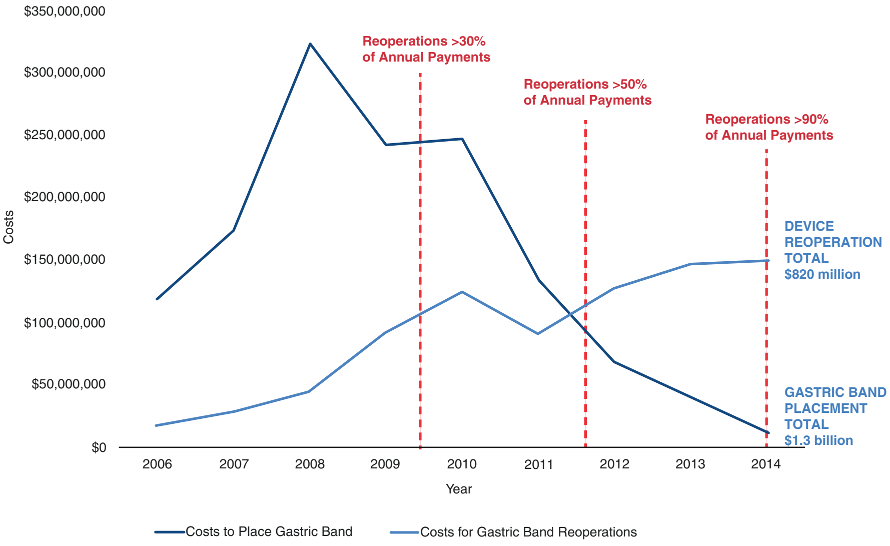


Fig. 3.11 Annual procedures for gastric band placement and device reoperations. (Reproduced with permission from the American Medical Association, Ibrahim et al, JAMA 2018.)

through insurance claims, but have little clinical granularity. Linking administrative claims directly to more granular clinical registry data creates a more thorough understanding of device performance [15, 16]. Leveraging granular clinical data along with long-term outcome assessment after device implantation is currently a burgeoning field that promises to improve the assessment of implantable devices for both patients and policy makers [17, 18].

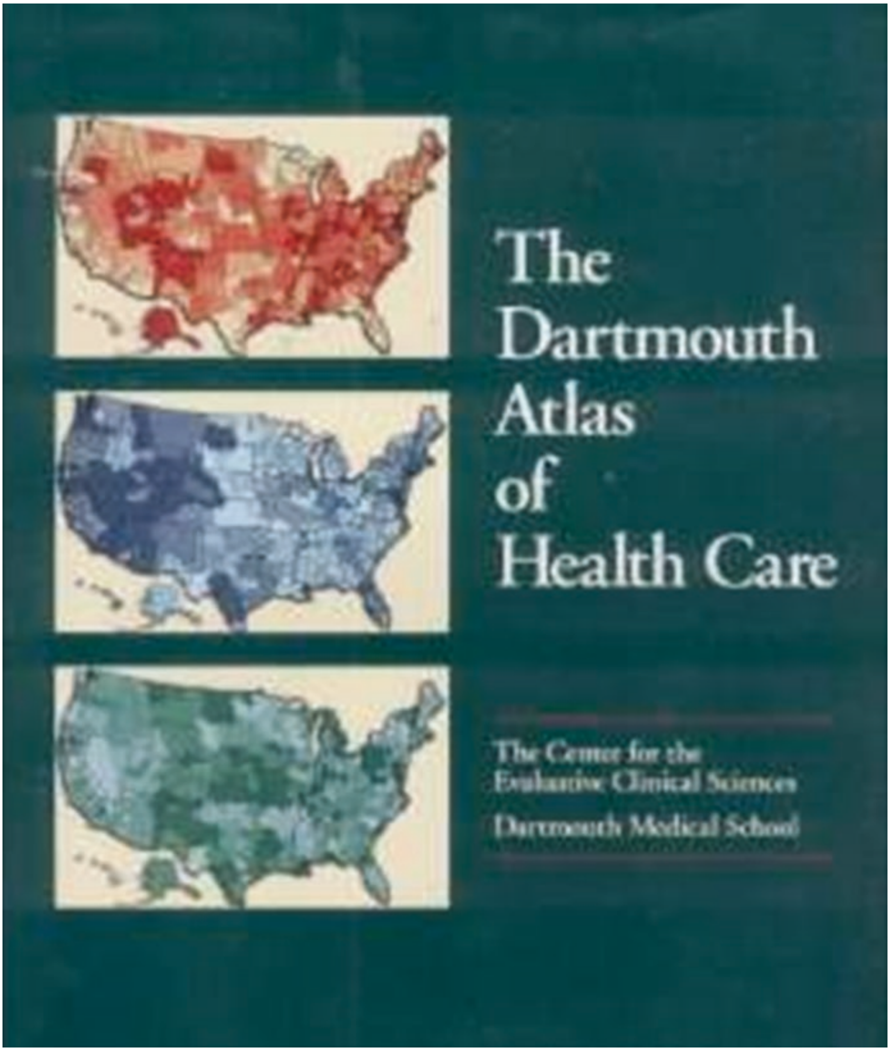
3.8 A Unique Tool in Implementing Change: The Dartmouth Atlas

The description of variations in healthcare began more than 30 years ago, and evidence surrounding these scope and impact of these variations has been building for more than three decades. Publications in major journals, attention from task forces and leaders in health policy, and lengthy consensus opinions have all stated that variation in health care delivery, especially surgical care, is not helpful, and potentially harmful. So why, then, has this trend continued?

Undoubtedly, translating this evidence to effective health policy has been challenging. Measuring variation is difficult, as is defining the implications of variation. The data involved can make an accountant’s head spin, and often the clearest messages only emerge after careful study of reams and reams of data.

To help increase the visibility of the difficulties surrounding variation in the provision of health care, and to help those in health policy grasp its true impact, Dr. Wennberg conceived an “atlas” of health care that would graphically convey this message to broad audiences. First published in 1996, the original edition of the Dartmouth Atlas of HealthCare used maps, charts, and tables to illustrate the relationships between geography, variation, and healthcare (Fig. 3.12) [19].

These compendiums, rather than simply aimed at medical audiences, were written for a broader appeal. Wennberg felt that health policy had to be understood not just by physicians, but by patients, payers, and policymakers to have the greatest impact. And, given the success of the original edition, several subsequent Atlases and reports have followed. These reports have garnered attention from leaders in



The Dartmouth Atlas of Health Care

The Center for the Evaluative Clinical Sciences

Dartmouth Medical School

Fig. 3.12 The Dartmouth Atlas of healthcare. (Reproduced with permission, personal communication, John Wennberg, AHA Publishing, Inc.)

health policy both nationally and internationally, and have served as a blueprint for health care reform aimed at limiting variation and unnecessary spending in the US health care system.

References

1. Bunker JP. Surgical manpower. A comparison of operations and surgeons in the United States and in England and Wales. N Engl J Med. 1970;282(3):135–44.
2. Wennberg J, Gittelsohn. Small area variations in health care delivery. Science. 1973;182(4117):1102–8.
3. Birkmeyer JD, Sharp SM, Finlayson SR, Fisher ES, Wennberg JE. Variation profiles of common surgical procedures. Surgery. 1998;124(5):917–23.
4. Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. The implications of regional variations in Medicare spending. Part 1: The content, quality, and accessibility of care. Ann Intern Med. 2003;138(4):273–87.
5. Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. The implications of regional variations in Medicare spending. Part 2: Health outcomes and satisfaction with care. Ann Intern Med. 2003;138(4):288–98.
6. Kwok AC, Semel ME, Lipsitz SR, Bader AM, Barnato AE, Gawande AA, et al. The intensity and variation of surgical care at the end of life: a retrospective cohort study. Lancet. 2011;378(9800):1408–13.
7. Weinstein JN, Lurie JD, Olson PR, Bronner KK, Fisher ES. United States' trends and regional variations in lumbar spine surgery: 1992–2003. Spine (Phila Pa 1976). 2006;31(23):2707–14.
8. Weinstein JN, Lurie JD, Tosteson TD, Tosteson AN, Blood EA, Abdu WA, et al. Surgical versus nonoperative treatment for lumbar disc herniation: four-year results for the spine patient outcomes research trial (SPORT). Spine (Phila Pa 1976). 2008;33(25):2789–800.
9. Goodney PP, Beck AW, Nagle J, Welch HG, Zwolak RM. National trends in lower extremity bypass surgery, endovascular interventions, and major amputations. J Vasc Surg. 2009;50(1):54–60.
10. Scali ST, Goodney PP, Walsh DB, Travis LL, Nolan BW, Goodman DC, et al. National trends and regional variation of open and endovascular repair of thoracic and thoracoabdominal aneurysms in contemporary practice. J Vasc Surg. 2011;53(6):1499–505.
11. Goodney PP, Travis LL, Malenka D, Bronner KK, Lucas FL, Cronenwett JL, Goodman DC, Fisher ES. Circ Cardiovasc Qual Outcomes. 2010;3(1):15–24. https://doi.org/10.1161/CIRCOUTCOMES.109.864736. Epub 2009 Dec 8. PMID: 20123667
12. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. N Engl J Med. 2009;361(14):1368–75.
13. Goodney PP, Holman K, Henke PK, Travis LL, Dimick JB, Stukel TA, Fisher ES, Birkmeyer JD. J Vasc Surg. 2013;57(6):1471–79, 1480.e1-3; discussion 1479–80. https://doi.org/10.1016/j.jvs.2012.11.068. Epub 2013 Feb 1. PMID: 23375611
14. Ibrahim AM, Dimick JB. Monitoring medical devices: missed warning signs within existing data. JAMA. 2017;318(4):327–8.
15. Hoel AW, Faerber AE, Moore KO, Ramkumar N, Brooke BS, Scali ST, et al. A pilot study for long-term outcome assessment after aortic aneurysm repair using vascular quality initiative data matched to Medicare claims. J Vasc Surg. 2017;66(3):751–9 e1.
16. Sedrakyan A, Campbell B, Merino JG, Kuntz R, Hirst A, McCulloch P. IDEAL-D: a rational framework for evaluating and regulating the use of medical devices. BMJ. 2016;353:i2372.
17. Columbo JA, Kang R, Hoel AW, Kang J, Leinweber KA, Tauber KS, et al. A comparison of reintervention rates after endovascular aneurysm repair between the vascular quality initiative registry, Medicare claims, and chart review. J Vasc Surg. 2019;69(1):74–9.e6. https://doi.org/10.1016/j.jvs.2018.03.42.

18. Columbo JA, Martinez-Camblor P, MacKenzie TA, Staiger DO, Kang R, Goodney PP, et al. Comparing long-term mortality after carotid Endarterectomy vs carotid stenting using a novel instrumental variable method for risk adjustment in observational time-to-event data. JAMA Netw Open. 2018;1(5):e181676.
19. Dartmouth Atlas of HealthcareOctober 1st, 2007: [October 1st, 2007 pp.]. Available from: https://www.dartmouthatlas.org.

Further Reading

Birkmeyer JD, Sharp SM, Finlayson SR, Fisher ES, Wennberg JE. Variation profiles of common surgical procedures. Surgery. 1998;124:917–23. *This paper was the first to demonstrate that small area analysis of variations in surgical practice was feasible on a national scale, and introduced the study of regional variation to surgeons*

Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. The implications of regional variations in medicare spending. Part 1: the content, quality, and accessibility of care. Ann Intern Med. 2003;138:273–87.

Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. The implications of regional variations in medicare spending. Part 2: Health outcomes and satisfaction with care. Ann Intern Med. 2003;138:288–98. *This two-part manuscript is among the most commonly referenced large studies of variation in Medicare spending and its effect on outcomes*

Wennberg J, Gittelsohn. Small area variations in health care delivery. Science. 1973;182:1102–8. *This was the first prominent publication demonstrating that small area analysis could identify important variations in the use and outcomes of medical care.*

Health Policy Research in Surgery

4

Karan R. Chhabra and Justin B. Dimick

Abstract

The purpose of this chapter is to provide an overview of health policy research in surgery. We will begin by considering common pitfalls in conducting health policy research. We will then provide research examples from two key areas: (1) Physician and hospital payment reform and (2) Surgical training and workforce policy. For each area of research, recent and impending policy changes will be discussed; examples of studies that have answered important questions provided; and important research questions that are not yet answered will be highlighted. Finally, we will close with a brief discussion of the research tools necessary to generate the right answers and where to find collaborators for those interested in pursuing research in this field.

4.1 Introduction

Despite being largely invisible to most practicing surgeons, health policy shapes every detail of the context in which we work, including (1) how we are paid, (2) how we are trained, and (3) whether we are incentivized for volume or value.

K. R. Chhabra, MD, MSc (✉)
National Clinician Scholars Program at the Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI, USA

Center for Healthcare Outcomes and Policy, University of Michigan, Ann Arbor, MI, USA

Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA
e-mail: kchhabra@bwh.harvard.edu

J. B. Dimick, MD, MPH
Department of Surgery, University of Michigan Health System, Ann Arbor, MI, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_4

However, despite shaping our environment, health policy is hard to see. Surgical researchers are often drawn to topics that have immediate relevance to their practice. Surgeons are therefore naturally drawn to research that compares the effectiveness of different approaches to managing disease. After all, most of us chose surgery over other specialties because of how tightly linked outcomes are to our interventions. Surgeons thrive on immediacy.

Nonetheless, the decisions made by policymakers create the reality in which we live, however remote these decisions are from our daily practice. Moreover, policymakers often make such decisions without good evidence. There is very little research on health policy in surgery and consequently, very little "evidence-based policymaking". But many important policies uniquely target surgeons and their patients, such as Medicare coverage decisions, payment reforms (readmissions penalties and bundled payments), and duty hour restrictions. We need surgeons focused on evaluating the effectiveness of health policy to better inform these decisions.

The purpose of this chapter is to provide an overview of health policy research in surgery. We will begin by considering common pitfalls in conducting health policy research. We will then provide research examples from two key areas: (1) Physician and hospital payment reform and (2) Surgical training and workforce policy. For each area of research, recent and impending policy changes will be discussed; examples of studies that have answered important questions provided; and important research questions that are not yet answered are identified. Finally, we will close with a brief discussion of the research tools necessary to generate the right answers and where to find collaborators for those interested in pursuing research in this field.

4.1.1 Common Pitfalls in Health Policy Research

There are several common pitfalls to be aware of when conducting health policy research. Because it is difficult (or nearly impossible) to randomize hospitals to different policy options, health care delivery system research often uses observational studies and so-called "natural experiments".

Ignoring the "Counterfactual". One important flaw of many policy evaluation studies is to fail to consider what economists call the counterfactual—what would have happened had the policy not been implemented. It is tempting to simply compare outcomes before vs. after implementation of a policy. Such "pre-post" studies may incorrectly attribute a significant improvement in outcomes to the policy when outcomes would have improved without the policy (i.e., if there was a trend towards improved outcomes that was not adjusted for). Techniques for overcoming this pitfall include differences-in-differences (DID) and interrupted time-series analysis (ITSA), both discussed below. If you are not familiar with these methods, it is extremely important to consult an economist (or econometrician) when performing these complex regression analyses.

Not Considering Unintended Consequences. Another opportunity for surgical health services researchers is to evaluate policy changes' unintended consequences. These can take the form of positive or negative "spillover" effects to other clinical areas. Positive spillover effects would be when a policy aimed at improving care in one clinical area (or one outcome) results in improvements for other areas (or outcomes). For example, policies aimed at reducing readmissions for vascular surgery could also result in fewer readmissions for other procedures (e.g., general surgery) as these patients are often cared for by the same nurses and on the same floors. Negative spillover effects, often called "multi-tasking" in the economics literature, result when resources are shifted to a targeted condition and care deteriorates in another clinical area. For example, policies aimed at reducing readmissions for vascular surgery could inadvertently increase readmissions for other surgical conditions if resources are taken from other surgical patients and care is improved only for the patients targeted by the policy. As clinicians on the front lines of patient care, surgeon-researchers are uniquely well-positioned to evaluate for potential spillover effects and other potential unintended consequences.

4.1.2 Key Research Methods in Health Policy

As mentioned above, one of the key challenges in policy research is drawing causal links between policy and changes in outcomes. Without randomization, it is often difficult to know the counterfactual change in outcomes (i.e. if a policy were not implemented).

Differences in Differences. The best way to avoid this pitfall is to include a control group of patients (or hospitals) that is unaffected by the policy. The control group represents the counterfactual and can be used to account for background trends. This is known as a "differences in differences" (DID) design. In DID analyses, the first difference is the change in the group exposed to the policy. The second difference is the change in the control group, i.e. the one not exposed to the policy. The change in the control group is subtracted from the change in the exposure group to achieve the DID estimate. Conceptually, this can be expressed as follows:

$$\text{DID} = \left(\text{post}_{\text{exposed}} - \text{pre}_{\text{exposed}}\right)\_ - \left(\text{post}_{\text{control}} - \text{pre}_{\text{control}}\right).$$

In reality, health policy research uses regression models rather than simple subtraction to calculate DID estimates, as models allow us to perform risk-adjustment and compute statistical significance.

One key assumption for this method is that the trends are parallel prior to the implementation of the policy, which should be directly tested prior to using this approach. If the trends are not parallel, then a DID approach will not yield the correct inference. Another important assumption is that of "common shocks": that is, any unpredictable events (outside of the policy) would affect both groups equally. Finally, one must consider spillover effects; if the policy also spilled over and influenced the control group, then DID analysis may be inappropriate. These limitations show the challenge of finding an ideal control group to allow DID analysis.

Interrupted Time Series Analysis. An alternative approach is needed when there is no suitable control group. In this setting, an interrupted time-series model (ITSA) is often useful. ITSAs generally model the *slope*, or the rate of change, of an outcome such as mortality or readmissions rates. They can also measure whether the *intercept*, or absolute level of an outcome, abruptly changes after an intervention. These models treat the slope before the intervention as the counterfactual trend, assuming that if the intervention had not taken place, the same trend would have continued. ITSA models can also incorporate a control group. A positive finding is when there is a statistically significant difference in slopes between the pre- and post- periods, and/or between the exposure and control group.

A table outlining the differences between ITSA and DID is below (Table 4.1).

4.1.3 Physician and Hospital Payment Reform

For young surgeons entering practice right now, rising health care costs and consequent reform to "bend the cost curve" will be the single driving policy force of their professional lives. For decades, rising health care expenditures have been called unsustainable, and it appears that they are finally living up to that reality.

The Affordable Care Act (ACA) brought about several changes that fundamentally changed how physicians and hospitals are paid by Medicare as well as major insurers. Most of these changes aim to move from a volume based payment system to one that pays for value. One of the dominant features of these reforms is to shift financial risk to providers. Such "at risk" payment models fundamentally change the incentives for improving quality of care. In traditional payment models, the payer bears most of the financial risk of complications. Payers would foot the bill for complications and the subsequent health care, including prolonged length of stay, physician consultations, home health care, and skilled nursing care after discharge.

The Center for Medicare and Medicaid Services (CMS), the US's largest healthcare payer, has several candidate policies that aim to improve quality and reduce costs in surgery, including selective referral, pay-for-performance, non-payment for adverse events (e.g., hospital acquired infections and readmissions), episode bundled payments, and accountable care organizations (Table 4.2). It is essential that researchers thoroughly evaluate the benefits and harms of these policy changes. Without such research, policymakers will not know what works, and what doesn't, as we move forward with future iterations of payment reform.

Below we include several examples from the literature that evaluate health care policies. These illustrate many of the concepts discussed above, including how these studies addressed common pitfalls in health policy research.

Example 1 *Bariatric surgery complications before vs after implementation of a national policy restricting coverage to centers of excellence. Dimick JB, Nicholas LH, Ryan AM, Thumma JR, Birkmeyer JD. JAMA 2013;309:792–799.*

This study from our research group evaluated the impact of the CMS national coverage decision for bariatric surgery, which was the most ambitious selective referral program in surgery to date. In 2006, CMS limited coverage of bariatric

Table 4.1 Key differences between differences-in-differences and interrupted time-series analysis

	Differences-in-differences (DID)	Interrupted time-series analysis (ITSA)
Counterfactual (i.e., what would happen without policy)	Pre-post difference in control group	Slope of outcome prior to intervention
Key assumptions	1. Parallel trends between exposure and control groups in pre- period 2. Common shocks after policy implementation 3. No major spillovers between exposure and control groups	1. Pre-intervention trend would continue if policy had not occurred 2. Intervention introduced at a single time point 3. Population characteristics do not change over time
Advantages	Highly rigorous when a control group is available Can combine other techniques for minimizing confounding (e.g. propensity score matching)	Easy to model more than 2 time periods (e.g. can incorporate an "implementation" period) Control group is possible but not required
Disadvantages	Requires a control group that satisfies assumptions	Results highly affected by choice of time points Seasonal fluctuations may require complex adjustments

Table 4.2 Center for Medicare and Medicaid Services (*CMS*) Policies aimed at improving quality and reducing costs in surgery

Policy option	CMS adoption
Selective referral	National coverage decisions for selected procedures
Refer patients to specific providers, i.e., "centers of excellence"	*Example*: Bariatric surgery coverage linked to "center of excellence status"
Non-payment for adverse events	National programs already in place
Incentivize quality improvement by withholding payment for certain adverse outcomes	*Examples*: In October 2008, CMS discontinued additional payments for certain hospital-acquired conditions that were deemed preventable
Pay for performance	Multiple large pilot programs
Reward providers for high quality or low cost care	*Example*: Hospital readmissions reduction program (HRRP)
Bundled payment	Regional and national pilot programs
Incentivize efficient, coordinated care by bundling payments around an episode	*Examples*: Comprehensive Care for Joint Replacement (CJR), bundled payments for care initiative (BPCI)
Accountable care organizations	Pilot programs
Health care providers accept risk for reducing health care expenditure growth for population of Medicare beneficiaries	*Examples*: Pioneer accountable care organization (ACO) demonstration program; Medicare shared savings accountable care organization (ACO) program Multiple commercial insurance pilots

surgery to so-called centers of excellence (COEs) as defined by the American College of Surgeons (ACS) and American Society for Metabolic and Bariatric Surgery (ASMBS). Prior studies evaluating the program had shown benefits, with reductions in morbidity and mortality. However, these studies had failed to adequately account for pre-existing trends towards improved outcomes in bariatric surgery. In our study, a control group of non-Medicare patients undergoing bariatric surgery was used to adequately account for these trends. In this differences-in-differences analysis (discussed in detail above), there was no independent effect of the CMS policy on overall complications, serious complications, or reoperations. This study demonstrates the importance of adequately adjusting for pre-existing time trends. Without such an adjustment, policymakers would mistakenly attribute the improved outcomes to the policy.

Key Unanswered Questions Further research needs to demonstrate the extent to which this policy limited access for Medicare beneficiaries in need of bariatric surgery. It is possible that Medicare patients had to travel further for surgery. Vulnerable populations may experience a decline in availability of surgery if they could not afford to travel away from their homes. Since the policy had no measurable benefit, research demonstrating such harms should strongly motivate CMS to reconsider this policy.

Example 2 *Changes in hospital quality associated with hospital value-based purchasing. Ryan AM, Krinsky S, Maurer KA, Dimick JB. N Engl J Med 2017;376:2358–66.*

This study evaluated the impact of Medicare's largest pay-for-performance program, the Hospital Value-Based Purchasing (HVBP) program, on patient outcomes. HVBP adjusts hospitals' Medicare payments based on a complex formula incorporating both improvement and maintenance of an array of patient experience, process, and outcome measures. Prior studies at 9 and 30 months after program implementation had demonstrated null effects, but its longer-term impact on risk-adjusted outcomes had not been explored. We evaluated patient experience, process measures, and mortality at acute care hospitals subject to HVBP. Since most traditional inpatient hospitals were subject to this program, this study used an unconventional control group: critical access hospitals, which are small rural facilities exempt from many Medicare payment reforms. Though these hospitals are considerably different from HVBP-participating hospitals, this study used matching techniques to compare hospitals with similar trends in each study outcome prior to implementation of HVBP. This study also used a differences-in-differences design, enabled by the development of a control group, to ensure that temporal differences in outcomes were taken into account. It found no consistent improvement in outcomes with the implementation of HVBP.

Key Unanswered Questions The important questions around pay-for-performance include whether programs with larger incentives will have an impact on outcomes.

This study showed relatively small, yet complex payment incentives like HVBP may not have important effects on clinical quality. Analysis of larger, simpler incentives such as the Hospital Readmissions Reduction program has shown more profound effects. With programs that penalize hospitals for poor outcomes, it will be important to conduct studies to understand whether such policies improve or exacerbate racial and socioeconomic disparities in surgical outcomes. New physician-focused pay-for-performance programs have also recently been introduced, including the Medicare Access and CHIP Re-Authorization Act (MACRA) and the Merit-Based Incentive Payment System (MIPS) in 2015. Surgeon-researchers have an opportunity to study their impact on surgical quality and costs.

4.1.4 Surgical Training and Workforce Policy

Policy around surgical training, including duty hour restrictions, has dramatically changed how we train surgeons in the United States. The motivation for this policy change, beginning with the 80-hour workweek, was the perception that longer days and weeks lead to more fatigued surgical trainees who make more errors that threaten patient safety. This was followed by further reforms in 2011 that, among other changes, restricted interns to 16-hour in-house shifts. But relatively few studies have addressed the key question: *Did this policy have the intended consequences of improving patient safety*? We will discuss an example of a paper that examined this question below. However, it is also important to also ask whether the policy had any unintended consequences. For this particular policy, unintended consequences include the potential to make patient safety worse, by increasing hand-offs, or by impacting surgical education in way that makes surgical trainees less prepared for independent clinical practice.

Example 3 *Association of the 2011 ACGME Resident Duty Hour Reform With General Surgery Patient Outcomes and With Resident Examination Performance. Rajaram R, Chung JW, Jones AT, Cohen ME, Dahlke AR, Ko CY, Tarpley JL, Lewis FR, Hoyt DB, Bilimoria KY.* JAMA *2014;312:2374–84.*

To assess whether the implementation of the 80-hour workweek was associated with improved patient safety, Rajaram and colleagues conducted this large quasi-experimental study using the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database. These studies compared mortality rates before vs. after implementation of the 2011 duty hour restrictions. They found no changes in 30-day morbidity or mortality after implementation of the policy. Rather than simply using a pre-post design, comparing outcomes before vs. after the policy was implemented, this study employed a differences-in-differences approach that used the percentage of operations involving a resident as an elegant proxy for hospital teaching status. Specifically, they deemed hospitals with ≥95% of cases involving a resident to be “teaching,” and those with ≤1% of cases involving a resident “nonteaching.” Since the policy should have a larger impact on teaching hospitals, they were able to evaluate the impact of the policy on safety in a controlled fashion, with nonteaching hospitals representing the counterfactual trend in surgical outcomes.

Key Questions Left Unanswered This study was important and widely cited but did not definitively answer the question of whether the 2011 duty hour restrictions impacted the quality of surgical training. However, its findings were echoed by a cluster-randomized study called the Flexibility In Duty Hour Requirements for Surgical Trainees (FIRST) Trial, which randomly assigned residency programs to the 2011 restrictions vs. more flexible duty hours. The FIRST Trial similarly found no improvement in mortality at programs subject to the 2011 restrictions, and thus led to their reversal. This is an example of high-quality research using secondary data with results mirroring that of a far more resource-intensive policy randomized trial—both of which led to major policy change.

Another key health policy issue is the adequacy of the surgical workforce. There is active, passionate debate about whether there is an impending shortage of general surgeons. Many believe the aging baby boomer population and the increased need for surgery will lead to a shortage of surgeons. This conclusion is logical. However, there is another school of thought that thinks we already have too many surgeons and this leads to overutilization of surgical procedures, especially discretionary procedures. This theory of "supply sensitive care" has been popularized by the Dartmouth Atlas of Healthcare, and argues that we should not be concerned about a 10% shortage of surgeons when we have more than two-fold (100% differences) variations across regions of the United States. It is likely there are an oversupply in certain regions and an undersupply in others.

These diverging schools of thought are more than theoretical musing. The practical implications of fixing the problem using available policy levers bring about an important issue. If we try and fix the projected shortage by increasing the number of surgeons we train, it is very likely that we will exacerbate the distribution problems rather than alleviate the shortage-i.e., the surgeons we train will no doubt choose to live in the most desirable areas and the regions that are currently underserved will continue to be that way. In other words, increasing the overall supply of surgeons is a very blunt tool for fixing a shortage that only exists in a few regions. Once again, there have been numerous studies written that use various models to make workforce predictions in surgery. But very few have assessed the key question: *How many surgeons are actually needed to provide adequate care within a region*? Below we review one of the few studies that address this important question, albeit from an indirect perspective.

Example 4 *Perforated appendicitis among rural and urban patients: implications of access to care. Paquette IM, Zuckerman R, Finlayson SR. Ann Surg 2011;253:534–538.*

This study evaluated rates of perforated appendicitis in rural as compared to urban areas. Perforation was used as a proxy for delayed access to care in rural areas. Paquette and colleagues used the Nationwide Inpatient Sample (NIS) to compare rates of perforation across regions with different population density. They found that patients living in rural areas were more likely than those living in urban areas to present with perforated appendicitis (36% vs. 31%). Although this is indirect evidence, it suggests there may not be enough general surgeons in rural regions.

Key Question Left Unanswered There is very little research to guide evidence-based policymaking in decisions about surgical workforce. Future research needs to focus on better understanding how many surgeons are needed to provide access to all necessary care in a region. The study by Finlayson and colleagues is a good start but this work needs to be extended to a broader range of clinical conditions. In addition, research aimed at evaluating the impact of policies for increasing the supply of surgeons in underserved regions should be pursued. For example, it is unclear if incentives such as loan forgiveness to work in these areas translate into a long-term increase in the supply of surgeons in these areas.

4.2 What Research Tools Are Needed for Health Policy Research?

One key challenge to evaluating health policy is that it is nearly impossible to conduct true randomized experiments. It is very difficult to randomize hospitals and physicians to different payment structures. As discussed above, it is necessary to draw inferences about the effectiveness of policy changes from quasi-experimental or observational studies. *Economics and econometrics (this is what economists call statistics) provides robust methodological tools for designing these studies*. Longitudinal study designs, including differences-in-differences, and other panel data approaches are widely used in policy evaluation. *These build on standard linear and logistic regression models that can be learned in any basic statistics class*. Many of these econometric models (e.g., differences-in-differences) can be implemented in multivariate regression models as a simple interaction term. For example, in a study evaluating outcomes pre vs. post in an exposed and non-exposed group of hospitals, the difference-in-difference can be estimated as the interaction term post∗exposed (where post = 1 after implementation and exposed = 1 if the policy targeted the hospital where the patient had surgery).

4.3 Where Can I Find Collaborators for Health Policy Research?

For policy evaluation, the most natural collaborators are often health economists, which you may find in the medical school health services research department, a health management and policy department in a school of public health, or in the undergraduate department of economics. Collaborators from political science, policy analysis, and health policy will provide the policy context, advice about looking for the impact of unintended consequences, and will often also have quantitative skills in econometrics.

4.4 Where Should I Get Started in Health Policy Research?

The best place to start is to keep up with health policy. The *New York Times*, *Wall Street Journal*, and *Washington Post* provide robust coverage of health policy, and importantly, its impact on patients. If you want information early you may have to visit the *Federal Register* or directly read the legislation to find the important details of a policy.

The most relevant journals include *Health Affairs*, *NEJM*, and *JAMA*. *Health Affairs* is entirely focused on health policy and innovation in health care delivery, and has a health policy blog that is a good source of information. *NEJM* and *JAMA* often publish policy-relevant articles, including editorials (Perspectives in *NEJM* and Viewpoints in *JAMA*). *NEJM* also has a special section of its website dedicated to health policy and reform. Other journals that demonstrate rigorous methodology for policy evaluation but tend to have a broader focus include *Health Services Research* and *Medical Care*.

As discussed in other chapters, the key to developing a career in any research discipline is to find a mentor. Mentors who do health policy research in surgery are rare, and you may find greater success looking for non-surgeons (maybe even non-physicians).

4.5 Conclusion

Health policy is often invisible but shapes every detail of how we work. Because of unsustainable growth in health care expenditures, the pace of policy change is accelerating. The careers of surgeons training today will be characterized by constant innovation in our delivery system, particularly in how we get paid by Medicare and private payers. Sophisticated research on the effectiveness of these policy changes is needed to help policymakers make evidence based decisions. Despite the importance of this area of research, there are very few surgeons involved in rigorous policy evaluation, which provides a great opportunity for young surgeons to fill this void.

Further Reading

Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017;46(1):348–55.

Dimick JB, Nicholas LH, Ryan AM, Thumma JR, Birkmeyer JD. Bariatric surgery complications before vs. after implementation of a national policy restricting coverage to centers of excellence. JAMA. 2013;309:792–9.

Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. JAMA. 2014;312:2401–2.

Paquette IM, Zuckerman R, Finlayson SR. Perforated appendicitis among rural and urban patients: implications of access to care. Ann Surg. 2011;253:534–8.

Ryan AM, Krinsky S, Maurer KA, Dimick JB. Changes in hospital quality associated with hospital value-based purchasing. N Engl J Med. 2017;376:2358–66.

Rajaram R, Chung JW, Jones AT, Cohen ME, Dahlke AR, Ko CY, Tarpley JL, Lewis FR, Hoyt DB, Bilimoria KY. Association of the 2011 ACGME resident duty hour reform with general surgery patient outcomes and with resident examination performance. JAMA. 2014;312:2374–84.

Surgical Disparities: Beyond Non-Modifiable Patient Factors

5

Numa P. Perez, Claire A. Pernat, and David C. Chang

Abstract

Over the past several years, there has been an explosion of publications describing disparities in virtually all aspects of our healthcare system. Surgery is no exception, with recent studies reporting inequities in surgical care and outcomes related to race, gender, age, socioeconomic status, education, and geographic location. Progress, however, has remained slow, likely due to an inadequate understanding of the mechanisms that lead to these disparities. To effectively address health care disparities, it is imperative to differentiate between modifiable and non-modifiable factors and focus primarily on addressing the former. Since most people who will be reading this chapter are current or future surgeons, we would strongly recommend you focus on identifying modifiable causes of disparities that affect your patient population, followed by devising and testing innovative approaches that can reduce them in your practice and the healthcare system as a whole.

5.1 Introduction

Health care disparities are a popular and important topic of research, discussion, and policy implementation in all medical fields—including surgical care. Before we dive into our discussion about disparities, a few basic definitions are necessary to provide context for the reader. First, health disparities are differences in "health

N. P. Perez · D. C. Chang (✉)
Department of Surgery, Massachusetts General Hospital, Boston, MA, USA
e-mail: dchang8@mgh.harvard.edu

C. A. Pernat
Department of Surgery, Massachusetts General Hospital, Boston, MA, USA

Tufts University, Medford, MA, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_5

status that result from social disadvantage that is itself associated with characteristics such as race or ethnicity and socioeconomic status (SES)". On the other hand, health *care* disparities are differences in the "quality of health care experienced by those with social disadvantage" [1]. Lastly, health *equity* is "the attainment of the highest level of health for all people…with focused and ongoing societal efforts to address avoidable inequalities, historical and contemporary injustices, and the elimination of health and health care disparities" [2]. There exists a large body of literature documenting differences in treatment and clinical decisions related to race, sex, insurance, SES, and geographic setting/location, which lead to disparate access to and quality of surgical care, and ultimate disparate outcomes [3–7]. The focus of this chapter is not to be comprehensive but rather to highlight essential gaps in our current approach.

5.1.1 Brief History

Interest in surgical disparities is not a novel phenomenon. In fact, an article published in the *New England Journal of Medicine* in 1977 reported black patients at Johns Hopkins were four times more likely than white patients to be operated on by residents without appropriate supervision, while white patients undergoing similar operations were more often cared for by attending surgeons [8]. Some of the first policy initiatives in this field date as far back as 1985, when the seminal "Report of the Secretary's Task Force on Black & Minority Health" (a.k.a. Heckler report) was published, giving rise to the creation of the U.S. Department of Health and Human Services Office of Minority Health. In 1990, a similar change took place at the National Institutes of Health, leading to the establishment of the Office of Minority Programs. This was followed in 1998 by the Clinton administration's report titled "One America in the 21st Century: The President's Initiative on Race." The report explicitly targeted the elimination of health care disparities as a key goal [9]. This catapulted the passage of the Minority Health and Health Disparities Research and Education Act in 2000, leading to the creation of the National Center on Minority Health and Health Disparities, which later became the National Institute of Minority Health and Disparities (NIMHD). More recently in 2015, the NIMHD and the American College of Surgeons (ACS) hosted a "Symposium on Disparities in Surgical Care and Outcomes" that created a specific research agenda for addressing disparities in surgical care, outcomes, and treatment of our patients [10].

5.2 Concept of Modifiable Risk Factors

Despite considerable advances over the past decade in highlighting health care disparities, progress in creating actionable solutions remains slow. Britton et al. found that less than a quarter of surgeons take efforts to analyze, investigate, and address disparities within their own practices [11]. A primary reason for this is that much of surgical disparities research has focused on non-modifiable factors, such as patient age, gender, race, insurance, and geographic location (i.e. health disparities), while disregarding how providers and healthcare systems interact with such factors. This perspective, unfortunately, engenders a sense of futility. In fact, if we are thoughtful about the way we approach health care disparities, we are likely to realize non-modifiable and modifiable factors are generally two sides of the same coin, the main difference being to whom the burden of responsibility is assigned. For example, an access to care issue can be easily reframed as a system allocation problem; patient health literacy limitations can be rethought as deficiencies in the ability of providers to create rapport and educate their patients. Hence, it would be more productive if researchers focused their efforts on modifiable factors such as provider practice patterns and system-level barriers (i.e. health care disparities) that affect our patients in the context of the aforementioned non-modifiable features. This framework is more likely to lead to actionable solutions that can generate broad changes in practice, which can ultimately contribute to decreasing the disparity gap. Hereon, we will divide and examine modifiable factors into two main categories: factors related to access to care, and factors related to the quality of care.

5.2.1 Access to Care Factors

Inequitable access to care is a crucial, and perhaps the most important, determinant of health care disparities. Access is “the timely use of personal health services to achieve the best possible outcomes” [12]. In fact, several studies have demonstrated that under disparate access, advances in surveillance and treatment of diseases such as breast and colon cancer can actually widen the disparity gap, by improving outcomes preferentially for white and higher SES patients [13]. On the other hand, when access to care is equal, disparities can be reduced or eliminated entirely [14]. For example, Lee et al. evaluated pediatric patients in an integrated health system that provides equal access to all its patients and found no significant difference by race or income in appendiceal perforation rates [15]. This finding has been replicated in numerous further studies, as described in a review article published in 2013, which examined surgical outcomes disparities and found that when there is equal access to care across minority populations, the disparity gap decreases [3]. In fact, the ACS further highlighted this in a statement released in 2013, which affirms that “[o]ptimal access is the key to quality of care. Efforts to increase surgical presence and availability are crucial to providing the right care, at the right time, in the right place. Optimal quality, the centerpiece of the mission of the American College of Surgeons, is not achievable without optimal access” [16].

In keeping with the philosophy of focusing on modifiable risk factors, we believe it is important to change the way we think about access disparities, not as a “resource scarcity” problem, but instead as a “resource misallocation” problem.

5.2.1.1 Race-Based Disparities in Access to Care

Race has been one of the most studied factors as it pertains to access to care, but conflicting data exist on whether race-based disparities are due to race-alone or whether they reflect inherent differences in socioeconomic factors such as income and insurance coverage. For example, Taioli and Flores used the National Cancer Institute Surveillance, Epidemiology, and End Results (SEER) database to study the likelihood of undergoing surgical resection for lung cancer and found black patients were 30% less likely than their white counterparts to undergo surgery, despite adjusting for age and insurance [17]. Similarly, Wilson et al. found black Medicare recipients are less likely to undergo elective repair of abdominal aortic aneurysms (AAA) when compared to whites but are conversely more likely to undergo an urgent repair [18]. In 2015, Wang et al. used the Pediatric Health Information System (PHIS), a database of 43 tertiary children's hospitals, and found black and lower income children were significantly less likely to undergo abdominal imaging when presenting to the ED with abdominal pain and were conversely more likely to have perforated appendicitis [19, 20]. Moreover, the combined effect of race and income was greater than either of them alone.

We would argue that discussion about racial disparities, and debates about whether these disparities are due to race alone, or whether they reflect a combination of other differences in socioeconomic factors, is not very helpful, as they are all non-mutable factors. We argue that it would be more productive to focus on examining the potential mechanisms behind them, for example, provider referral patterns, or patient insurance coverage.

5.2.1.2 Provider Referral Patterns

As specialists, surgeons are rarely the first providers to interact with a patient during the workup and diagnosis process. Even when the possibility exists that a patient's illness requires surgical intervention, the screening and initial detection process was likely carried out by a primary care provider, who was then tasked with referring the patient for surgical evaluation. Therefore, it stands to reason that one of the first steps in the process where disparities in access to surgical care could arise would be at the provider referral stage. For example, Murphy and colleagues used the SEER database to study patient with pancreatic adenocarcinoma and found black patients were less likely to be referred for consultation to medical, radiation, and surgical oncologists, despite accounting for age, sex, marital status, SES, and Charlson Comorbidity Score (CCS) [21]. Simpson and colleagues identified similar disparities in multidisciplinary oncologic evaluation for black patients with metastatic colorectal cancer [22].

Though many factors likely contribute to disparate referral patterns, *implicit racial bias* on the part of providers is one with increasing evidence. Implicit biases are the attitudes or stereotypes that affect our understanding, actions, and decisions unconsciously, producing behavior that can markedly diverge from an individual's endorsed principles and beliefs [23]. Research has shown that implicit biases are more likely to manifest when individuals rely on System 1 mode of thinking, which is fast, automatic, and unconscious. Though physicians are generally expected to have lower rates of explicit bias than the general population, implicit bias is certainly prevalent within the field [24]. Given the fast-paced, stressful nature of surgery, it stands to reason that surgeons would be even more vulnerable to implicit biases than other specialists. Several clinical vignette-based studies have identified correlations between implicit bias and disparate treatment recommendations for a host of adult and pediatric conditions [25, 26]. Similarly, evidence of the existence of implicit biases has been found in referral patterns for bariatric surgery as well as prescription of narcotic pain medications for conditions otherwise commonly associated with drug abuse [27, 28]. Addressing implicit biases and their effects on clinical decision making should thus be of the utmost importance for addressing disparities in access to surgical care.

On the other hand, studies have found differences in patient perceptions of surgery that can certainly affect their likelihood of seeking surgical care. For example, black patients who suffer from morbid obesity have been found to be less likely to have considered metabolic and bariatric surgery (MBS) prior to presenting to their primary care office, primarily due to a view of surgery as a high-risk option of last resort [27]. However, sticking to the framework that for every patient factor there are potential solutions on the provider side, we could imagine improvements in provider communication and rapport could ameliorate potential differences in patient preferences and views of surgery.

5.2.1.3 Patient Insurance Coverage

Insurance is another potentially modifiable risk factor to mitigate documented access disparities, with some studies finding racial disparities are indeed mitigated by accounting for insurance coverage [15, 29–31]. Insurance affects access to care in two distinct ways: by impeding access altogether hence causing progression of disease and by limiting access to high volume surgeons/centers. For example, Garner et al. found that children with public insurance or uninsured, were more likely to be diagnosed with well-differentiated thyroid cancer at a significantly higher stage of disease when compared to their privately insured counterparts. Furthermore, children who were uninsured at the time of diagnosis experienced a 10-day delay in treatment initiation (28 days vs. 18 days) when compared to children who had private insurance [32]. Similar findings have been reported among adults with well-differentiated thyroid cancer, rectal cancer, breast cancer, carotid artery disease, and peripheral vascular disease [33–37].

Even when individuals who are underinsured are diagnosed and referred for surgical intervention, they often do not have access to the same hospitals and providers as their privately insured counterparts. For example, Huang et al. studied patients in the state of California who were diagnosed and treated for colorectal cancer from 1996 to 2006 and found individuals with public or no insurance were significantly less likely to be treated at high volume centers. Importantly, treatment at a high volume center was independently associated with an 8% decrease in the chance of dying compared to other settings [38]. Similar findings have been reported for the treatment of esophageal, gastric, lung, and pancreatic cancer, as well as critical limb ischemia, and mechanical thrombectomy after stroke [18, 39–41].

Over the past decade, efforts at the national and state level to drive insurance expansion and ameliorate the aforementioned access disparities have been underway, but data regarding their effects have so far presented conflicting results. For instance, studies examining the effects of the Affordable Care Act (ACA) insurance expansion found it led to improved access to care, earlier in the course of disease, with a higher likelihood of receiving optimal care [42]. Similar findings were reported by Loehrer and colleagues, who examined the effects of the Medicaid expansion in the state of Massachusetts, and found underinsured patients were more likely to undergo a number of complex cancer operations after the expansion, but were no more likely to undergo these operations at high volume hospitals than before the expansion [43].

Despite these conflicting results, efforts at the state and national level to address these insurance-based access disparities should continue and should include a close examination of the way hospitals negotiate reimbursements with both private and public insurers, focusing primarily on quality-based and not procedural metrics to justify such reimbursements [44].

5.2.2 Quality Factors

Differences in quality of care and surgical outcomes have been identified related to several factors such as race, insurance, and SES. Staying with the mindset that rather than capitulating these disparities out of context, it is best to attempt to place them within the mechanisms at the provider and system level that may contribute to them, we will divide surgical quality disparities into three categories: provision of quality care, provision of equitable care, and provision of appropriate care.

5.2.2.1 Provision of Quality Care

As mentioned above, minority patients often have less access to high volume, high-quality surgeons, which can ultimately affect their outcomes. For example, in 2016 Mehta et al. used a national database to evaluate outcomes after coronary artery bypass graft (CABG) and reported that black patients were more likely to be operated on by surgeons with higher risk-adjusted mortality rates, which resulted in higher overall postoperative morbidity and mortality [45]. In 2014, Hauch et al. used the National (Nationwide) Inpatient Sample (NIS) to study outcomes after thyroid surgery. They found Hispanic patients were more likely to be operated on by low-volume surgeons, and in certain regions throughout the United States, so did black patients. These low-volume surgeons had significantly longer LOS and higher complications rates. Similarly, patients with Medicare and lower income were more likely to be treated at low-volume centers [46]. In trauma care, younger black patients were found to be more likely than young white patients to be treated at low-performing facilities, which led to higher mortality [47].

This clustering effect of minority patients receiving care at lower volume centers is likely a substantial determinant of these observed health disparities. Popescu et al., found that physicians who tend to care for low-income and minority cancer patients generally have lower rates of compliance with guidelines [48].

In the end, surgical expertise can only be achieved through surgical experience, so the answer to this problem is not to bypass low-volume providers. Instead, we must focus our efforts towards developing practice guidelines that drive quality improvement and standardize care. One great example of this is the Enhanced Recovery After Surgery (ERAS) programs, which have now been applied to a wide array of surgical specialties to include colorectal surgery, gynecology, thoracic surgery, orthopedic surgery, urology, etc. ERAS uses standardized perioperative strategies to decrease the physiological stress invoked by surgery and to encourage early recovery. Wahl et al. demonstrated that using ERAS decreased the postoperative length of stay for black patients, which eliminated previously observed racial disparities [49]. These findings suggest that ERAS and other methods of standardization may provide a way to decrease surgical outcome disparities, as long as disparities related to access to care are simultaneously addressed. Further work should be done focusing on standardization of care beyond ERAS, to eliminate ongoing disparities in quality of care.

5.2.2.2 Provision of Equitable Care

Once a patient presents for surgical evaluation, one would expect providers to issue similar treatment recommendations to patients of similar surgical risk and disease state, but several studies have demonstrated otherwise. In 2018, Stapleton and colleagues used the New York Statewide Planning and Research Cooperative Systems (SPARCS) database to study treatment differences among patients with peripheral vascular disease, and found black patients were 45% more likely to receive an amputation than white patients, despite adjusting for other patient demographics, comorbidities, disease severity and year of procedure. Interestingly, this difference was more pronounced among hospitals and surgeons with higher overall volumes, as well as among surgeons who treat less black patients [7]. Downing et al. reported lower rates of limb salvage similarly after extremity sarcoma surgery for black patients when compared to whites [50]. Black et al. reported lower rates of axillary sentinel lymph node biopsy for pathologically node-negative breast cancer among black patients compared to whites, as well as among Medicaid patients and those with lower educational or income level. This resulted in higher rates of axillary lymph node dissection and lymphedema for these women [51]. Dubecz et al. found higher rates of inadequate lymphadenectomy after gastrointestinal cancer surgery for patients with lower income [52]. Lastly, race and socioeconomic based disparities have been identified regarding access to minimally invasive routes of a wide range of procedures to include appendectomy, cholecystectomy, colon surgery and AAA repair [53–57].

As described in the previous section concerning disparities in access to care, implicit bias on the part of the provider has been found to be associated with disparate treatment recommendations for different patient populations, at least using clinical vignette studies. No studies to date have been conducted to identify specific relationships between provider implicit bias and disparate real-life treatment decisions and clinical outcomes, though studies like those mentioned in the previous paragraph at the very least hint at its existence.

However, these disparate treatment recommendations may not always originate with the surgeon and may instead be heavily influenced by differences in insurance reimbursement rates and policies. For example, Patel el at. found significant disparities in the receipt of recommended radiation and/or chemotherapy after diagnosis or surgery for breast, endometrial, and colon cancer related to government-sponsored health insurance, which persisted despite adjustment for patient demographics [58]. Loehrer et al. studied the effects of insurance coverage on the timely provision of appropriate care for patients with acute cholecystitis and found that the Medicaid expansion in the state of Massachusetts led to an elimination of racial disparities in the receipt of immediate cholecystectomy observed prior to the expansion, which persisted in control states that did not undergo an insurance reform [59]. Nevertheless, a study published in 2015 by Mahmoudi et al. revealed the complex interaction between race, insurance, and disparities in the provision of care. They studied the effects of Medicaid expansion in the state of New York, which took place in 2001, on the likelihood of breast reconstruction after mastectomy. They found that despite an overall increase in the probability of patients undergoing immediate post-mastectomy breast reconstruction, the gap between white patients and both black and Hispanic patients increased by 6% and 5% respectively. This gap highlights the fact that ameliorating the barriers to care caused by lack of or underinsurance must be accompanied by intentional measures to address racial differences in patient education and access to pertinent consultation services (i.e. plastic surgeon) [60].

5.2.2.3 Provision of Appropriate Care

Speaking of equitable care requires addressing the concept of appropriate care. According to the RAND/UCLA Appropriateness Method, "[a]n appropriate procedure is one in which "the expected health benefit (e.g., increased life expectancy, relief of pain, reduction in anxiety, improved functional capacity) exceeds the expected negative consequences (e.g., mortality, morbidity, anxiety, pain, time lost from work) by a sufficiently wide margin that the procedure is worth doing, exclusive of cost" [61]. The concept of appropriateness of care thus attempts to address both the overuse and underuse of surgical procedures, with disparities related to both identified in current literature. Halm et al. used the RAND/UCLA Appropriateness Method to study patients in New York State who underwent carotid endarterectomies (CEA) between 1998 and 1999 and found black and Hispanic patients had higher rates of inappropriate CEA as well as lower rates of CEA being performed for appropriate indications. The excess in inappropriate CEAs for nonwhite patients was explained primarily by the fact that they had higher levels of preoperative comorbidities [50]. Conversely, a systematic review published in 2015 reported opposite findings to the Halm study and identified 59 unique studies, a substantial proportion of which (32) found evidence of more overuse of inappropriate care by whites than racial/ethnic minorities [62].

As surgeons, it is imperative we strive to provide care that is appropriate, as determined by national standards and best-practice guidelines, to all our patients alike, incorporating conscious checks and balances to combat our own inherent implicit biases.

5.3 Culturally Competent Science

While most of the work in disparity research has been focused on the care delivery process, much less attention has been directed to the scientific bases of our current clinical knowledge [63]. However, lack of awareness of disparities in the scientific process may lead to greater harms to patients than flaws in the delivery process. Problems can range from defining a treatment guideline that is not generalizable and applicable to non-white populations, to inappropriate definition of normal versus disease states. For example, the breast cancer screening guideline, as defined in the U.S. and based on a majority white population, has been shown to miss the peak incidence of all other non-white patient groups, raising the concern that a physician could be delivering appropriate care according to published guidelines, and still cause harms to non-white patients [64]. Another example of this Eurocentric approach when creating guidelines is the perception that certain foods from non-white populations are "not healthy", while most "healthy" foods come from the European diet. Problems with definitions can also go in the other direction and potentially mask clinically significant concerns. For example, African American patients are known to have lower white blood cell counts at baseline than white patients, hence the current definition of leukocytosis could under-diagnose it among African American trauma patients [65]. Similar concerns could be raised regarding the concept of "normal" pelvic anatomy, and how it may influence management of normal deliveries in non-white mothers [66].

One of the root causes of the problem may be an assumption that science is objective. However, while data themselves may be objective, their interpretations are not, and this subjectivity in the scientific process is likely surprising to many. Subtle biases, likely unintentional, can lead to dramatically different interpretations of the same objective data. For example, the majority population may be subtly considered to be the "normal" group, and trends in non-white groups considered "abnormal", such that the prevalence of ADHD can alternatively be judged as being under-diagnosed in non-white populations, instead of over-diagnosed in whites. It is imperative, therefore, that we remain aware of cultural competency issues, not just in the healthcare delivery process, but also in the scientific process.

Another important contributor to this bias is the lack of diversity in the scientific community. Modern science in the last two centuries has taken place mostly in the U.S. and Europe, where the majority population is white. The long-term solution to this problem will likely only come when we encourage students from diverse backgrounds from around the world to not just become physicians, but become physician-scientists [63].

As a surgeon-scientist today, there are certain steps you can take to move towards culturally competent science, for example diversifying your study populations. However, this change alone will not address the problem of generalizability because we often try to come to a single conclusion in our studies, which can result in trends in non-white populations getting "averaged out". It is more important to incorporate the concept of heterogeneity of treatment effects, and consider multiple analyses and multiple conclusions, each tailored to a different patient population. We need to keep in mind that what works for one population, or for one health system, may not work for another. One approach to drive home this point is to consider using different reference groups in statistical analyses, rather than the default group of white male patients. Another approach could be to stratify on non-modifiable risk factors such as race, and avoid merely adjusting for them in a regression model.

5.3.1 Final Comment

Considerable progress has been made over the past decade in highlighting health care disparities. However, there is still a long way to go until we have a healthcare system that eliminates undesirable variations in patient outcomes. We need to move beyond descriptive studies of disparities. For every description of disparities along non-modifiable risk factors we need to challenge ourselves to think of a different perspective that would shed light on modifiable ones. Lastly, we need to challenge the root causes beyond the care delivery process, and examine the lack of diversity in the scientific process and how these factors contribute to disparity.

References

1. Beal AC. High-quality health care: the essential route to eliminating disparities and achieving health equity. Health Aff. 2011;30:1868–71.
2. U.S. Department of Health & Human Services. Office of Minority Health. National Stakeholder Strategy for Achieving Health Equity. 2018. https://www.minorityhealth.hhs.gov/npa/files/Plans/NSS/CompleteNSS.pdf Accessed April 23, 2019.
3. Haider AH, Scott VK, Rehman KA, Velopulos C, Bentley JM, Cornwell EE III, Al-Refaie W. Racial disparities in surgical care and outcomes in the United States: a comprehensive review of patient, provider, and systemic factors. J Am Coll Surg. 2013;216:482–92.
4. Birkmeyer JD, Reames BN, McCulloch P, Carr AJ, Campbell WB, Wennberg JE. Understanding regional variation in the use of surgery. Lancet. 2013;382:1121–9.
5. McCulloch P, Nagendran M, Campbell WB, Price A, Jani A, Birkmeyer JD, Gray M. Strategies to reduce variation in the use of surgery. Lancet. 2013;382:1130–9.
6. Soden PA, Zettervall SL, Curran T, Vouyouka AG, Goodney PP, Mills JL, Hallett JW, Schermerhorn ML, Society for Vascular Surgery Vascular Quality Initiative. Regional variation in patient selection and treatment for lower extremity vascular disease in the vascular quality initiative. J Vasc Surg. 2017;65:108–18.
7. Stapleton SM, Bababekov YJ, Perez NP, Fong ZV, Hashimoto DA, Lillemoe KD, Watkins MT, Chang DC. Variation in amputation risk for black patients: uncovering potential sources of Bias and opportunities for intervention. J Am Coll Surg. 2018;226:641–9.e1
8. Egbert LD, Rothman IL. Relation between the race and economic status of patients and who performs their surgery. N Engl J Med. 1977;297:90–1.
9. Franklin JH, Chavez-Thompson L, Johnson Cook SD, Kean TH, Oh AE, Thomas R, Winter WF. One America in the 21st Century. The President's Initiative on Race. 1998.
10. Haider AH, Dankwa-Mullan I, Maragh-Bass AC, et al. Setting a National Agenda for surgical disparities research. JAMA Surg. 2016;151:554–63.
11. Britton BV, Nagarajan N, Zogg CK, Selvarajah S, Schupper AJ, Kironji AG, Lwin AT, Cerullo M, Salim A, Haider AH. Clinical science awareness of racial/ethnic disparities in surgical outcomes and care: factors affecting acknowledgment and action. Am J Surg. 2016;212:102–8.

12. De Jager E, Hons M, Levine AA, et al. Disparities in surgical access: a systematic literature review, conceptual model, and evidence map. J Am Coll Surg. 2019;228:276–98.
13. Saldana-Ruiz N, Clouston SAP, Rubin MS, Colen CG, Link BG. Fundamental causes of colorectal cancer mortality in the United States: understanding the importance of socioeconomic status in creating inequality in mortality. Am J Public Health. 2013;103:99–104.
14. Krieger N, Chen JT, Waterman PD. Decline in US breast cancer rates after the women's health initiative: socioeconomic and racial/ethnic differentials. Am J Public Health. 2010;100:S132–9.
15. Lee SL, Shekherdimian S, Chiu VY, Sydorak RM. Perforated appendicitis in children: equal access to care eliminates racial and socioeconomic disparities. J Pediatr Surg. 2010;45:1203–7.
16. Statement on Optimal Access. Am Coll Surg. 2013. https://www.facs.org/about-acs/statements/72-optimal-access. Accessed. April 23, 2019.
17. Taioli E, Flores R. Appropriateness of surgical approach in Black patients with lung cancer—15 years later, little has changed. J Thorac Oncol. 2017;12:573–7.
18. Attenello FJ, Adamczyk P, Wen G, He S, Zhang K, Russin JJ, Sanossian N, Amar AP, Mack WJ. Racial and socioeconomic disparities in access to mechanical revascularization procedures for acute ischemic stroke. J Stroke Cerebrovasc Dis. 2014;23:327–34.
19. Haider AH, Schneider EB, Sriram N, et al. Unconscious race and social class bias among acute care surgical clinicians and clinical treatment decisions. JAMA Surg. 2015;150:457–64.
20. Wang L, Haberland C, Thurm C, Bhattacharya J, Park KT. Health outcomes in US children with abdominal pain at major emergency departments associated with race and socioeconomic status. PLoS One. 2015;10:e0132758.
21. Murphy MM, Simons JP, Ng SC, Mcdade TP, Smith JK, Shah SA, Zhou Z, Earle CC, Tseng JF. Racial differences in cancer specialist consultation, treatment, and outcomes for locoregional pancreatic adenocarcinoma. Ann Surg Oncol. 2009;16:2968–77.
22. Simpson DR, Martínez ME, Gupta S, Hattangadi-Gluth J, Mell LK, Heestand G, Fanta P, Ramamoorthy S, Le Q-T, Murphy JD. Racial disparity in consultation, treatment, and the impact on survival in metastatic colorectal Cancer. J Natl Cancer Inst. 2013;105:1814–20.
23. Staats C, Capatosto K, Wright R, Jackson V (2016) 2016 State of the science: implicit bias review. https://www.issuelab.org/resource/2016-state-of-the-science-implicit-bias-review.html. Accessed 23 Apr 2019.
24. Sineshaw HM, Ng K, Flanders WD, Brawley OW, Jemal A. Factors that contribute to differences in survival of Black vs white patients with colorectal cancer. Gastroenterology. 2018;154:906–915.e7.
25. Green AR, Carney DR, Pallin DJ, Ngo LH, Raymond KL, Iezzoni LI, Banaji MR. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. J Gen Intern Med. 2007;22:1231–8.
26. Sabin JA, Greenwald AG. The influence of implicit bias on treatment recommendations for 4 common pediatric conditions: pain, urinary tract infection, attention deficit hyperactivity disorder, and asthma. Am J Public Health. 2012;102:988–95.
27. Wee CC, Huskey KW, Bolcic-Jankovic D, Colten ME, Davis RB, Hamel M. Sex, race, and consideration of bariatric surgery among primary care patients with moderate to severe obesity. J Gen Intern Med. 2014;29:68–75.
28. Singhal A, Tien YY, Hsia RY. Racial-ethnic disparities in opioid prescriptions at emergency department visits for conditions commonly associated with prescription drug abuse. PLoS One. 2016;11:1–14.
29. Mannix R, Bourgeois FT, Schutzman SA, Bernstein A, Lee LK. Neuroimaging for pediatric head trauma: do patient and hospital characteristics influence who gets imaged? Acad Emerg Med. 2010;17:694–700. https://doi.org/10.1111/j.1553-2712.2010.00797.x.
30. Weber TL, Ziegler KM, Kharbanda AB, Payne NR, Birger C, Puumala SE. Leaving the emergency department without complete care: disparities in American Indian children. BMC Health Serv Res. 2018;18:267. https://doi.org/10.1186/s12913-018-3092-z.
31. Baker DW, Stevens CD, Brook RH. Determinants of emergency department use: are race and ethnicity important? Ann Emerg Med. 1996;28:677–82. https://doi.org/10.1016/S0196-0644(96)70093-8.

32. Garner EF, Maizlin II, Dellinger MB, et al. Effects of socioeconomic status on children with well-differentiated thyroid cancer. Surgery. 2017;162:662–9.
33. Zagzag J, Kenigsberg A, Patel KN, Heller KS, Ogilvie JB. Thyroid cancer is more likely to be detected incidentally on imaging in private hospital patients. J Surg Res. 2017;215:239–44.
34. Pulte D, Jansen L, Brenner H. Population-level differences in rectal Cancer survival in uninsured patients are partially explained by differences in treatment. Oncologist. 2017;22:351–8.
35. Lukavsky R, Sariego J. Insurance status effects on stage of diagnosis and surgical options used in the treatment of breast Cancer. South Med J. 2015;108:258–61.
36. Brinjikji W, El-Sayed AM, Kallmes DF, Lanzino G, Cloft HJ. Racial and insurance based disparities in the treatment of carotid artery stenosis: a study of the Nationwide inpatient sample. J Neurointerv Surg. 2015;7:695–702.
37. Loehrer AP, Hawkins AT, Auchincloss HG, Song Z, Hutter MM, Patel VI. Impact of expanded insurance coverage on racial disparities in vascular disease: insights from Massachusetts. Ann Surg. 2016;263:705–11.
38. Huang LC, Tran TB, Ma Y, Ngo JV, Rhoads KF. Factors that influence minority use of high-volume hospitals for colorectal cancer care. Dis Colon Rectum. 2015;58:526–32.
39. Stitzenberg KB, Meropol NJ. Trends in centralization of Cancer surgery. Ann Surg Oncol. 2010;17:2824–31.
40. Sun M, Karakiewicz PI, Sammon JD, Sukumar S, Gervais M-K, Nguyen PL, Choueiri TK, Menon M, Trinh Q-D. Disparities in selective referral for cancer surgeries: implications for the current healthcare delivery system. BMJ Open. 2014;4:e003921.
41. Henry AJ, Hevelone ND, Belkin M, Nguyen LL. Socioeconomic and hospital-related predictors of amputation for critical limb ischemia. J Vasc Surg. 2011;53:330–339.e1.
42. Loehrer AP, Chang DC, Scott JW, Hutter MM, Patel VI, Lee JE, Sommers BD. Association of the Affordable Care Act Medicaid Expansion with Access to and quality of care for surgical conditions. JAMA Surg. 2018;153:e175568. https://doi.org/10.1001/jamasurg.2017.5568.
43. Loehrer AP, Chang DC, Chang GJ. Health reform and utilization of high-volume hospitals for complex Cancer operations. J Oncol Pract. 2018;14:39–50.
44. Schulman KA, Milstein A. The implications of “Medicare for all” for US Hospitals. JAMA. 2019;321:1661. https://doi.org/10.1001/jama.2019.3134.
45. Mehta RH, Shahian DM, Sheng S, O’Brien SM, Edwards FH, Jacobs JP, Peterson ED. Association of Hospital and Physician Characteristics and care processes with racial disparities in procedural outcomes among contemporary patients undergoing coronary artery bypass grafting surgery. Circulation. 2016;133:124–30.
46. Hauch A, Al-Qurayshi Z, Friedlander P, Kandil E. Association of socioeconomic status, race, and ethnicity with outcomes of patients undergoing thyroid surgery. JAMA Otolaryngol Head Neck Surg. 2014;140:1173–83.
47. Hicks CW, Hashmi ZG, Hui X, Velopulos C, Efron DT, Schneider EB, Cooper L, Haut ER, Cornwell EE, Haider AH. Explaining the paradoxical age-based racial disparities in survival after trauma: the role of the treating facility. Ann Surg. 2015;262:179–83.
48. Popescu I, Schrag D, Ang A, Wong M. Racial/ethnic and socioeconomic differences in colorectal and breast Cancer treatment quality. Med Care. 2016;54:780–8.
49. Wahl TS, Goss LE, Morris MS, et al. Enhanced recovery after surgery (ERAS) eliminates racial disparities in postoperative length of stay after colorectal surgery. Ann Surg. 2018;268:1026–35.
50. Downing S, Ahuja N, Oyetunji TA, Chang D, Frederick WAI. Disparity in limb-salvage surgery among sarcoma patients. Am J Surg. 2010;199:549–53.
51. Black DM, Jiang J, Kuerer HM, Buchholz TA, Smith BD. Racial disparities in adoption of axillary sentinel lymph node biopsy and lymphedema risk in women with breast Cancer. JAMA Surg. 2014;149:788–96.
52. Dubecz A, Solymosi N, Schweigert M, Stadlhuber RJ, Peters JH, Ofner D, Stein HJ. Time trends and disparities in lymphadenectomy for gastrointestinal cancer in the United States: a population-based analysis of 326,243 patients. J Gastrointest Surg. 2013;17:611–9.
53. Loehrer AP, Song Z, Auchincloss HG, Hutter MM. Massachusetts health care reform and reduced racial disparities in minimally invasive surgery. JAMA Surg. 2013;148:1116–22.

54. Oyetunji TA, Nwomeh BC, Ong'uti SK, Gonzalez DO, Cornwell EE, Fullum TM. Laparoscopic appendectomy in children with complicated appendicitis: ethnic disparity amid changing trend. J Surg Res. 2011;170:e99–e103.
55. Damle RN, Flahive JM, Davids JS, Maykel JA, Sturrock PR, Alavi K. Examination of racial disparities in the receipt of minimally invasive surgery among a National Cohort of adult patients undergoing colorectal surgery. Dis Colon Rectum. 2016;59(11):1055–62.
56. Greenstein AJ, Romanoff AM, Moskowitz AJ, Sosunov EA, Khaitov S, Egorova NN. Payer status and access to laparoscopic subtotal colectomy for ulcerative colitis. Dis Colon Rectum. 2013;56:1062–7.
57. Osborne NH, Mathur AK, Upchurch GR, Dimick JB. Understanding the racial disparity in the receipt of endovascular abdominal aortic aneurysm repair. Arch Surg. 2010;145(11):1105–8.
58. Parikh-Patel A, Morris CR, Kizer KW. Disparities in quality of cancer care: the role of health insurance and population demographics. Medicine (Baltimore). 2017;96:e9125. https://doi.org/10.1097/MD.0000000000009125.
59. Loehrer AP, Song Z, Auchincloss HG, Hutter MM. Influence of health insurance expansion on disparities in the treatment of acute Cholecystitis. Ann Surg. 2015;262:139–45.
60. Mahmoudi E, Giladi AM, Wu L, Chung KC. Effect of federal and state policy changes on racial/ethnic variation in immediate postmastectomy breast reconstruction. Plast Reconstr Surg. 2015;135:1285–94.
61. Fitch K, Bernstein SJ, Aguilar MD, Burnand B, LaCalle JR, Lazaro P, van het Loo M, McDonnell J, Vader J, Kahan JP (2001) The RAND/UCLA Appropriateness Method User's Manual.
62. Kressin NR, Groeneveld PW. Race/ethnicity and overuse of care: a systematic review. Milbank Q. 2015;93:112–38. https://doi.org/10.1111/1468-0009.12107.
63. Chang DC, General M, Peabody FW, Molina G. Culturally competent science. JAMA Surg. 2018;153:699–701.
64. Stapleton SM, Oseni TO, Bababekov YJ, Hung YC, Chang DC. Race/ethnicity and age distribution of breast Cancer diagnosis in the United States. JAMA Surg. 2018;153(6):594–5.
65. Chang DC, Cornwell EE 3rd, Phillips J, Paradise J, Campbell K. Early leukocytosis in trauma patients: what difference does it make? Curr Surg. 2003;60(6):632–5.
66. Yin S. Why textbooks may need to update what they say about birth canals. New York Times, October 27, 2018.

Suggested Readings

Dimick J, Ruhter J, Sarrazin MV, Birkmeyer JD. Black patients more likely than whites to undergo surgery at low-quality hospitals in segregated regions. Health Aff. 2013;32(6):1046–53.

Haider AH, Scott VK, Rehman KA, Velopulos CG, Bentley JM, Cornwell EE III, Al-Refaie W. Racial disparities in surgical care and outcomes in the United States: a comprehensive review of patient, provider and systemic factors. J Am Coll Surg. 2013a;216(3):482–92.

Haider AH, Weygandt PL, Bentley JM, Monn MF, Rehman KA, Zarzaur BL, Crandall ML, Cornwell EE III, Cooper LA. Disparities in trauma care and outcomes in the United States: a systematic review and meta-analysis. J Trauma Acute Care Surg. 2013b;74(5):1195–205.

Morris AM, Rhoads KF, Stain SC, Birkmeyer JD. Understanding racial disparities in cancer treatment and outcomes. J Am Coll Surg. 2010;211(1):105–13.

National Research Council. Unequal treatment: confronting racial and ethnic disparities in health care. Washington, DC: The National Academies Press; 2003.

Measuring Surgical Quality

6

Ryan P. Merkow and Karl Y. Bilimoria

Abstract

Surgery is a high stakes area of medicine where the actual care provided (i.e., the procedure performed) places patients at risk for significant morbidity and mortality. Measuring and improving surgical care in the U.S. is increasingly the focus of various stakeholders of healthcare quality, including payers, oversight organizations, hospitals, clinicians, and most importantly patients. However, measuring surgical quality is not straightforward. In this chapter, we will define concepts underlying surgical quality measurement, and highlight many of the important considerations that should be understood and adopted when studying surgical quality. Finally, we discuss future areas that we believe are important to further move surgical quality measurement and improvement forward.

Surgical services represents over 40% of all health care spending in the United States with over 50 million inpatient procedures performed annually [1, 2]. With the aging population requiring more surgical procedures, adoption of new, less invasive surgical techniques (e.g., robotics), and as surgical indications expand, the annual number of procedures performed will continue to increase. However, surgery is a high stakes area of medicine where the actual care provided (i.e., the procedure performed) places patients at risk for significant morbidity and mortality. Decades of data has repeatedly demonstrated that the risks and benefits are largely driven by differences in patient selection, the hospital in which the patient receives care, and

R. P. Merkow (✉) · K. Y. Bilimoria
Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery,
Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
e-mail: Ryan.Merkow@Northwestern.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_6

the experience of the individual care providers (e.g., the surgeon). In addition, the occurrence of a single postoperative complication can result in a dramatic difference in cost [3, 4].

Therefore, it is not surprising that over the last decade a growing national interest in surgical quality measurement and improvement has occurred. Although current methods evaluating surgical quality have relevance, new concepts and strategies are also emerging to more efficiently measure and translate identifiable problems in quality in to actual real time improvements. In this chapter, we will review the basic quality measurement framework and explore new areas of surgical quality measurement and improvement.

6.1 Surgical Quality Measurement Framework

The conventional approach when evaluating quality is classically based on Donabedian's conceptual framework: structure, process, and outcomes [5, 6]. Care can be evaluated based on the setting in which it is delivered (structure), the actual care delivered (process), or its end points (outcomes). Each pillar of quality assessment has distinct pros and cons, and it is becoming clear that no program should focus on only one type of measurement tool to improve and sustain surgical quality.

Early approaches to surgical quality measurement and improvement focused on structure and process of care such as the volume-outcome relationship or the delivery of guideline recommended treatment (e.g., adjuvant chemotherapy in stage III colon cancer). The primary benefits in using these indicators is they are more easily measurable, and do not require sophisticated risk adjustment to control for differences in case-mix. However, alone, these approaches have a number of shortcomings. First, they are based on the assumption that a direct link exists to improved outcomes. Yet, in real-world practice, this assumption has been challenging to consistently demonstrate, and it is unclear how to interpret the soft links [7]. For example, a variety of stakeholders of healthcare quality have used surgical volume for purposes of surgical quality assessment. On average, there may be an association between volume and performance, but the true factors leading to high-quality care remain undefined. Other process measures, such as the SCIP measures, were shown to lack a direct link for individual measures, but when taken as a bundle of care, there was in fact an association with improved outcomes. This has also been demonstrated repeatedly with the expansion of enhanced recovery pathways in colorectal surgery [8].

It should also be emphasized that a direct link to improved outcomes may not be necessary for a structural or process measure to be considered important. For example, having the necessary structural resources (e.g., imaging modalities, specialists, multidisciplinary conferences) available to provide complex cancer care is considered standard without the need for any outcome linkages [9, 10].

The last of Donabedian's three part framework is studying the end points, or outcomes, of surgical quality. This approach has obvious face validity for stakeholders of surgical quality. Outcomes are generally straightforward to interpret, particularly for patients who are increasingly being provided with performance data to make decisions on where to go for their treatment. Moreover, health care providers can unambiguously assess how their care compares with their peers, as well as the success of their process-improvement efforts. Nevertheless, outcomes assessment also has a number of shortcomings that should be considered [11].

First, it is not always clear what outcomes should be used to measure surgical quality. Although outcomes represent the end point of care, they must also be meaningful surrogate measures of quality. For example, many outcomes- based initiatives focus solely on inpatient mortality rates owing to its low data collection burden, limited need for definition standardization, and accessibility from administrative data sources. However, inpatient mortality alone is a rare event and does not capture other relevant and perhaps more preventable complications such as 30-day surgical site infections. Other outcomes, such as venous thromboembolism, may be more reflective of how vigilant a hospital is at identifying events (i.e., surveillance bias) and therefore may not represent a true marker of surgical quality [12]. Other methodological considerations also exist. For example, outcomes require robust risk adjustment such that outcomes can be fairly compared between providers and hospitals that may take care of sicker patients [13]. However, it is not always clear what risk-adjustment variables should be included, or what patient cohort should make up the denominator. The answers are highly relevant, particularly to assure health care providers that their performance is being accurately assessed. A recent example is whether or not socioeconomic status should be included when measuring hospital performance based on readmissions [14].

An additional concern when assessing outcomes relates to reliability, which represents the amount of variability in an assessment that is due to a real difference in performance compared with variability due to measurement error [15]. Uncommon events (e.g., mortality) and procedures that are performed infrequently (e.g., esophagectomy) have low reliability, making accurate quality measurements challenging. This issue is particularly relevant to the measurement of individual surgeon quality assessment.

A final consideration is the source and expense of outcomes information. Clinical data has obvious advantages over administrative claims sources; however, it is also more expensive and requires substantially more resources and expertise to collect. However, it is particularly well suited for collaboratives and patient safety organizations. Nevertheless, many hospitals that have the greatest need for high-quality clinical outcomes assessment simply cannot afford it. Funding partnerships with insurance companies has demonstrated feasibility in several states. On the other hand, administrative data are not standardized and are generated as a result of billing for health care services. However, administrative data has a number of strengths including the ability to study patients longitudinally over time, measure and compare health care payments, and evaluate larger populations of patients and hospitals.

6.2 Data Sources

There are several data sources available to assess surgical quality. Data sources can be organized into two broad groups: clinical and administrative. Clinical data is considered the gold standard as it is generally collected by trained and audited personnel for the purposes of quality assessment. Therefore, the variables collected will be standardized and uniform. For example, the definition of a surgical site infection may vary widely between surgeons. Without a standard definition, it may not be valid to compared surgeons or hospitals. The best example of clinical data is the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) [16]. Administrative data is generally collected as a result of another process, such as billing, such as Medicare data [17]. This data is not standardized and there are important biases that impact the collection and reporting of surgeon and hospital quality using this data source.

Nevertheless, both types of data have important pros and cons. Clinical data is often expensive to collect, and outcomes such as mortality and readmission are probably collected just as well by administrative data. Moreover, many hospitals that have the greatest need for high-quality clinical outcomes assessment simply cannot afford it. Funding partnerships within the context of collaborative and/or with insurance companies has demonstrated feasibility in several states. Administrative data only reliably identifies outcomes that can be billed for. For example, superficial surgical site infection is poorly reported as there are few opportunities to bill for this complication, whereas readmission is almost always captured in administrative data. Administrative data has additional strengths including the ability to study patients longitudinally over time, measure and compare health care payments, and evaluate larger populations of patients and hospitals. These limitations and others must be considered when measuring surgical quality.

6.3 Statistical Considerations

Prior to analyzing data, the researchers' objectives and analytic strategy should be planned and discussed both for clinical relevance and statistical appropriateness [18]. Two of the most important concerns is managing confounding and bias when measuring surgical quality. Confounding is generally easier to manage as it can be more easily tested for and addressed with appropriate risk adjustment and techniques such as hierarchical modeling. Selection bias is significantly more challenging to address. It is often impossible to know the magnitude and direction of this type of bias in observational research. For example, when comparing two procedure techniques, it is impossible to go back in time to determine whether patients were selected for one procedure or the other for any particular reason. Propensity score analyses, a statistical tool which helps balance comparative groups based on known confounding variables may minimize selection bias, however can never completely address differences in selection.

A second important consideration when measuring surgical quality is handling missing data. Certain variables, particularly in administrative datasets, may be missing in over 50% of observations. These variables should generally be discarded. Other variables may have substantial amounts of missing information (greater than 10% but less than 50%) but may still be informative. Options for managing missing data include coding dummy variables, excluding observations with missing data, simply not using the variable in the analysis or the preferred method of imputation [19].

Estimating hospital level practice patterns and outcomes is another important consideration. Deciphering statistical noise from actual meaningful differences between hospitals is not straightforward, particularly for lower volume hospitals. Such imbalances in the signal to noise ratio, even after risk adjustment, makes interpreting results challenging [20]. A statistical technique using hierarchical modeling allows for more fair and reliable estimates as it accounts for instances in which the "second level" (usually the hospital) has very few cases by adjusting to the grand mean of the particular outcome of interest. Hierarchical modeling is considered standard when the "second level" of interest can be assigned [15].

6.4 New Approaches to Surgical Quality Measurement

Other Measurement strategies in recent years have brought new insights to surgical quality assessment. One of the most interesting new advances is the idea that quality of care should include the individual patient's perspective, or patient reported outcomes (PROs) [21]. For example, a cancer surgeon may judge quality based on whether the surgical margins were free of tumor or whether an adequate lymph node dissection was achieved. An orthopedic surgeon may judge the success of the operation based on device failure or wound infection rates. The individual patient may be equally or more concerned about their overall quality of life, functional outcome, and/or pain control. Development of patient-reported outcome methodology is now underway and has clear importance in surgery.

Other important aspects of quality include the appropriateness [22], accessibility and timeliness of care [23]. Hospitals may have better-than-expected outcomes; however, if patients have to wait months for definitive treatment and do not receive the right care at the right time, it can be easily argued that surgical quality is deficient.

The value, or quality per unit cost, of care should also be implemented into comparative assessment programs. Currently, there is no mechanism in place that allows hospitals to fairly compare their surgical value with other hospitals in a standardized fashion [24]. One significant barrier is there is limited available data on the actual cost of complications. Most analyses to date use hospital charges or insurance payments. Charges have little association with actual costs of care and are not able to provide meaningful comparative data. Insurance payments may more closely approximate actual costs of care, however, are widely variable depending on payer and are influenced by hospital and patient specific factors (e.g., insurance coverage) that limit its utility in cost analyses. Medicare payments are additionally limited in that they are based on diagnosis-related groups, blunting the relationship between individual complications and costs.

Finally, many strategies exist to measure surgical quality; however, taken separately, each individual metric may have a limited or weak relationship on overall surgical quality. One approach is to combine a diverse set of measures spanning all aspects of quality into a composite index. True composite measures are powerful tools—they derive the reliability necessary from each component and may better forecast future performance. For example, measuring mortality alone may not produce a reliable estimate of surgical quality; however, a composite index that included mortality, other serious complications, a select number of important process of care and structural measures, patient-centered outcomes, and standardized costs may better reflect true health care provider performance. Although there are certainly major limitations to this approach, including how to weight the different components, composite indices should be considered in future measurement strategies.

6.5 Conclusion

Reliable and accurate surgical quality measurement is necessary to improve care and reduce costs. Surgical quality assessment has unquestionably improved over the past several decades; however, to make substantive and sustainable improvements, measurement approaches must continue to evolve. With the growing focus on surgical quality and value by all stakeholders, we are hopeful about the future of surgical quality measurement and improvement.

Acknowledgment *Funding*: RPM is supported by the Agency for Healthcare Research Quality (K12HS026385) and an Institutional Research Grant from the American Cancer Society (IRG-18-163-24).

References

1. National Quality Forum. NQF-Endorsed Measures for Surgical Procedures, 2015–2017. Available at: https://www.qualityforum.org/Publications/2017/04/Surgery_2015-2017_Final_Report.aspx. Last accessed 15 Jan 2019.
2. Centers of Disease Control and Prevention. National Hospital Discharge Survey: 2009 Table, Procedures by selected patient characteristics - Number by procedure category and age. Available at https://www.cdc.gov/nchs/fastats/hospital.htm. Last accessed 15 Jan 2019.
3. Vonlanthen R, Slankamenac K, Breitenstein S, et al. The impact of complications on costs of major surgical procedures: a cost analysis of 1200 patients. Ann Surg. 2011;254(6):907–13.
4. Healy MA, Mullard AJ, Campbell DA Jr, et al. Hospital and payer costs associated with surgical complications. JAMA Surg. 2016;151(9):823–30.
5. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q 1966;44(3):Suppl:166–206.
6. Donabedian A. The quality of care. How can it be assessed? J Am Med Assoc. 1988;260(12):1743–8.
7. Stulberg JJ, Delaney CP, Neuhauser DV, et al. Adherence to surgical care improvement project measures and the association with postoperative infections. J Am Med Assoc. 2010;303(24):2479–85.
8. Owodunni OP, Hampton J, Bettick D, et al. High compliance to an enhanced recovery pathway for patients >/=65 years undergoing major small and large intestinal surgery is associated with improved postoperative outcomes. Ann Surg. 2018:1.
9. The American College of Surgeons, Cancer Programs. The Commission on Cancer. Available at: http://www.facs.org/cancer/coc/cocar.html. Last accessed 15 Aug 2019.

10. McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med. 2003;348(26):2635–45.
11. Merkow RP, Bilimoria KY, Ko CY. Surgical quality measurement: an evolving science. JAMA Surg. 2013;148(7):586–7.
12. Bilimoria KY, Chung J, Ju MH, et al. Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure. JAMA. 2013;310(14):1482–9.
13. Merkow RP, Hall BL, Cohen ME, et al. Relevance of the c-statistic when evaluating risk-adjustment models in surgery. J Am Coll Surg. 2012;214(5):822–30.
14. Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health Aff (Millwood). 2014;33(5):778–85.
15. Cohen ME, Ko CY, Bilimoria KY, et al. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. J Am Coll Surg 2013;217(2):336–346 e331.
16. The American College of Surgeons. ACS National Surgical Quality Improvement Program. Available at: https://www.facs.org/quality-programs/acs-nsqip. Last accessed 15 Jan 2019.
17. CMS Hospital Quality Initiatives. Available at https://www.cms.gov/HospitalQualityInits/. Last accessed 15 Jan 2019.
18. Merkow RP, Rademaker AW, Bilimoria KY. Practical guide to surgical data sets: National Cancer Database (NCDB). JAMA Surg. 2018;153(9):850–1.
19. Hamilton BH, Ko CY, Richards K, et al. Missing data in the American College of Surgeons National Surgical Quality Improvement Program are not missing at random: implications and potential impact on quality assessments. J Am College Surg 2010;210(2):125–139 e122.
20. Adams JL, Mehrotra A, Thomas JW, et al. Physician cost profiling--reliability and risk of misclassification. N Engl J Med. 2010;362(11):1014–21.
21. Health Measures: Transforming How Health Is Measured. Available at: http://www.healthmeasures.net/explore-measurement-systems/promis. Last accessed 15 Jan 2019.
22. Lee CN, Ko CY. Beyond outcomes--the appropriateness of surgical care. J Am Med Assoc. 2009;302(14):1580–1.
23. Institute of Medicine CoQoHiA, editor. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press; 2001.
24. Reinhardt UE. The pricing of U.S. hospital services: chaos behind a veil of secrecy. Health Aff (Millwood). 2006;25(1):57–69.

Using Data for Local Quality Improvement

7

Caroline E. Reinke and Rachel R. Kelz

Abstract

Effective quality improvement requires identification of a problem, highly reliable and valid data, and stakeholder engagement. Using best practices to improve care has been successful at local and regional levels. This chapter provides a high-level overview of data sources and quality improvement tools that can be utilized for local quality improvement.

7.1 Introduction

Achieving the best outcome for each surgical patient is dependent on providing excellent, comprehensive and coordinated care across the five phases of surgical care: preoperative (consultation and evaluation), immediate preoperative (24–48 h prior to the operation), intraoperative, postoperative, and postdischarge [1]. The concept of quality improvement has been critical to the surgical community, dating back a century to the days of Codman [2, 3]. While there have been many advancements in surgical outcomes over time in local, regional, and national settings, quality improvement is a continuous process with ample opportunity for continued progress.

Skeptics condemn the term quality improvement dismissing it as a process where both improvement and setbacks can occur. However, this amplifies the great

C. E. Reinke
Department of Surgery, Carolinas Medical Center, Charlotte, NC, USA

R. R. Kelz (✉)
Department of Surgery, Hospital of the University of Pennsylvania, Philadelphia, PA, USA
e-mail: Rachel.Kelz@uphs.upenn.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_7

misunderstanding about quality improvement methodology and underlines the need for additional education in this domain. Regardless of the chosen methodology, all quality improvement protocols include measurement of the intended goals and balancing measures to minimize or eliminate adverse outcomes that transpire unexpectedly. This ensures that the net gain from each initiative is positive or the process is halted.

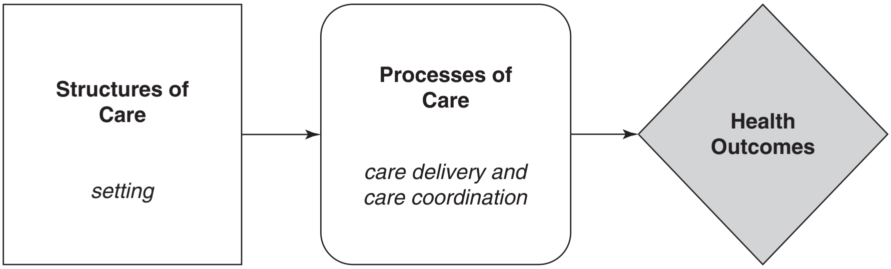
This chapter will provide a brief overview of quality improvement, different data sources and their structure, and how to use the data to identify the problem, engage stakeholders, and track the net impact of quality improvement initiatives at the local level. A detailed understanding of the data is important for successful quality improvement. Data are plentiful in the modern age of healthcare, but each data source has nuances that must be well understood in order for it to be used correctly and effectively. None of the data can provide the whole picture, and most commonly a combination of data sources will be needed to inform quality improvement initiatives and track improvements.

7.2 Quality Improvement

The U.S. Department of Health and Human Services Health Resources and Services Administration defines quality improvement as "systematic and continuous actions that lead to measurable improvement in health care services and the health status of targeted patient groups." [4] Quality improvement efforts require utilization of data to inform the process and an engaged team to share data and lead changes, and, therefore, is often most effective at the local level. With the development of larger surgical quality collaboratives, often statewide or regional, there has been increased efforts and successes at leading quality improvement efforts at the higher level [5–9]. These quality improvement efforts are lead and supported at a regional level by providing best practices resources, guides, and data collection/reporting support. However, the implementation of the work requires leaders and stakeholders familiar with the individual culture and resources at the local level to successfully implement changes.

The Surgical Care Improvement Project (SCIP) was one of the largest efforts at improving surgical care at the national level [10]. It began in 2002 as the Surgical Infection Prevention Project (SIP) and in 2006 was expanded to include measures aimed at reducing venous thromboembolism and cardiac events [11]. The goal was to standardize surgical quality improvement measures in a way that could be implemented on a national level. Although the SCIP may not have reached its goal of a reduction in surgical complications by 25% by 2010, it certainly increased awareness and changed practice across many hospitals in the United States.

7.3 Data



Donabedian classified measures used to assess and compare quality as a structure, process, or outcome [12, 13].

Structures of care reference the physical resources available at a certain location for patient care. For surgical patients, structures include hospital resources (critical care beds, operating room supplies, etc.), available personnel (surgical subspecialists, consultants, ancillary staff) and surgical volume. Processes of care refer to how things are done and are often the most easily impacted by providers. Outcomes are the results the patients experience after undergoing care. Table 7.1 outlines examples of each, as well as advantages and disadvantages [14].

7.3.1 Data Sources

For local quality improvement initiatives, local data will be needed to best understand the care provided and the outcomes achieved. The most common data sources are administrative data (sometimes referred to as billing or discharge data), quality registry data, and data available through the medical record. Some facilities have the capacity to mine their electronic health records while others must gather this information by hand.

Administrative data is collected by the hospital as part of its routine business of providing hospital-based care. Information available through this type of data includes patient demographics, International Classification of Diseases (ICD) diagnoses, ICD procedure codes, relevant dates (admission, discharge, procedure) and charges. Administrative data are relatively easy to analyze and often serve a critical role at determining the magnitude of a problem within a local setting. Administrative data notably lacks some patient characteristics that are often of interest for quality improvement such as body mass index (BMI), Do Not Resuscitate (DNR) status, and social determinants of health. The discharge or claims data also do not include process measures although, with some clever manipulation, proxies for some measures can be

Table 7.1 Using structure, process, and outcomes to measure surgical quality with examples, advantages, and disadvantages of each [14] (used with permission)

	Structure	Process	Outcomes
Examples	Procedure volume	Perioperative β-blockers in high-risk surgical patients	Morbidity and mortality rates
	Fellowship-trained surgeons	Use of internal mammary graft during coronary artery bypass graft	Functional health status
	"Closed" intensive care units		Patient satisfaction Cost
Primary advantage(s)	Expedient, inexpensive proxies of surgical outcomes	Reflect care that patients actually receive—may seem "fairer" to providers	Buy-in from surgeons—the "bottom line" of what they do
		Actionable from provider perspective, clear link to quality improvement activities	Outcomes measurement alone may improve outcomes
Disadvantages	Most variables not actionable from provider perspective	Little information about which processes are important for specific procedures	Numbers too small to measure with adequate precision procedure-specific outcomes for most hospitals and procedures
	Imperfect proxies for outcomes—reflect average results for large groups of providers, not individuals		Outcomes measures that are not procedure-specific less useful for purposes of quality improvement

abstracted. For example, while you cannot determine if a patient received chemotherapy, you can often tell if there was a bill generated for its administration. Some outcomes are easily identifiable and highly accurate in this data, such as inpatient death and length of stay. Other outcomes -identified by an ICD diagnosis code - are only identifiable if they happen during a hospital stay and with variable accuracy [15].

Many hospitals upload their administrative data to a national organization, such as Vizient (https://www.vizientinc.com/) or Premier (https://www.premierinc.com/), to receive comparative statistics on their performance. These companies allow hospitals to leverage risk-adjusted comparison of their local data to other hospitals, including outcomes and cost, to better understand where there may be opportunities for improving the value of care delivery. Benchmarking administrative data, as well as collaborative relationships built through such data sharing, are useful for learning about successes that other hospitals have had and sharing knowledge. These data are often shared through government agencies to eventually become publicly available in datasets such as the Nationwide Inpatient Sample [16].

In the aggregated form, the administrative data becomes useful for researchers. However, the NIS cannot be used for local quality improvement efforts as critical identifiers are stripped from the dataset to prevent the identification of individuals.

There are also a variety of sources of data available for other analysis. Many common datasets used for research are available through governmental agencies and describe various aspects of health and healthcare across the United States. Examples include the Nationwide Inpatient Sample, the Medicare, Surveillance, Epidemiology and End Results Program (SEER), and the National Health and Nutrition Examination Survey (NHANES). These datasets provide information about patient demographics and outcomes at a national level, and occasionally can even provide some process measures (SEER). However, due to the de-identified nature of these datasets they cannot be used to study outcomes and practices at a local level. A checklist for improving the use of surgical databases for research and detail on individual databases was recently published in JAMA Surgery [17].

Quality registries are increasingly used for tracking outcomes after surgical care. The data included in quality registries are obtained and entered by trained chart reviewers. The most well-known of these is the National Surgery Quality Improvement Program (NSQIP), which began in the Veterans Affairs hospitals in 1991 [18]. There are currently more than 700 hospitals enrolled in NSQIP across multiple countries [16]. NSQIP participants get Semi-Annual Reports providing detailed risk-adjusted benchmarking to similar hospitals across the program. The NSQIP Participant User File (PUF) is frequently used for research and the PUF User Guides provide detailed information about the variables collected [19]. Collection of NSQIP outcomes and use of this information to impact local care has been shown to have a positive impact in many different settings [20, 21]. One notable limitation of the NSQIP is the lack of process measures. While many sites have incorporated the measurement of these into the "custom fields" option in NSQIP, other options, such as studying process measures through merged data with a computerized physician order entry (CPOE) dataset, exist. There are a number of other quality registries available, which are typically procedure-focused. Examples include the Vascular Quality Initiative (VQI/M2S), Collaborative Endocrine Surgery Quality Improvement Program (CESQIP), Scientific Registry of Transplant Recipients (SRTR), and the American Urological Association Quality Registry (AQUA), each focused on outcomes (and sometimes process measures) specific to their field.

Two additional quality registries are notable for their focus on process measures in additional to outcomes –the AHRQ Safety Program for Improving Surgical Care and Recovery (ISCR) and the ERAS® Interactive Audit System (EIAS). A recent collaboration between the Agency for Healthcare Research and Quality and the Johns Hopkins Medicine Armstrong Institute for Patient Safety and Quality created the Improving Surgical Care and Recovery (ISCR) program. [22] The ISCR program implements evidence-based pathways and includes education to facilitate implementation as well as the collection of both process and outcomes measures. The ERAS Society is an international society with a USA Chapter [23]. Both programs focus on implementation of evidence-based practice and tracking of process measures, yet are similar to the other quality registries that rely on chart review for identification of compliance with the measures.

The National Healthcare Safety Network (NHSN) is another dataset used by many hospitals to track healthcare acquired infections (HAI). Outcomes are reported to the Centers for Disease Control and Prevention (CDC) and standardized rates are tracked. Acute care hospitals that participate in the NHSN can track multiple healthcare acquired infections, including surgical site infections after multiple specific procedures, based on ICD-10-PCS codes (https://www.cdc.gov/nhsn/faqs/faq-ssi-proc-codes.html - accessed 12/26/2018). Currently, over 17,000 facilities use NHSN to evaluate and track their progress in prevention of healthcare acquired infections. As with other data sources, understanding how occurrences are identified and nuances of the definitions are critical, particularly at your individual hospital. Hospitals may identify surgical site infections by chart review of all cases or may choose specific triggers to identify an event (such a positive cultures or readmissions). As always, understanding the definitions are critical to understanding the results, especially when investigating discrepancies in data sources. Using colon surgical site infection as an example, “stool in the abdomen, without pus or abscess” is not counted as an intraabdominal (IAB) infection in this dataset (https://www.cdc.gov/nhsn/faqs/faq-ssi.html - Q14).Variation in definitions across different datasets reflects the real-world ambiguity that can be present in monitoring outcomes. Tracking progress over time within a single data source and understanding the specifics of inclusion criteria, case identification, and outcome definitions will aid in adequate characterization of the problem and allocation of resources for individual quality improvement efforts.

The other type of data commonly used for local quality improvement projects is data available from the Computer Physician Order Entry (CPOE) and/or the Electronic Medical Record (EMR). With the implementation of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, many hospitals made the transition to electronic order entry with the additional incentive available [24]. These data sources are typically quite abundant and detailed, which can have significant limitations. Depending on the information technology (IT) resources at an individual facility, the richness of the data may prohibit the ability to easily query the database for a particular question. However, with substantial IT support, the EMR can provide data on most process measures of interest in surgical patients.

Understanding the data currently available and available in a future state is critical to assessing current performance, establishing need, informing the intervention, and tracking progress as outlined in the next section.

7.4 Identifying the Problem and Engaging Stakeholders

To begin a local quality improvement project, it’s important to both identify the area of focus and understand the data available to assess current performance. The area of focus can be chosen by many different methods. Sometimes an opportunity for improvement will be shared by hospital administrators based on governmental agency priorities (mortality, length of stay, cost, readmission, surgical site infection). For sites already participating in a data registry, it may be identified based on review of risk-adjusted benchmarked reports. Areas for improvement can also be identified based on clinical experience (e.g., provider perception of having more than expected surgical site infections). Once a topic has been chosen, it is important to assess current performance and establish a baseline. If the area of improvement was chosen based on already available data, this has been done for you! If it's a need identified by front-line providers, a conversation will need to be held with local administrators and leaders to identify data sources available to establish current performance. This data will allow you to establish the need for an intervention and create your "burning platform" for change.

Next, it is important to understand best practices around the topic you choose. Best practice guidelines are available for many topics. NSQIP has specific guidelines on their website for current participants, and the American College of Surgeons provides guidelines for management of geriatric patients [25]. Review of best practices will help identify specific processes that impact outcomes. Once process measures are identified, you must determine how you will track these. Some may require chart review, while others may be able to be pulled from the EMR. Reviewing your current processes, and any gaps in compliance with recommended processes, will inform your intervention. Finally, once an intervention has been chosen, the plan must be implemented, and progress tracked.

There are many different structures used to formalize the process described above. One of the most commonly used is the Plan-Do-Study-Act cycle. This process is elegant in its simplicity and insightful in its acknowledgement that quality improvement is an ongoing process, not a one-time activity [26]. A very similar structure is used in Six Sigma work and is referred to as the DMAIC (Define, Measure, Analyze, Improve, Control) process [27]. Most healthcare systems will have someone on the quality improvement team that is well-versed in one or more of these techniques. All the structures work well, and it is most important to have a shared language between team members. For additional education and training in the PDSA cycle, the Institute for Healthcare Improvement has excellent resources [28].

Ultimately, what will make a local quality improvement project successful is meaningful data, engagement of stakeholders, support from the organization, and an unwavering focus on the patient.

7.5 Future Directions

Quality improvement has always been critical to the care of surgical patients, but our knowledge and skill around quality improvement continues to improve. Despite the years of experience, the work is not done. As individual organizations continue their quality improvement work, risk-adjusted national comparisons will continue to raise the bar. Many quality registries have focused on 30-day outcomes, but for some surgical diseases (e.g. hernia repair or oncologic procedures) a much longer time period for tracking outcomes is likely to be most important to patients [29]. There is increasing interest in patient-reported outcomes as well as quality of life measurement [29, 30]. Outcomes after non-operative management of surgical diseases remains poorly understood [31]. Additionally, recent work to improve quality has focused on the use of video-based coaching to improve surgical technique in the operating room, and area that was previously difficult to measure [32].

In addition to novel outcomes measures, future quality improvement work will likely increasingly be supported by regional infrastructures. While quality improvement will always need to be shared and implemented on a local level, the ability of regional structures to do the work of reviewing and agreeing upon best practices and sharing that with multidisciplinary teams is of value. This has been demonstrated by the work of the Michigan Collaborative as well as Illinois Surgical Quality Improvement Collaborative. As accountable care organizations (ACOs) and clinically integrated networks (CINs) grow, these are other platforms in which data-sharing and engagement of a wide range of stakeholders may greatly enhance the ability to successfully implement quality improvement projects.

References

1. Hoyt DB, Ko CY, editors. Optimal resources for surgical quality and safety: American College of Surgeons; 2017.
2. Berwick DM. Measuring surgical outcomes for improvement: was Codman wrong? JAMA. 2015;313(5):469–70.
3. Codman EA Study of hospital efficiency: as demonstrated by the case report of the first five years of a private hospital. 1975. Joint Commission; 1996.
4. Health Resources & Services Administration. Quality improvement 2011. Available from: https://www.hrsa.gov/sites/default/files/quality/toolbox/508pdfs/qualityimprovement.pdf.
5. Campbell DA, Englesbe MJ, Kubus JJ, Phillips LR, Shanley CJ, Velanovich V, et al. Accelerating the pace of surgical quality improvement: the power of hospital collaboration. Arch Surg. 2010;145(10):985–91.
6. Guillamondegui OD, Gunter OL, Hines L, Martin BJ, Gibson W, Clarke PC, et al. Using the National Surgical Quality Improvement Program and the Tennessee Surgical Quality Collaborative to improve surgical outcomes. J Am Coll Surg. 2012;214(4):709–14.
7. Illinois Surgical Quality Improvement Collaborative. ISQIC Projects. Available from: https://www.isqic.org/isqic-projects [Accessed 19 December 2018]
8. O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. JAMA. 1996;275(11):841–6.
9. Share DA, Campbell DA, Birkmeyer N, Prager RL, Gurm HS, Moscucci M, et al. How a regional collaborative of hospitals and physicians in Michigan cut costs and improved the quality of care. Health Aff (Millwood). 2011;30(4):636–45.
10. Rosenberger LH, Politano AD, Sawyer RG. The surgical care improvement project and prevention of post-operative infection, including surgical site infection. Surg Infect. 2011;12(3):163–8.
11. Munday GS, Deveaux P, Roberts H, Fry DE, Polk HC. Impact of implementation of the surgical care improvement project and future strategies for improving quality in surgery. Am J Surg. 2014;208(5):835–40.
12. Agency for Healthcare Research and Quality. Types of Health Care Quality Measures. Available from: https://www.ahrq.gov/talkingquality/measures/types.html.

13. McDonald KM, Sundaram V, Bravata DM, Lewis R, Lin N, Kraft SA, et al. Closing the quality gap: a critical analysis of quality improvement strategies (Vol. 7: care coordination). AHRQ Technical Reviews. Rockville, MD; 2007.
14. Birkmeyer JD, Dimick JB, Birkmeyer NJ. Measuring the quality of surgical care: structure, process, or outcomes? J Am Coll Surg. 2004;198(4):626–32.
15. Lawson EH, Louie R, Zingmond DS, Brook RH, Hall BL, Han L, et al. A comparison of clinical registry versus administrative claims data for reporting of 30-day surgical complications. Ann Surg. 2012;256(6):973–81.
16. The Healthcare Cost and Utilization Project (HCUP). NIS Description of Data Elements. Available from: https://www.hcup-us.ahrq.gov/db/nation/nis/nisdde.jsp [Accessed 27 November 2018]
17. Haider AH, Bilimoria KY, Kibbe MR. A checklist to elevate the science of surgical database research. JAMA Surg. 2018;153(6):505–7. https://doi.org/10.1001/jamasurg.2018.0628.
18. Khuri SF, Daley J, Henderson W, Hur K, Demakis J, Aust JB, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. Ann Surg. 1998;228(4):491–507.
19. American College of Surgeons. ACS NSQIP Participant Use Data File. Available from: https://www.facs.org/quality-programs/acs-nsqip/participant-use.
20. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. Ann Surg. 2009;250(3):363–76.
21. Nimeri AA, Bautista J, Philip R. Reducing healthcare costs using ACS NSQIP-driven quality improvement projects: a success story from Sheikh Khalifa Medical City (SKMC). World J Surg. 2018:1–8.
22. American College of Surgeons. AHRQ Safety Program for Improving Surgical Care and Recovery. Available from: https://www.facs.org/quality-programs/iscr [Accessed 19 December 2018]
23. ERAS® USA ERAS® Society USA Chapter. Available from: http://erasusa.org/ [Accessed 19 December 2018]
24. Friedman DJ, Parrish RG, Ross DA. Electronic health records and US public health: current realities and future promise. Am J Public Health. 2013;103(9):1560–7.
25. American College of Surgeons. The coalition for quality in geriatric surgery project. Available from: https://www.facs.org/quality-programs/geriatric-coalition/resources.
26. Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed JE. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. BMJ Qual Saf. 2014;23(4):290–8.
27. Tolga Taner M, Sezen B, Antony J. An overview of six sigma applications in healthcare industry. Int J Health Care Qual Assur. 2007;20(4):329–40.
28. Institute for Healthcare Improvement. Resources. Available from: http://www.ihi.org/resources/Pages/default.aspx [Accessed 27 November 2018]
29. Fong ZV, Alvino DM, CF-d C, Nipp RD, Traeger LN, Ruddy M, et al. Health-related quality of life and functional outcomes in 5-year survivors after Pancreaticoduodenectomy. Ann Surg. 2017;266(4):685–92.
30. Arnold MR, Coakley KM, Fromke EJ, Groene SA, Prasad T, Colavita PD, et al. Long-term assessment of surgical and quality-of-life outcomes between lightweight and standard (heavyweight) three-dimensional contoured mesh in laparoscopic inguinal hernia repair. Surgery. 2018, 165:820–4.
31. Wandling MW, Ko CY, Bankey PE, Cribari C, Cryer HG, Diaz JJ, et al. Expanding the scope of quality measurement in surgery to include nonoperative care: results from the American College of Surgeons National Surgical Quality Improvement Program emergency general surgery pilot. J Trauma Acute Care Surg. 2017;83(5):837–45.
32. Greenberg CC, Dombrowski J, Dimick JB. Video-based surgical coaching: an emerging approach to performance improvement. JAMA Surg. 2016;151(3):282–3.

Part II

Emerging Areas of Research

Dissemination and Implementation Science

8

Lesly A. Dossett and Dana A. Telem

8.1 The Rationale for Dissemination and Implementation Science

Although evidence-based interventions are recognized to minimize unwarranted healthcare variation and promote best practices, trial results or guideline publication alone does not guarantee practice change [1]. Many gaps exist between evidence-based recommendations and real-world practices. The reasons for this are likely multifactorial. Providers may be unaware of recommendations, struggle with overcoming local barriers to their implementation, or even disagree with the relevance to their patient populations. As such, the intended impact of guidelines is often suboptimal, failing to lead to measurable and sustainable behavior change. Thus, the need for innovative, evidence-based dissemination and implementation strategies designed to eliminate the differences between theoretical and actual practice is clear [2].

The National Institutes of Health (NIH) defines implementation as 'the use of strategies to adopt and integrate evidence-based health interventions and change practice patterns within specific settings. Dissemination is the process of spreading information and knowledge to these settings. Implementation and dissemination (D&I) science is therefore the study of methods to promote the integrations of research findings and evidence into healthcare policy and practice. D&I exists at the intersection of health services, translational and basic science research, and the intended stakeholders, including providers. Effective implementation bridges the gap between science and practice through the development of strategic interventions aimed at overcoming identified barriers to practice change by taking individual and organizational behaviors into account as demonstrated in Fig. 8.1.

L. A. Dossett (✉) · D. A. Telem
University of Michigan, Ann Arbor, MI, USA
e-mail: ldossett@med.umich.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_8

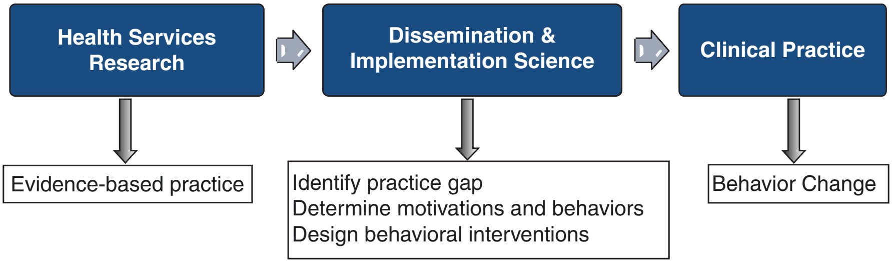

Health Services Research
Dissemination & Implementation Science
Clinical Practice
Evidence-based practice
Identify practice gap
Determine motivations and behaviors
Design behavioral interventions
Behavior Change

Fig. 8.1 Conceptual model of dissemination and implementation research

8.2 Dissemination and Implementation Terms and Definitions

One challenging aspect of D&I research is a lack of standardized terms and definitions, which vary not only by field, but by country. In this chapter we use the term D&I which is consistent with the growing field within the US. In Canada and Europe, "knowledge translation" or "scaling up" are often used to describe similar processes [3].

The subjects of D&I activities are interventions with proven efficacy and effectiveness (i.e. evidenced-based interventions, EBI). These may include programs, policies, processes or guidelines. Rogers identifies five perceived attributes of an innovation that are likely to influence the speed and extent of its adoption—(1) relative advantage, (2) compatibility, (3) observability, (4) trialability and (5) complexity [4]. Relative advantage refers to the degree to which the stakeholders view the EBI as compared to the process it replaces while compatibility is the perceived fit of the EBI into the adopting setting.

Processes for D&I include diffusion, dissemination, implementation and de-implementation. *Diffusion* is the passive, untargeted, unplanned and uncontrolled spread of new intervention. In contrast, *dissemination* is an active approach of spreading EBIs to the target audience via determined channels using planned strategies. *Implementation* is the process of putting to use or integrating EBIs within a setting. *De-implementation* is defined as stopping or abandoning practices that have not been proven effective or have proven harmful.

8.3 Conceptual Models and Frameworks

A variety of D&I models, which includes frameworks and theories, exist. A theory is a systematic way of understanding events or behaviors by indicating relationships between constructs and concepts. A theory common in D&I health research is Rogers Diffusion on Innovations theory. Frameworks guide investigators to consider constructs in systemic ways to develop and evaluate interventions. One commonly used framework in D&I research is the Consolidated Framework for Implementation Research (CFIR) [5].

The selection of a model or theory should best fit the research question by taking into account three main criteria (Table 8.1). Broadly, when considering a model one must account for whether the research will require dissemination and/or implementation, the socio-ecological level of the desired change, and the measurement constructs. Over 60 models are described for D&I research and many benefits are garnered from using an existing, validated model. Adaptations of models; however, are often required to improve the appropriateness of the selected model to the population or setting of interest. D&I models are dynamic entities and should be refined based on relevant factors needed for delivery of evidence based interventions. For instance, if planning for a technology based intervention, considerations to mechanisms such as availability of internet connection, technical ability of the target population, and availability of technology are critical.

Examples of commonly used frameworks in public health include, but are not limited to, the Framework for Reach, Effectiveness, Adoption, Implementation and

Table 8.1 Dissemination and implementation model selection based on individual criteria

Category	Socio-ecological level	Measurement construct(s)
Dissemination and implementation	Individual	Acceptability/feasibility
Implementation only	Community	Adopter/decision maker characteristics
Dissemination only	Policy	Awareness
	Organization	Champion/field agent
	System	Compatibility
		Context
		Intervention development
		Engagement
		External validity/generalizability
		Fit
		Identification
		Knowledge and knowledge synthesis
		Maintenance and sustainability
		Pre-implementation
		Reach
		Outcomes—health, clinical, satisfaction
		Strategies
		Adaptation and evaluation
		Adoption
		Barriers and facilitators
		Dose
		Cost
		Evaluation
		Fidelity
		Goals

Maintenance (RE-AIM) which is a dissemination and implementation framework targeting change at the individual, organizational and community level; the Precede-Proceed; Dynamic Sustainability; Practical; Robust Implementation Sustainability Model (PRISM) which represents and implementation dominant framework targeting the individual and organizational level; the Consolidated Framework for Implementation Research (CFIR) which is an implementation only framework addressing organizational and community change; and the Theoretical Domains Framework (TDF) which is an excellent instrument to comprehensively understand and assess the facilitators and barriers of change at the individual level. The NIH provides an excellent and comprehensive online tool to aid researchers in selecting an appropriate model and can be found at: https://prevention.nih.gov/research-priorities/dissemination-implementation.

8.4 Study Design in D&I Research

D&I studies are the last stage of research in the science to practice continuum, preceded by the efficacy and effectiveness studies that are distinct and address different questions than D&I studies which focus on making programs work. Critical in understanding this distinction is the concept of implementation strategies—typically multi-level and multi-component on and within a delivery system that aim to increase the adoption of an EBI into routine use or care. Implementation strategies are broadly categorized into nine domains—engage consumers, use evaluative and iterative strategies, change infrastructure, adapt and tailor to the context, develop stakeholder interrelationships, utilize financial strategies, support clinicians, provide interactive assistance and train and educate stakeholders. It is rare that a single discrete domain would be used; rather a package of components is typically selected to address multiple barriers to implementation.

Several distinct phases exist within D&I including exploration, adoption/preparation, implementation and sustainment and these phases typically correspond to distinct research questions. The process of implementing new EBI in the real world can be complex and involve numerous decisions on the part of the evaluators and key decision makers in the system. The exploration and adoption/preparation phases (also called pre-implementation) include selection of the EBI to be used, the target stakeholders, the expected reach of an EBI within a particular setting, and other resources being allocated. Given this range of options and decisions for implementers, the pre-implementation phase offers a number of study design options to evaluate needs and capacities for different alternatives including decision analysis, economic analysis and simulation. In the adoption/preparation, implementation and sustainment phases, there are three main categories of designs based on the type of comparison—(1) within-site design, (2) between-site design and (3) within- and between-site design.

8.4.1 Hybrid Designs

The concept of hybrid designs, which combine effectiveness and implementation, was proposed as a way to more quickly move interventions into routine adoption. The concept also encourages consideration of building intervention for implementation as early as possible by seeking end-user input into the intervention before its effectiveness is completely known. When considering hybrid designs, the “intervention” refers to the clinical or prevention practice while “strategy” refers to the implementation-support activities or tools. Curran and colleagues originally proposed three types of hybrid designs—type 1, 2 and 3. Type 1 test intervention effectiveness while gathering information on implementation and Type 3 tests implementation strategies while gathering information on intervention effectiveness. Type 2 tests the intervention while also studying an implementation strategy. In these designs, the emphasis on intervention effectiveness outcomes is strongest in Type 1 and becomes a secondary objective in Type 3, whereas implementation strategy is a secondary aim in Type 1 and becomes the focus in Type 3.

8.5 Measurement in D&I Research

The constructs for measurement in D&I research are derived from conceptual models and these models distinguish between the evidenced based interventions (EBI) being introduced and the dissemination and implementation strategies being used to introduced them. D&I outcomes serve as intermediate outcomes proximal to the changes in behavior or health the EBI is intended to produce. Distinguishing D&I effectiveness from EBI effectiveness is critical—if interventions fail to produce the desired outcome we must be able to determine if this is a result of an ineffective intervention or if an effective intervention was poorly deployed.

The most commonly used taxonomy of outcomes was developed by Proctor and colleagues which puts forward eight conceptually distinct implementation outcomes—*acceptability, appropriateness, feasibility, adoption, fidelity, implementation cost, penetration and sustainability* [6]. *Acceptability* is the perception among stakeholders that a given intervention is agreeable, palatable or satisfactory. It should be assessed based on the stakeholder’s knowledge or experience with an EBI, and can be measured from the perspective of multiple stakeholders. *Appropriateness* is the perceived fit, relevance or compatibility of the EBI for a given practice setting, provider or consumer. Although it is conceptually distinct to acceptability, the two outcomes are rarely meaningfully different—that is, rarely would an EBI be acceptable but not appropriate and visa versa. *Feasibility* is the extent to which a new program or policy can be used or carried out within a certain setting or population. This outcome is often considered in retrospect as a reason for implementation success or failure, often reflecting the feasibility of required implementation resources. Adoption is the intention, decision or action to try and employ an EBI. It is often referred to as “uptake” and can be measured at the either the organizational or individual level. *Fidelity* is the degree to which and EBI was implemented as originally intended. It is typically described in terms of adherence to the programs protocol, the amount of the program delivered and the quality of delivery. *Implementation cost* is the incremental or additional expense of implementing the EBI and is most useful when comparing multiple D&I strategies. *Penetration* is the integration of a practice within a setting, for example, the number of persons using a service divided by the persons eligible for the service. Penetration can also be measured at the level of the provider. Sustainability is the extent to which a newly implemented EBI is maintained or institutionalized.

When considering implementation outcomes several other factors are important to consider including (a) the level of analysis, (b) the salient D&I phase and (c) the measurement approach. The *level of analysis* is important for its implications for the measure's target stakeholders, source of data and unit of analysis. These include analysis at the level of the individual, organization or policy. The phase of D&I for which an outcome may be measured includes the exploration, preparation, implementation, or sustainment phases. Some outcomes may need to be measured in multiple phases of the implementation. Finally, a variety of measurement approaches may be used including qualitative interviews or focus groups, quantitative surveys, mixed methods approaches, record archives and administrative data.

References

1. Balas EA, Boren SA. Managing clinical knowledge for health care improvement. Yearb Med Inform. 2000:65–70.
2. Fisher ES, Shortell SM, Savitz LA. Implementation science: a potential catalyst for delivery system reform. JAMA. 2016;315:339–40.
3. McKibbon KA, Lokker C, Wilczynski NL, et al. A cross-sectional study of the number and frequency of terms used to refer to knowledge translation in a body of health literature in 2006: a tower of babel? Implement Sci. 2010;5:16.
4. Rogers E. Diffusion of innovations. 5th ed. New York: Free Press; 2003.
5. Damschroder LJ, Lowery JC. Evaluation of a large-scale weight management program using the consolidated framework for implementation research (CFIR). Implement Sci. 2013;8:51.
6. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Admin Pol Ment Health. 2011;38:65–76.

Studying Organizational Culture in Surgery

9

Nikhil Panda and Alex B. Haynes

Abstract

Organizational culture encompasses the shared ethos of a group of individuals that guide how they achieve a common goal. The study of organizational culture originated in industries outside of medicine and surgery; however, there is a growing focus on studying the culture within complex organizations in the delivery of safe and high-quality patient care. In this chapter, we review methods to study and understand organizational culture, specifically safety culture, in surgery and highlight potential impact on patient outcomes.

9.1 Introduction

Organizations have long been noted to have an ambiguous, yet specific culture, reflecting "the customary and traditional way of thinking and doing things [1]." Perhaps because organizational culture is in itself conceptually enigmatic, or has been historically challenging to quantify as a single metric, studies of and tools for measuring organizational culture are relatively contemporary. Originating outside of medicine and surgery in the business fields of industry, manufacturing, and commerce, careful evaluations of culture quickly revealed its critical role in the ability

N. Panda, MD
Department of Surgery, Massachusetts General Hospital, Harvard Medical School,
Boston, MA, USA

Safe Surgery Program, Ariadne Labs, Boston, MA, USA

A. B. Haynes, MD, MPH (✉)
Safe Surgery Program, Ariadne Labs, Boston, MA, USA

Department of Surgery and Perioperative Care, Dell Medical School,
University of Texas Austin, Boston, MA, USA
e-mail: alex.haynes@austin.utexas.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_9

of an organization to perform effectively and efficiently. When strong, organizational culture can be a crucible of innovation; when weak, it may represent the greatest liability to progress [2]. These reflections have led to an even greater emphasis on the study of organizational culture in many fields, especially health care. How does one define ‘good’ or ‘strong’ culture within a health care organization or hospital system? How does culture evolve and change over time?

9.2 Organizational Culture in Health Care

While the study of culture originated outside of medicine and surgery, modern health care is delivered within a complex system and health care organizations manifest their own unique cultures. The interactions of the delivery system (itself comprised of a blend of institutions, individual providers, patients, and other stakeholders) with the payment and regulatory systems can have a profound influence on the delivery of care. The structure and organization of health care delivery systems has evolved over time and with it, the understanding of how clinical processes and outcomes are affected by these structures. Surgical care delivery is no exception. The most illustrative example is the environment of the operating room, which is among the most complex in the hospital. Here, a multidisciplinary team assembles to provide high-acuity patient care. This team may consist of individual clinicians who work together regularly, or who may be meeting each other for the first time at the beginning of or even during a surgical procedure. New members are brought on board on a regular basis and integrated into the workflow as other members exit. The intensity of the surgical care delivery in the operating room rarely provides an opportunity for any of these individuals to parse out the beliefs, values, and preferences of the others. The “glue” that holds this system together is culture.

There is a nascent, but growing body of literature that describes the importance of certain organizational cultural characteristics in regard to the delivery of consistent, high-quality surgical care. This can include factors related to hospital leadership and its vision, including board composition and prioritization of initiatives. It may relate to the characteristics of individual leaders within the surgical divisions or entire health care systems. There are factors associated with the ways that management flows from these upper level leaders to the front-line clinical workers. Among care teams in the operating room and elsewhere, there are cultural factors related to the way that members of these teams interact and care is delivered. As academic surgeons interested in health services research, an understanding of organizational culture can be helpful when investigating variation in the quality and delivery of surgical care, the processes of implementation or de-implementation, the introduction of safety initiatives, patient and provider experiences, and a host of other areas of study. In order to integrate the study of this culture into an individual’s academic portfolio, one must first understand what surgical organizational culture describes, the tools available to measure culture within organizations, and the potential benefits in terms of patient outcomes.

9.3 What Is Surgical Organizational Culture?

Surgical organizational culture can refer to many things, from ethnographic observations of the interpersonal psychosocial dynamics of surgeons and surgical trainees to norms of communication and practice in hospitals. For the purposes of this chapter, we will be focusing on those aspects of culture that can be attributed to the organization itself and studied in an empiric fashion. In particular, the focus here will be on the culture surrounding the processes of direct care provided to patients. Often, this is defined specifically as safety culture. The definition of safety culture adopted by the Agency for Healthcare Research and Quality (AHRQ), derived from the nuclear power industry, is, "the product of individual and group values, attitudes, perceptions, competencies, and patterns of behavior that determine the commitment to, and the style and proficiency of, an organization's health and safety management. Organizations with a positive safety culture are characterized by communications founded on mutual trust, by shared perceptions of the importance of safety, and by confidence in the efficacy of preventive measures [3]."

One can sense from this definition that safety culture, like other dimensions of organizational culture, is generally vague, referring to the often intangible and nearly unquantifiable drivers of an organization's behaviors. Amir Ghaferi, a leader in the field of organizational culture, describes culture as the, "forces in an organization that operate in the background – it's 'context' [4]." Therefore, surgical safety culture is understandably challenging to measure. How does one quantify a group's set of values, how these values drive certain actions rather than others, and how this evolves? Additionally, and similarly difficult, is designing and implementing initiatives to change culture, as this requires changing a group of individuals' behaviors. In the next section, we introduce several methods that have been developed to measure safety culture and ways these have been used to improve patient outcomes.

9.4 Qualitative and Quantitative Tools for Measuring Safety Culture

Approaches to studying organizational culture can include both qualitative and quantitative methodologies. Qualitative research typically involves ethnographic field observations, interviews, focus groups, among other techniques for understanding the interactions of the complex surgical ecosystem. There are many examples of this type of work in health care and specifically in surgery, ranging from Charles Bosk's seminal book, *Forgive and Remember* [5], to more contemporary studies of implementation of safety programs, such as Mary Dixon-Woods' investigation of how culture in English intensive care units shaped the ability to implement a central line safety initiative [6]. These approaches can yield incredibly vivid and specific explanations of organizational culture and subcultures, and can be especially powerful when linked to empiric studies. The skillset and resources needed for this work are possessed by relatively few surgeon-scientists, but is more commonly achieved through collaborations with other experienced researchers with backgrounds in sociology, anthropology, and qualitative evaluation. These interdisciplinary collaborations can be very fruitful, not only for the research output, but also for the experience with qualitative methodology gained by the surgical researcher.

While qualitative methods can yield granular data and rich content, they are often impractical to utilize on a large scale or on repeated basis. It can also be difficult to empirically correlate the findings of this type of work with quantitative datasets on surgical processes and outcomes. This has created an appetite for instruments that can be repeatedly administered and compared between organizations in an empiric fashion. One such tool is the Safety Attitudes Questionnaire (SAQ), an AHRQ-funded instrument developed and validated by Brian Sexton, Eric Thomas, and Robert Helmreich [7]. The SAQ is partially derived from a tool used in the aviation industry. Modified for the healthcare environment, it is comprised of 60 items scored on a five-point Likert scale spanning six domains: (1) Teamwork Climate, (2) Safety Climate, (3) Perceptions of Management, (4) Job Satisfaction, (5) Working Conditions, and (6) Stress Recognition. The tool has been tested in a variety of healthcare environments, including intensive care units, inpatient wards, ambulatory clinics, and operating rooms, and has demonstrated good psychometric properties. There is also a 36-item short form, as well as specific versions for the operating room and labor and delivery suite, although the formal testing of these modifications has been less rigorous.

Another similar tool is the Hospital Survey on Patient Safety (HSOPS), developed directly by AHRQ [8]. Similarly composed of Likert-scored items and responses, the HSOPS has 42 items grouped into 12 composite measures. This tool is intended for the hospital-setting, although there are related surveys for other settings including the outpatient clinic and ambulatory surgical units. Other surveys, such as that developed by Sara Singer and colleagues, look specifically at some of the actions that frontline workers perceive as creating a climate of safe care delivery [9].

9.5 Safety Culture and Surgical Outcomes

Understanding the safety culture of surgical organization and the ability to measure this empirically is only a first step. There is now a growing body of literature that links safety culture with outcomes after surgery. For example, one study used the HSOPS tool in seven hospitals participating in a colorectal surgery improvement project, finding that there was an association between surgical site infection rate and perceived safety culture in 9 of 12 domains of the survey [10]. Another study, embedded within the Safe Surgery 2015 South Carolina collaborative, identified an association between perception of safe surgical practice using the instrument developed by Singer et al., with risk-adjusted postoperative mortality after inpatient surgery [11]. For each one point difference (on a 7 point Likert scale) in the domains of respect, clinical leadership, and assertiveness, there was a 14–29% difference in risk-adjusted mortality. Another recent study looked at a single institution's changes in safety culture, as measured by a modified version of the SAQ, after a program to improve culture and outcomes. The authors found that there was a sustained improvement in SAQ scores after implementation of this program, which paralleled an improvement in outcomes reflected in National Surgical Quality Improvement Program (NSQIP) measures [12].

9.6 Next Steps: Opportunities for Health Services Researchers

These studies represent the insights that can be gained from studying and understanding safety culture and surgical outcomes. There is much further work that can be done to examine this phenomenon in a more granular fashion. Identifying links between processes of care delivery and safety culture could help to understand the reasons for failure to rescue, thought to be a major contributor to variations in postoperative mortality [13]. Attention is needed to develop scalable interventions to improve the safety culture within diverse surgical organizations with various levels of resources, such that patient outcomes are improved. This will necessitate engagement with implementation science, communication, education, and others along with the study of culture. Mixed-methods approaches will be particularly valuable in these studies, and represents a critical opportunity for academic surgeons to lead investigations.

Additionally, there are opportunities to draw from the studies in other industries, such as business and education, to better understand other areas of culture beyond safety within health care organizations. For example, the World Management Survey group is evaluating management culture and hospital performance, including linking this aspect of organizational culture with clinical processes and patient outcomes, mirroring previous work been done in manufacturing [14]. Finally, and perhaps the area of greatest potential, is introducing opportunities for continuing education and training in organizational culture to various front-line stakeholders (e.g., clinicians, trainees, and other staff) within complex health care systems.

9.7 Conclusion

Ultimately, studying and understanding the various dimensions of organizational culture, coupled with intentional efforts to improve it, can lead to innovation that will improve the quality and safety of surgical care delivery. For the surgeon-scientist focusing on health services research, much further work is needed to identify scalable tools to measure surgical cultures and subcultures, design interventions to enhance culture and educate others, and study the complex interplay between surgical processes, culture, and patient outcomes.

References

1. Jaques E. The changing culture of a factory. London: Tavistock Publications; 1951. p. 251.
2. Groysberg B, Lee J, Price J, Yo-Jud Cheng J. The Culture Factor [Internet]. Harvard Business Review. Available from: https://hbr.org/2018/01/the-culture-factor [cited 2019 May 9].

3. Factors SG on H. Organising for safety: Third report of the ACSNI (advisory committee on the safety of nuclear installations). England: Sudbury; 1993.
4. Ghaferi AA. Success in academic surgery. In: Dimick JB, Greenberg CC, editors. Available from: http://www.springer.com/series/11216 [cited 2019 May 9].
5. Bosk CL. Forgive and remember: managing medical failure. Chicago: University of Chicago Press; 1979.
6. Dixon-Woods M, Leslie M, Tarrant C, Bion J. Explaining matching Michigan: an ethnographic study of a patient safety program. Implement Sci. 2013;8(1):70. https://doi.org/10.1186/1748-5908-8-70. [cited 2019 May 9]
7. Sexton JB, Helmreich RL, Neilands TB, Rowan K, Vella K, Boyden J, et al. The safety attitudes questionnaire: psychometric properties, benchmarking data, and emerging research. BMC Health Serv Res. 2006;6(1):44. https://doi.org/10.1186/1472-6963-6-44. [cited 2019 May 9]
8. Hospital Survey on Patient Safety Culture. Agency for Healthcare Research &; Quality [Internet]. Available from: https://www.ahrq.gov/sops/surveys/hospital/index.html [cited 2019 May 9].
9. Singer SJ, Jiang W, Huang LC, Gibbons L, Kiang MV, Edmondson L, et al. Surgical team member assessment of the safety of surgery practice in 38 South Carolina hospitals. Med Care Res Rev. 2015;72(3):298–323. Available from: http://journals.sagepub.com [cited 2019 May 9]. https://doi.org/10.1177/1077558715577479.
10. Fan CJ, Pawlik TM, Daniels T, Vernon N, Banks K, Westby P, et al. Association of Safety Culture with surgical site infection outcomes. J Am Coll Surg. 2016;222(2):122–8. https://doi.org/10.1016/j.jamcollsurg.2015.11.008.
11. Molina G, Berry WR, Lipsitz SR, Edmondson L, Li Z, Neville BA, et al. Perception of safety of surgical practice among operating room personnel from survey data is associated with all-cause 30-day postoperative death rate in South Carolina. Ann Surg. 2017;266(4):658–66. Available from: http://insights.ovid.com/crossref?an=00000658-201710000-00014 [cited 2019 May 9]
12. Babic B, Volpe AA, Merola S, Mauer E, Cozacov Y, Ko CY, et al. Sustained culture and surgical outcome improvement. Am J Surg. 2018;216(5):841–5. Available from. https://doi.org/10.1016/j.amjsurg.2018.02.016.
13. Ghaferi AA, Dimick JB. Importance of teamwork, communication and culture on failure-to-rescue in the elderly. Br J Surg [Internet]. 2016 .[cited 2019 May 9];103(2):e47–e51. Available from: doi:https://doi.org/10.1002/bjs.10031
14. World Management Survey | Benchmark your manufacturing firm, hospital, school, or retail outlet against others in your country, industry or size class [Internet]. Available from: https://worldmanagementsurvey.org/ [cited 2019 May 9].

Patient Reported Outcomes in Surgical Research

10

Larissa K. F. Temple and Arden M. Morris

Abstract

Patient reported outcomes (PROs) are defined as the outcomes of treatment reported directly by patients and thereby reflect the results of treatment that are most meaningful to patients. As a critical component of Patient Centered Care, PROs are increasingly important to clinical care, quality assessment and improvement, and health services research (HSR). Most of the formative assessment of PROs focuses on accuracy and validity—most importantly, does the instrument measure what it is intended to measure? Does it comprehensively measure the critical domains in the area of inquiry, is it as parsimonious as possible to reduce respondent burden, do the results generalize to the represented population? In short, how close does the PRO come to truly answering the research question? The current chapter explores such questions and offers keys to selection and management of PROs, establishment of internal and external validity, and interpretation of findings.

10.1 Patient Reported Outcomes: A Critical Component of Patient Centered Care

Patient reported outcomes (PROs) are defined as the outcomes of treatment reported directly by patients and thereby reflect the results of treatment that are most meaningful to patients. As a critical component of Patient Centered Care,

L. K. F. Temple
Division of Colorectal Surgery, Department of Surgery, University of Rochester Medical School, Rochester, NY, USA
e-mail: Larissa_Temple@URMC.Rochester.edu

A. M. Morris (✉)
S-SPIRE Center, Department of Surgery, Stanford University Medical School, Stanford, CA, USA
e-mail: ammsurg@stanford.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_10

Table 10.1 A typology of patient reported outcomes

Category	Dimensions	Exemplar instrument
Health related quality of life	Well-being, function, role	SF36
Symptoms	Pain, fatigue	Visual acuity scale
Patient reported experiences	Satisfaction with surgical care	S-CAHPS
Health behaviors	Harmful or beneficial activities	CAGE questionnaire

SF-36: Short Form 36; *S-CAHPS*: Consumer Assessment of Healthcare Providers and Systems Surgical Care Survey; *CAGE*: Cut-Annoyed-Guilty-Eye

PROs are increasingly important to clinical care, quality assessment and improvement, and health services research (HSR). PROs can be categorized traditionally as health related quality of life, symptoms, experiences of care (*i.e.*, Patient Reported Experiences), and as health behaviors (Table 10.1). All are meant to directly reflect the patient perspective and to provide a counter-balance to the more dominant clinician perspective in the medical record or the investigator perspective in the research record. In extensive previous work among patients with cancer, for example, clinician-reported patient symptoms are typically downgraded in frequency and intensity relative to patient-reported symptoms. However, when randomized to an intervention that allowed regular and direct patient reporting of symptoms, patients required significantly fewer emergency department visits and unplanned hospitalizations and were able to remain on chemotherapy longer than patients in the usual care control group [1]. These findings indicate that understanding patient experiences and perspectives without a clinician/investigator filter may have substantial clinical implications. PROs should affect how we weigh treatment effectiveness against toxicity in clinical trials and in HSR. In particular, the power of merging PROs with clinical data embodies the fundamental goals of patient centeredness. Ultimately, without capturing PROs, it is not possible to understand comprehensively the value of surgical care to the end users - that is, to the patients who undergo it.

In clinical and health services research over the past 10 years, methods to measure PROs have seen tremendous advancement. The National Quality Forum has outlined a hierarchy of methodological rigor for selection of PROs (Fig. 10.1) [2]. At their most simple, PROs are received directly from patients and need not have any formal measurement properties, as demonstrated by Likert scale questions often appended to studies. By contrast, PRO-Measures (PRO-Ms) require establishment of psychometric properties to demonstrate reliability and validity and can (should) be used in research, but first must undergo more formalized interpretation by developers. PRO-Performance Measures (PRO-PM) should demonstrate the ability to delineate differences in PROs based upon provider and/or hospital systems. PRO-PMs must be even more rigorously developed and tested given the intended purpose of differentiating quality of care and potentially used for accountability.

Measures of PROs, and patient reported experiences under the PRO umbrella, capture many domains and may be generic and/or specific to an outcome related to an intervention and/or a specific disease. Generic measures (i.e. symptoms, function, quality of life) permit comparison among diverse diseases and establishment

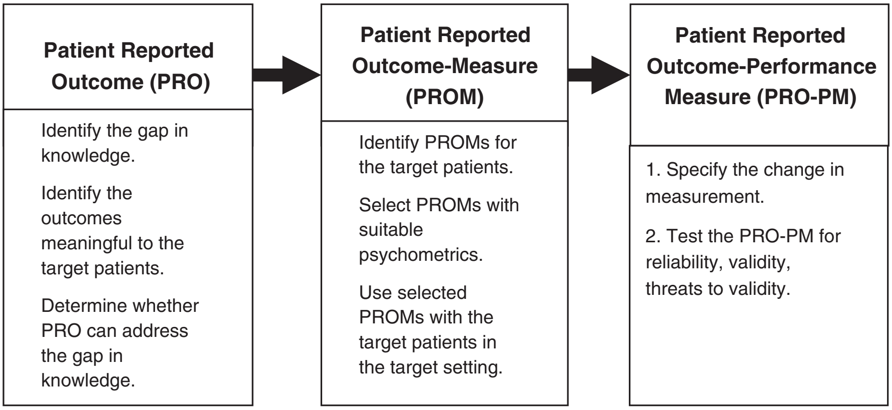

Patient Reported Outcome (PRO)
Identify the gap in knowledge.
Identify the outcomes meaningful to the target patients.
Determine whether PRO can address the gap in knowledge.
Patient Reported Outcome-Measure (PROM)
Identify PROMs for the target patients.
Select PROMs with suitable psychometrics.
Use selected PROMs with the target patients in the target setting.
Patient Reported Outcome-Performance Measure (PRO-PM)
1. Specify the change in measurement.
2. Test the PRO-PM for reliability, validity, threats to validity.

Fig. 10.1 A Hierarchy of Methodological Rigor for Selection of Patient Reported Outcomes. *Adapted from the National Quality Forum (www.qualityforum.org)*

of construct validity. Unique outcomes associated with disease processes and/or treatments, however, may not be detected without disease-specific measures.

10.2 Development of Patient Reported Outcomes

The first step in design of a study or program that includes PROs involves decision making about the use of existing scales versus development of new measures. Such a determination is informed first by the study question and design but also must take into account pragmatic realities related to time, financial costs, and other research team resources. Use of existing measures decreases the planning and preparation time, facilitates sample size calculations, and enables comparisons of data. However, if the pre-existing measure that has been selected is not responsive to the actual research question, the data will be meaningless. While development of a new measure may ensure that the items included are responsive to the research question, measure development is time intensive and laborious, and the new measure will require extensive validation before it can be appropriately incorporated into a study as a primary outcome. Therefore, investigators must carefully consider the financial and opportunity costs to the use of newly developed vs. established measures.

Regardless of whether PROs are to be developed or selected, preparation for a PRO study begins with an exhaustive literature search of existing measures as well as non-validated measures. If a PRO is to be developed, patient engagement from the time of inception is critical—and sometimes under-appreciated. Ideally, patients should be involved in every step of the PRO or PROM development process in order to ensure relevance of the question and outcome, fidelity to the patient perspective, and comprehensibility of the instrument, and to assist in interpretability by the investigator (Table 10.2). All too often, however, the study team lacks resources and time for an ideal PRO-M development process. A recent study of PRO-M developers

Table 10.2 Steps in development of patient reported outcomes

Action	Purpose	Participants
Question development	Motivate the study	Research team and stakeholders
Literature review	Identify and review existing measures	Research team
Develop domains document	Map the areas of inquiry	Research team and stakeholders
Qualitative interviews or focus groups	Ensure complete capture of important domains	Research team, stakeholders, patients
Question development	Generate a preliminary draft of the survey	Research team and stakeholders
Cognitive testing and survey refinement	Ensure readability, comprehensibility, fidelity to domains	Research team and patients with the target condition
Item reduction	Minimize respondent burden	Research team

suggested that more than 25% included no patient input whatsoever [3]. Among those that did, patients were primarily involved through interviews or focus groups seeking feedback only on item development and comprehensibility. Engagement of the broader community of stakeholders is also critical to development of PROs. Depending upon the study question and intended purpose of the PRO, stakeholders may include patients, physicians, health care systems and/or payers.

As a next step in measure development after the literature review, many research teams will create a document of the domains to be included in the PRO-M drawn directly from the literature search, the underlying conceptual model, research question, and patient and stakeholder input. Items for a survey can be developed directly from such a "domains document", often taking the form of a Likert or Likert-type scale. Likert and Likert-type scales are ordinal scales that measure agreement with a statement or the intensity of an experience on a linear continuum. The Likert-type scale for patient reported outcomes or survey items is popular due to ease of comprehension by the respondent and analysis by the research team. After creating a preliminary draft of items for the PRO or PRO-M, investigators ascertain inclusion of all relevant domains by conducting qualitative interviews to collect additional items from a representative sample of patients for whom the PRO is intended. Then, after the identification of items and question creation, internal validity should be assessed through cognitive interviewing of patient respondents. In this way, investigators can ensure that patients' interpretations of the items in the PROM truly reflect the questions of the study team. Finally, some form of item reduction, such as factor analysis [4], should be performed to minimize respondent burden.

While laborious and expensive, the development of a PRO may provide tremendous advances in understanding surgical outcomes under the right circumstances. Moreover, rapid advances in the development of PROs within the last 10 years have eased the burden on research teams and patient respondents [5]. In addition to semi-structured interviews, item generation can be augmented by crowdsourcing, which is increasingly used in an online format [6].

10.3 Pre-Existing Patient Reported Outcomes

Given the effort required for development, many investigators instead ultimately select PROs that have been previously developed, tested, and published. PRO researchers may be tempted to pull items from different PROs to create a piecemeal measure; however, the psychometric validity of an extant PRO only applies to use of the measure as it was developed. An instrument with variations in the order or wording of items may perform differently from the original. As well, the items in measures developed with item response theory (IRT) [7] may be "pulled" from a statistical perspective but few studies that have examined how to score these items in isolation. Therefore, whenever possible, we recommend use of existing measures as they were designed. As well, research teams are wise to address, up front, any copyright issues and to obtain the scoring manual before implementing a PRO or PRO-M.

As noted above, if existing PROs are available and appropriate to the research question, the researcher must make explicit decisions about whether to use generic or disease-specific PROs. Disease- specific PROs focus on the disease and patient population of interest and, therefore, provide data that is germane to affected patients and their providers. By contrast, generic measures are useful to compare across populations and treatments, to compare different populations across a similar treatment, and to establish benchmarks. Initiatives like PROMIS (http://www.healthmeasures.net/explore-measurement-systems/promis), suggest a broad desire to use generic measures to evaluate disease states. Alternatively, initiatives like the International Consortium for Health Outcomes Measurement (www.ichom.org) demonstrate the validity of disease-specific measurement, which may be more actionable for specific patients but don't have the same potential for quality or performance improvement that generic measures have. Depending on the research goals, many PRO researchers use a combination of generic and disease specific measures. Similar to factor analysis noted above, item reduction and validation can be facilitated by use of innovative psychometric techniques such as IRT, which models the relationship between item responses and latent constructs to develop the most parsimonious, least burdensome PRO-M instrument.

10.4 Psychometric Issues

The psychometric properties of a PRO speak to its soundness or internal validity, which is critical to the collection of meaningful data [8]. The psychometric properties of the measure are most likely to be optimized using a mixed methods approach with both qualitative and quantitative research methods. A qualitative approach is particularly helpful to establish face and content validity. Face validity simply indicates the degree to which the PRO actually measures what it claims to measure; content validity indicates the degree to which all the relevant items based on underlying theoretical concepts are included in the PRO. Of note, several measures have been developed using items from long standing "legacy measures" based on the argument that the items have been vetted previously and therefore face and content validity have been established. It is possible, however, that legacy measures may be missing important items and/or updates are warranted.

Several important psychometric components of PROs must be addressed quantitatively. Reliability refers to the measure's ability to distinguish true differences between groups and/or changes over time. Establishment of the simplest form of reliability, test-retest reliability, is a minimum standard in measure development. Specifically, test-retest reliability involves application of the PRO on more than one occasion when responses are expected to be consistent and determination of stability of responses with Crohnbach's alpha. Criterion validity is an assessment of a PROs actual performance and measured two ways. First, concurrent validity refers to a measure's relationship to benchmark instruments that purport to measure the same construct or domains. Ideally results from a newly developed measure correlate closely with results from a gold standard (or criterion) measure if one exists. Second, predictive validity indicates how well a measure deployed at one point in time can predict results at a future point in time.

Construct validity, indicating how well a PRO captures a theoretical trait, is a more complex psychometric property. While content validity described above refers to whether the content of a measure reflects the important concepts of a trait, construct validity refers to whether a measure accurately reflects the central theoretical phenomenon under study and does not inadvertently measure other related phenomena. Establishment of construct validity is not possible with a single test. Instead, it requires a body of work to be developed that demonstrates the utility of the PRO. The most important two considerations within construct validity are convergent validity—indicating that constructs that are expected to be related actually are related—and discriminant validity—indicating that two constructs that are unrelated actually are unrelated. In testing discriminant validity, the researcher also must evaluate how the measure distinguishes between groups of patients, for example by comparing scores of patients believed *a priori* to be different. Finally, depending on the PRO, it is important to show that the PRO is sensitive to change over time.

The majority of PRO-Ms are still developed using classic psychometric theory. It is important, however, to appreciate that there are newer psychometric methods for PRO development, such as previously mentioned IRT and Rasch modeling [7]. As noted, both are useful in computer adaptive testing which enables surveys to increase precision without adding questions. Unlike classical techniques, reliability and validity in this context rests on the individual items rather than the overall instrument. By contrast to IRT, which doesn't calibrate between the items, Rasch techniques employ a hierarchical calibration. Although the statistical theory behind IRT and Rasch differ, both are designed to optimize precision and minimize survey burden, thereby increasing response rates.

Ultimately, in developing or choosing a measure, PRO researchers must accept that at least some psychometric properties will be imperfect and as in all research, compromises are unavoidable. The study team decisions and assumptions therefore should be made explicit and contingencies considered for the possibility that the

PRO does not work as expected. As well, it is important to note that sometimes the null hypothesis cannot be rejected; that is, the hypothesized outcome differences may not exist between the subject groups.

10.5 External Validity

The external validity of a PRO refers to how well the results can be generalized to a population of interest, other settings, or other points in time. The most important first step to optimize external validity, or generalizability, occurs during selection of respondents. The subject group should be compared in multiple ways against the target population to ensure similarity, or alternatively, can be selected from the target population in a random fashion. The research team should always determine a sampling strategy explicitly and in advance of instrument deployment, accounting for financial and other resources.

The most important second step to optimize external validity is to achieve adequate response rates. Even the best sampling strategy can be compromised by poor response rates. Poor response rates threaten external validity because of potential differences between those who respond and those who do not. In this way, although the sampling strategy may identify an outstanding representation of the population of interest, the actual data collected may not.

Pre-testing the content and format of a PRO-M is critical to ensure the best possible survey design and usability. In general, once a patient starts a survey, she is likely to finish the survey. Advance planning by the study team can help in several ways. Patients that are reporting outcomes should feel that the PRO items are important and have the potential to improve their care—the survey should reflect what the patient is experiencing. Particularly in clinical care, opportunities for respondent feedback to the research team will increase patients' willingness to complete a PRO. For example, embedding the PRO data collection and discussion of the results with patients as part of their clinical care has been shown to improve patient satisfaction with the visit [1]. More recent data demonstrate that pre-notification and/or incentives may no longer be as important as they once were and that appropriate incentives for specific cohort may vary [9, 10].

Historically, survey response rates have been optimized using Dillman strategies with repeated attempts at contact through several modes [11]. Paper and pencil surveys, for example, appear to be more highly endorsed with a personalized letter, incentives, and a planned strategy for repeated mailings. In the current era of information overload, surveys must stand out—and response rates for electronic surveys depend on user-friendly formatting and personalization. PRO-Ms should take less than 10 minutes to complete and should have a very easy user interface. They should be adaptable to completion on various electronic modalities. Alternative back-up modes (i.e. telephone, paper) should be considered to improve data collection from otherwise under-represented groups. Using available technology to decrease the respondent burden will likely improve data collection.

10.6 Interpretation of Patient Reported Outcomes

Interpretation of the data is the last major task in a PRO research project—this is the critical transition between acquiring data and generating new knowledge. In their simplest form, PROs have been reported as mean scores and absolute or relative differences between groups. CONSORT guidelines [12] acknowledge that for comparison across studies, absolute differences may be more meaningful than relative differences because absolute differences do not rely on baseline measures, which may be different among groups.

Presentation of mean scores has been supplanted more recently by the concept of minimum clinically important difference (MCID), based on important work from early evidence-based medicine researchers [13]. The minimum clinically important difference refers to the smallest difference between scores that patients themselves perceive to be beneficial or harmful. The MCID can be determined in two ways—the distribution method based on the standard deviation (*e.g.*, standardized effect sizes) among measurements or the anchor-based method which relies on comparison against other previously established benchmarks (*e.g.*, generic quality of life measures). Each method has advantages—for example, the distribution method is relatively quick and easy to calculate—however, only the anchor-based method considers the patients' perceptions and therefore is preferred by the PRO research community. To encompass the spectrum of patient experiences and perceptions, researchers who use the anchor-based method should compare against multiple instruments. The study team then must decide which anchor to use to calculate MCID based on the mean score of patients who report a minimum detectable difference. Some research teams have reconciled the distribution- and anchor-based methods, for example through standardizing the anchor-based MCID by dividing by the pooled standard deviation [14].

Once the effect sizes have been determined and MCID established, they can be compared across groups or studies based on Cohen's guidelines for interpreting effect size estimates with small, medium, and large effect sizes corresponding to D = 0.2, 0.5 and 0.8. This numerical distinction may be somewhat arbitrary and the onus is on the researcher to ensure during interpretation that the findings are salient for patients and their clinicians.

10.7 Conclusion

Ideally, patient reported outcomes should be included in any study that pertains to the processes and outcomes of care of patients. Reliable PRO data can facilitate the accuracy of predictions, can provide a foundation for patient and clinician expectations, and are fundamental to engagement in authentic shared decision making for treatment. PROs are also necessarily imperfect. They attempt to pin down an experience, perspective, or value at one point in time and, as in any discipline, are merely proxies for the real-life experience or outcome that is being measured. Nonetheless, based on their critical importance to patient centered care and the new, sophisticated methods under development to enhance accuracy and precision and reduce burden, we can expect that PROs will have an increasingly central position in health services and clinical research.

References

1. Basch E, et al. Symptom monitoring with patient-reported outcomes during routine Cancer treatment: a randomized controlled trial. J Clin Oncol. 2016;34(6):557–65.
2. *Patient Reported Outcomes (PROs) in Performance Measurement*. 2013.
3. Wiering B, de Boer D, Delnoij D. Patient involvement in the development of patient-reported outcome measures: the developers' perspective. BMC Health Serv Res. 2017;17(1):635.
4. Veenstra CM, et al. A composite measure of personal financial burden among patients with stage III colorectal cancer. Med Care. 2014;52(11):957–62.
5. Cook KF, et al. PROMIS measures of pain, fatigue, negative affect, physical function, and social function demonstrated clinical validity across a range of chronic conditions. J Clin Epidemiol. 2016;73:89–102.
6. Ranard BL, et al. Crowdsourcing--harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med. 2014;29(1):187–203.
7. Lai JS, et al. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-fatigue scale. Qual Life Res. 2003;12(5):485–501.
8. Vetter TR, Cubbin C. Psychometrics: trust, but verify. Anesth Analg. 2019;128(1):176–81.
9. Edwards PJ, Roberts I, Clarke MJ, DiGuiseppi C, Wentz R, Kwan I, Cooper R, Felix LM, Pratap S. Methods to increase response to postal and electronic questionnaires. Cochrane Database of Syst Rev. 2009, Issue 3. Art. No.: MR000008. https://doi.org/10.1002/14651858.MR000008.pub4.
10. Dykema J, et al. Effects of incentives and prenotification on response rates and costs in a national web survey of physicians. Eval Health Prof. 2011;34(4):434–47.
11. Dillman D, Smyth J, Christian L. Internet, phone, mail, and mixed-mode surveys: the tailored design method. New Jersey: Wiley; 2014.
12. Calvert M, et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 2013;309(8):814–22.
13. Guyatt GH, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371–83.
14. Waljee JF, et al. Variation in patient-reported outcomes across hospitals following surgery. Med Care. 2015;53(11):960–6.

Landmark Papers

Basch E, Torda P, Adams K. Standards for patient-reported outcome-based performance measures. JAMA. 2013;310(2):139–40.

Guyatt GH, et al. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371–83.

Pezold ML, et al. Defining a research agenda for patient-reported outcomes in surgery: using a Delphi survey of stakeholders. JAMA Surg. 2016;151(10):930–6.

Snyder CF, et al. The role of informatics in promoting patient-centered care. Cancer J. 2011;17(4):211–8.

van der Meij E, et al. Using PROMIS for measuring recovery after abdominal surgery: a pilot study. BMC Health Serv Res. 2018;18(1):128.

Wiering B, de Boer D, Delnoij D. Patient involvement in the development of patient-reported outcome measures: a scoping review. Health Expect. 2017;20(1):11–23.

Studying What Happens in the Operating Room

11

Yue-Yung Hu and Caprice C. Greenberg

Abstract

The majority of surgical adverse events may be traced to the operating room (OR). However, traditional surgical health services research approaches are limited in their ability to assess what is happening in the OR. In this chapter, we present an overview of point-of-care research, including strategies for data collection and analysis. Seminal papers in the field of intraoperative research are discussed.

11.1 Why Study What Happens in the OR?

In 1999, the Institute of Medicine published *To Err is Human: Building a Safer Health System* [1], reporting that 44,000–98,000 patients die annually in the United States as a result of preventable medical errors – more than motor vehicle collisions, breast cancer, and AIDS. Adverse events are estimated to cost Americans $393–$958 billion dollars a year, representing 45% of our healthcare expenditures [2].

Between half and two-thirds of inpatient adverse events are attributable to surgical care [3–5], and 75% percent of these errors occur intraoperatively [6]. Yet, the

Y.-Y. Hu (✉)
Division of Pediatric Surgery, Ann & Robert H. Lurie Children's Hospital of Chicago, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Surgical Outcomes & Quality Improvement Center, Department of Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
e-mail: yueyunghu@luriechildrens.org

C. C. Greenberg
Wisconsin Surgical Outcomes Research Program, Department of Surgery, University of Wisconsin Hospitals & Clinics, Madison, WI, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_11

vast majority of research addressing surgical quality and safety does not give specific attention to the highest risk point-of-care: the operating room (OR).

The most commonly assessed risk factors – patient characteristics (e.g., comorbidities), operative risk factors (e.g., procedural complexity), surgeon volume, institutional volume, institutional infrastructure (e.g., availability of resources like critical care or pediatric anesthesiology) – are all fixed in place before a patient's operation begins; as such, they do not speak to what is happening in the OR. Moreover, they present few opportunities for modification.

Safety has been traditionally equated with the absence of postoperative morbidity or mortality. Using administrative (e.g., Medicare claims) or clinical (e.g., American College of Surgeons National Surgical Quality Improvement Project medical chart abstraction) data, the incidence of surgical complications may be estimated, from which the quality of care is inferred. However, in characterizing care as binary – the presence or absence of complications, and therefore, of quality – we lose information that may be used to prevent or mitigate them. Such studies are unable to discern the impact of individual events on outcomes and/or whether they occurred pre-, intra-, or postoperatively. They also overlook adverse events that have been compensated (e.g., near misses). Because not all complications are preventable, studying such resilience is of particular interest to those dedicated to improving care.

11.2 What Is There to Study in the OR?

11.2.1 Complexity of Intraoperative Care

Point-of-care research has been underutilized in surgery, likely due to the complexity of the OR. A safe operation depends upon the performance of individuals – surgeons, anesthesiologists, nurses – who work against time and production pressures, fatigue, and an ever-expanding body of medical knowledge. These individuals must function within intra- and inter-disciplinary teams in the OR, and to do so, they must overcome disparate levels of experience, competing priorities, ad hoc rather than fixed team membership, and institutional and/or professional cultures that may dictate behavioral and/or communication norms. Finally, these individuals and teams in the OR depend upon accurate and timely transfer of information and equipment from other locations within the hospital, often after a period of time has elapsed (e.g. preoperative anesthesia testing) [7].

11.2.2 Human Factors Engineering

Because of this complexity, the OR is particularly amenable to human factors research. Human factors is a field of engineering concerned with the "understanding of interactions among humans and other elements of a system," including physical, cognitive, social, and organizational aspects. It is responsible for safety advances in other complex work domains requiring high reliability, such as aviation and nuclear reactor control [8].

11.2.3 Conceptual Models

James Reason's *Swiss cheese model* (Fig. 11.1) conceptualizes a system as a stack of Swiss cheese slices. Each slice represents a layer of defense against error (active failures), and the holes in each slice represent unique vulnerabilities (latent failures) in each layer of defense. Errors are mitigated by multiple layers of defense; only when vulnerabilities align between layers does risk materialize into harm [9].

The *Donabedian model* describes three sequential categories for examining care quality: structure (the context in which healthcare is delivered, e.g., facilities, financing, staff, equipment), process (the interaction between patients and providers), and outcomes (the effects of care on patients and populations) [10]. These concepts have been foundational to health care quality research; however, the model has been criticized as overly linear.

The *Systems Engineering Initiative for Patient Safety (SEIPS) model* (Fig. 11.2) builds on the Donabedian platform and describes five domains: tools/technology,

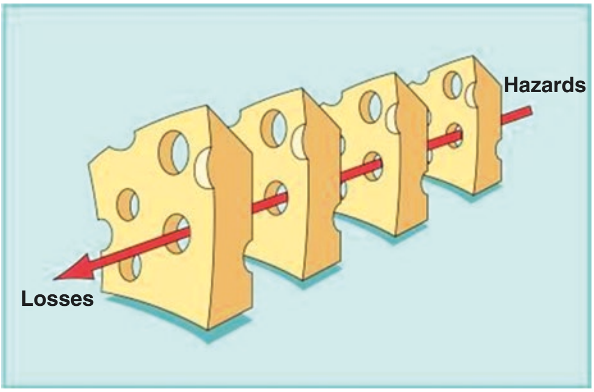


Fig. 11.1 James Reason's Swiss cheese model [9]

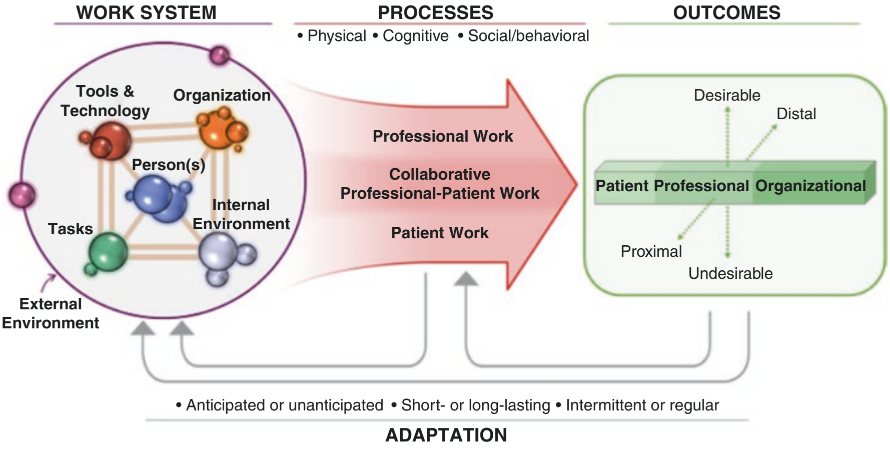

WORK SYSTEM
PROCESSES
• Physical • Cognitive • Social/behavioral
OUTCOMES
Tools & Technology
Organization
Person(s)
Internal Environment
Tasks
External Environment
Professional Work
Collaborative Professional-Patient Work
Patient Work
Desirable
Distal
Patient Professional Organizational
Proximal
Undesirable
• Anticipated or unanticipated • Short- or long-lasting • Intermittent or regular
ADAPTATION

Fig. 11.2 SIEPS 2.0 model [11]

organization, tasks, environment, and people. Three concepts added in the SEIPS 2.0 model – configuring (dynamic, hierarchical, and interactive properties), engagement (individuals and teams may perform activities separately and collaboratively), and adaptation (a feedback mechanism explaining how dynamic systems evolve in both planned and unplanned ways) – address concerns about over-linearity in the Donabedian model [11].

11.3 How Do We Study What Happens in the OR?

11.3.1 Data Sources

Questionnaires allow researchers to obtain answers to specific questions from large numbers of subjects, anonymously (which may increase honesty), expeditiously, and inexpensively. They must be rigorously designed and tested prior to deployment to optimize psychometric properties. If the intent of the question is not clear and/or the response options inadequately inclusive, the resultant data is uninterpretable. We therefore recommend using previously validated instruments whenever possible. For example, the Safety Attitudes Questionnaire assesses perceptions of teamwork in the OR [12]. Occasionally, no existing validated tool exists, and one must develop a new instrument, taking care to perform rigorous pilot and validity testing to ensure reliability and accuracy. A major limitation of even well-constructed surveys is the difficulty in investigating new questions or research themes that arise. Follow-up questions require administration of a second survey, and paired analysis requires two identical subject populations as well as identifiers to link them.

Interviews of key informants are a rich source of data, as respondents may elaborate on their experiences and perspectives in great detail. The interaction between interviewer and interviewee allows for new themes to be further explored as they emerge. Because they are conducted individually, confidentiality may be easier to assure. *Focus groups* are essentially group interviews. The added interaction between focus group members allows for disparate opinions and experiences to be explored in depth. However, the lack of anonymity in challenging cultural norms and/or the views of thought leaders may prevent individuals from being forthcoming. Finally, both interviews and focus groups are subject to recall bias, and it may be difficult or impossible to ascertain respondents' accounts of events.

Field observation, in which research personnel embed themselves *in vivo* in the work environment, allows prospective data collection at the point-of-care. Several validated instruments exist to enable observers to measure communication [13], teamwork (NOTECHS [14], OTAS [15]), and disruptions [16] in the OR. However, the Hawthorne effect may limit the authenticity of the events that are observed. Recall bias, although likely lessened by the third party perspective (vs. first person narrative obtained in interviews/focus groups), may still exist. Because multiple events occur rapidly and often simultaneously, they may be difficult to fully observe.

Physical constraints prevent multiple observers from being present, and the observers selected may have limited comprehension of unfolding events, depending on the scope of their domain knowledge.

Video recording addresses some of the challenges inherent to field observation. It generates a comprehensive record that allows for repeated viewing by multiple observers, minimizing data loss and/or misinterpretation. By replacing observers with equipment, the Hawthorne effect may be minimized, but not eliminated. Other disadvantages include the expense, a time-consuming analysis, and sensitivity around confidentiality/medicolegal discoverability issues.

11.3.2 Sampling

The sampling schema chosen depends upon the study objectives. If the goal is to make a statistical generalization, then the sample should be representative of the relevant population. For this reason, random sampling is typically used for quantitative research. In contrast, analytic generalizations, which relate to conceptual power rather than representativeness, has historically been linked to non-random sampling. However, these are not inviolable rules; non-random sampling may be appropriate for some quantitative research projects, just as random sampling may be best for some qualitative research projects.

There are five *random, probabilistic sampling* schema: simple, stratified, cluster, systematic, and multi-stage. *Non-random, purposive sampling* techniques include maximum variation, homogenous, critical case, theory-based, confirming/disconfirming, snowball/chain, extreme case, typical case, intensity, politically important case, random purposeful, stratified purposeful, criterion, opportunistic, mixed purposeful, convenience, quota, and multi-stage purposeful [17]. See Table 11.1 for a concise explanation of these sampling schema. Examples are given in Sect. 11.4 of this text.

Table 11.1 Major Sampling Schemes in Mixed-Methods Research [17]

Sampling scheme	Description
Simple[a]	Every individual in the sampling frame (i.e., desired population) has an equal and independent chance of being chosen for the study.
Stratified[a]	Sampling frame is divided into sub-sections comprising groups that are relatively homogeneous with respect to one or more characteristics and a random sample from each stratum is selected.
Cluster[a]	Selecting intact groups representing clusters of individuals rather than choosing individuals one at a time.
Systematic[a]	Choosing individuals from a list by selecting every *kth* sampling frame member, where *k* typifies the population divided by the preferred sample size.
Multi-Stage Random[a]	Choosing a sample from the random sampling schemes in multiple stages.
Maximum Variation	Choosing settings, groups, and/or individuals to maximize the range of perspectives investigated in the study.

(continued)

Table 11.1 (continued)

Sampling scheme	Description
Homogenous	Choosing settings, groups, and/or individuals based on similar or specific characteristics.
Critical Case	Choosing settings, groups, and/or individuals based on specific characteristic(s) because their inclusion provides the researcher with compelling insight about a phenomenon of interest.
Theory-Based	Choosing settings, groups, and/or individuals because their helps the researcher to develop a theory.
Confirming/ Disconfirming	After beginning data collection, the researcher conducts subsequent analyses to verify or contradict initial results.
Snowball/Chain	Participants are asked to recruit individuals to join the study.
Extreme Case	Selecting outlying cases and conducting comparative analyses.
Typical Case	Selecting and analyzing average or normal cases.
Intensity	Choosing settings, groups, and/or individuals because their experiences relative to the phenomena of interest are viewed as intense but not extreme.
Politically Important Case	Choosing settings, groups, and/or individuals to be included or excluded based on their political connections to the phenomena of interest.
Random Purposeful	Selecting random cases from the sampling frame and randomly choosing a desired number of individuals to participate in the study.
Stratified Purposeful	Sampling frame is divided into strata to obtain relatively homogeneous sub-groups and a purposeful sample is selected from each stratum.
Criterion	Choosing settings, groups, and/or individuals because they represent one or more criteria.
Opportunistic	Researcher selects a case based on specific characteristics (i.e., typical, negative, or extreme) to capitalize on developing events occurring during data collection.
Mixed Purposeful	Choosing more than one sampling strategy and comparing the results emerging from both samples.
Convenience	Choosing settings, groups, and/or individuals that are conveniently available and willing to participate in the study.
Quota	Researcher identifies desired characteristics and quotas of sample members to be included in the study.
Multi-Stage Purposeful Random	Choosing settings, groups, and/or individuals representing a sample in two or more stages. The first stage is random selection and the following stages are purposive selection of participants.
Multi-Stage Purposeful	Choosing settings, groups, and/or individuals representing a sample in two or more stages in which all stages reflect purposive sampling of participants.

[a]Represent random (i.e., probabilistic) sampling schemes

11.3.3 Analysis

Surgical health services researchers have long performed *quantitative research*, to which numerical measurement and statistical analysis are central. Certain features of intraoperative care are amenable to such techniques. For example, the number of exits/entries into the OR, the number of interdisciplinary communications, and the amount of time spent answering pages may all be counted or calculated and analyzed for an association – most frequently using mathematical modeling – with an outcome of interest (e.g., the number of adverse events).

Many aspects of intraoperative care are inadequately described in quantitative terms. Often, the etiology of a complication is multifactorial and therefore not measurable in a straightforward way. We previously described [18] a scenario in which a patient undergoing a massive retroperitoneal sarcoma resection had unrecognized blood losses, was under-resuscitated, and suffered cardiac ischemia. Although on the surface, the error seems simple, the underlying factors were many, varied, and layered. The anesthesiology resident was left unsupervised for a period of time, as the attending anesthesiologist had supervisory responsibilities in other ORs (coordination). Seeing the suction canisters, the anesthesiology resident asked the surgical attending if there was bleeding. The surgeons denied "surgical bleeding," by which they meant there was continuous oozing, but no single source on which to intervene (communication). The surgeons did not seek information about the patient's hemodynamic status, and the anesthesiology resident did not hear their discussion about the changing demands of the case (monitoring/vigilance). Further probing by the anesthesiology resident was likely inhibited by the culture of hierarchy in the OR (status asymmetry). The urgency of the situation was not recognized and resuscitation did not begin until the anesthesiology attending returned (knowledge/training). Such non-structured data is better analyzed using *qualitative research* techniques. Qualitative researchers strive to make sense of phenomena in the real, uncontrolled world – in particular, how people understand them and the meaning they attribute to them. Qualitative techniques may be appropriate for exploratory, hypothesis-generating work; these findings may then be further investigated using traditional quantitative techniques.

Mixed-methods researchers may utilize both quantitative and qualitative techniques to triangulate a research question. This powerful combination is commonly encountered in point-of-care research.

11.3.4 Practical Considerations

Those seeking to perform point-of-care research may benefit from *collaboration* with research experts from relevant fields, e.g., human factors engineers, cognitive psychologists, business psychologists, psychometricians, and qualitative researchers. Ideally, these collaborators should be involved early in the study design process, as their input may significantly alter and improve the quality of data collected as well as its analysis and interpretation.

Additionally, as the OR is a high-stress environment in which workers have been subject to observation or monitoring only as a regulatory function, intraoperative event analysis may constitute a sensitive research activity. To ensure provider participation, *engagement* is critical. Key administrative and clinical collaborators should be identified and involved early to design research protocols that minimally disrupt workflow and/or inconvenience providers. The support of these collaborators will provide face validity for the project at the time of review, as well as ensure up-to-date and rigorous methodological considerations.

Medicolegal risk is an oft-cited barrier to provider participation in point-of-care research. In our discussions with legal experts, such documentation is more likely to provide evidence that the standard of care was delivered than the contrary. On the other hand, if the standard of care was violated, settling without a lengthy trial would be better for all involved. In either case, video may function in the best interest of healthcare providers. Nevertheless, those concerned about litigation will require further reassurance. For our study using intraoperative video, we instituted the following: (a) video recordings were intentionally low-resolution to limit provider identifiability; (b) videos were transcribed, eliminating provider identifiers, then destroyed within 90 days, as per study protocol; (c) case identifiers were maintained in a coded fashion to allow for chart review at 30 days, then destroyed, as per protocol, (d) a Certificate of Confidentiality was obtained from the Department of Health and Human services to protect against involuntary disclosure [18].

11.4 Seminal Papers on What Happens in the OR

11.4.1 deLeval MR, Carthey J, Wright DJ, Farewell VT, Reason JT. Human Factors and Cardiac Surgery: A Multicenter Study. *Journal of Thoracic and Cardiovascular Surgery* 2000;119(4Pt1):661–72

Data Sources: Quantitative (surgeon volume), questionnaire (Surgical Team Assessment Record), and direct field observation.

Sampling: Homogeneous & Critical Case – 243 neonatal arterial switch operations performed by all 21 cardiac surgeons in the United Kingdom during an 18 month study period were included. Convenience – 50 of the cases were not observed due to concurrent scheduling. 20 observed cases were removed from analysis because the observer had insufficient domain knowledge.

Analysis: Human factors observers noted individual and team performance, intra- and inter-disciplinary team communication, situational and organizational data, major events ("likely to have serious consequences for the safety of the patient"), minor events ("disrupted the 'surgical flow' of the procedure but…not expected to have serious consequences for the safety of the patient"), and compensation or lack thereof. Multivariable logistic regression was performed to predict death and/or near misses. Examples of major and minor events are given.

Findings: Neither the STAR questionnaire nor surgeon volume were significant predictors of death and/or near misses. The odds of death and/or a near miss increased by a factor of 6.2 for a compensated major event and by a factor of 40 for an uncompensated major event. Minor events, whether compensated or not, increased the risk by a factor of 1.4 and had a multiplicative effect.

11.4.2 Pisano GP, Bohmer RMJ, Edmondson AC. Organizational Differences in Rates of Learning: Evidence from the Adoption of Minimally Invasive Cardiac Surgery. *Management Science* 2001;47 (6);752–68

Data Sources: Quantitative (type of operation, number of CABG grafts, Higgins score, procedural volume) and interviews (of cardiac surgeons, cardiologists, anesthesiologists, nurses, perfusionists, and administrators).

Sampling: Homogeneous & Critical Case – 660 cases across 16 hospitals in which a new minimally invasive cardiac surgical technology was deployed.

Analysis: A multivariable linear regression model was constructed to predict procedure time. Interviews were used to obtain data on each site's approach to adoption, management practices, and organizational climate. Two hospitals, one fast-learning and one slow-learning, are described.

Findings: Despite receiving highly standardized training on the new technology, there was substantial variation between institutions in both initial operative time and the rate of change in operative time. Interviews revealed that differences in learning could be attributed to differences in organizational processes. The fast-learning hospital sent a team with a history of working together to be trained and kept that team intact for the first 30 cases. Prior to introduction of the procedure, meetings were held to standardize terminology and clarify individual roles. Preoperative briefings and postoperative debriefings were instituted for the first 10 and 20 cases, respectively. The surgeon was noted to encourage "a high degree of cooperation among members of the team." The slow-learning hospital sent a team for training based on staff availability. The operative team was inconstant for the first six cases, and members had little advanced notice for preparation. The surgeon indicated that his focus was on "mastering the technical aspects of the operation rather than managing the overall adoption process."

11.4.3 Catchpole K, Mishra a, Handa a, Mcculloch P. Teamwork and Error in the Operating Room: Analysis of Skills and Roles. *Annals of Surgery* 2008;247:699–706

Data Sources: Direct field observation.

Sampling: Typical Case & Opportunistic – 26 laparoscopic cholecystectomies and 22 carotid endarterectomies were selected for observation because "the operations are common enough to allow data collection on frequent occasions, are moderately complex, have recognizable complications that can be monitored, and all involve surgeons, nurses, and anesthetists."

Analysis: Observers scored team performance using the Oxford NOTECHS system. Outcomes recorded included errors in surgical technique, errors/problems outside of the operative field, and events with potential or actualized harm to patients or staff. Multivariable regression models were constructed.

Findings: Improvements in surgeons' leadership and management scores were associated with reduced operative time, while improvements in anesthetists' leadership and management scores were associated with increased operative time. Improved nursing leadership and management scores were associated with lower other procedural problems/errors. Higher surgical situational awareness resulted in fewer surgical technical errors.

11.4.4 Hu YY, Arriaga AF, Roth EM, Peyre SE, Corso KA, Swanson RS, Osteen RT, Schmitt P, Bader AM, Zinner MJ, Greenberg CC. Protecting Patients from an Unsafe System: The Etiology and Recovery of Intraoperative Deviations in Care. *Annals of Surgery* 2012;256 (2):203–10

Data Sources: Intraoperative video.

Sampling: Opportunistic – 10 operations with published complication rates of >20% were selected in order to maximize the chances of observing an unanticipated event.

Analysis: A surgeon, a cognitive psychologist, and an educational psychologist reviewed transcripts of the videos to identify deviations in care, which were classified as delays (halts in forward progress >2 minutes), safety compromises (episodes of increased risk of harm to the patient), or both. Classifications were confirmed by clinical domain experts (2 surgeons, 1 anesthesiologist, and 1 OR nurse). Contributing and compensatory factors were identified for each deviation and attributed to the patient, the providers, or the organization/environment. Summary/descriptive statistics (i.e., counts and frequencies) and representative examples were given.

Findings: 33 deviations (10 delays, 17 safety compromises, 6 both) were identified, averaging one in every 80 minutes of operative time. The most common contributing factors were communication and organizational structure. Providers compensated for 97% of deviations, whereas the system was responsible for none of the compensation.

11.4.5 Jones LK, Jennings BM, Higgins MK, deWaal FBM. Ethological Observations of Social Behavior in the operating Room. *Proceedings of the National Academy of Sciences* 2018; 115(29):7575–80

Data Sources: Direct field observation.

Sampling: Convenience and Quota – 200 cases at 3 teaching hospitals within 6 driving miles of one another, with a limit of ≤4 cases per attending surgeon.

Analysis: The attending surgeon's gender, age, and academic affiliation and the gender composition of the OR were noted. Non-technical, interprofessional communication behaviors were time-stamped and classified as cooperation, conflict, or neither. Directionality (source and recipient for each behavior) was noted. Generalized multilevel models were constructed to predict cooperative and conflict behaviors.

Findings: Of all communications, 59% were cooperative and 2.8% conflicts. 34.5% of procedures had conflicts. The probability of cooperation decreased as the percentage of males in the room increased. Cooperation was higher and conflict was lower if the attending surgeon's gender differed from the primary gender composition of the rest of the team; conversely, cooperation was lower and conflict was higher if the surgeon and team were gender concordant. The effect was stronger for male than for female surgeons.

11.5 Conclusion

A better understanding of what is happening in the OR, the surgical point-of-care, is needed to improve surgical quality and safety. To study this domain of care, surgical health services researchers should consider qualitative and mixed-methods techniques, as well as multidisciplinary collaboration. It should be noted that point-of-care research may be performed across the care continuum; optimizing individual, team, and system performance is critical to improving quality and safety in domains of care outside, as well as inside, of the OR. We anticipate that work in this area will continue to increase and suggest that surgical health services researchers familiarize themselves with the relevant research concepts and techniques.

References

1. Kohn LT, Corrigan J, Donaldson MS. To err is human: building a safer health system. Washington, DC: National Academy Press; 2000.
2. Goodman JC, Villarreal P, Jones B. The social cost of adverse medical events, and what we can do about it. Health Aff (Millwood). 2011;30:590–5.
3. Gawande AA, Thomas EJ, Zinner MJ, Brennan TA. The incidence and nature of surgical adverse events in Colorado and Utah in 1992. Surgery. 1999;126:66–75.
4. Leape LL, Brennan TA, Laird N, et al. The nature of adverse events in hospitalized patients. Results of the Harvard medical practice study II. N Engl J Med. 1991;324:377–84.
5. Thomas EJ, Studdert DM, Burstin HR, et al. Incidence and types of adverse events and negligent care in Utah and Colorado. Med Care. 2000;38:261–71.
6. Rogers SO Jr, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery. 2006;140:25–33.
7. Vincent C, Moorthy K, Sarker SK, Chang A, Darzi AW. Systems approaches to surgical quality and safety: from concept to measurement. Ann Surg. 2004;239:475–82.
8. What is Ergonomics? (Accessed 17 Jan 2019, at https://www.iea.cc/whats/index.html.)
9. Reason J. Human error: models and management. BMJ. 2000;320:768–70.
10. Donabedian A. The quality of care. How can it be assessed? JAMA. 1988;260:1743–8.
11. Holden RJ, Carayon P, Gurses AP, et al. SEIPS 2.0: a human factors framework for studying and improving the work of healthcare professionals and patients. Ergonomics. 2013;56:1669–86.
12. Makary MA, Sexton JB, Freischlag JA, et al. Operating room teamwork among physicians and nurses: teamwork in the eye of the beholder. J Am Coll Surg. 2006;202:746–52.

13. Lingard L, Regehr G, Espin S, Whyte S. A theory-based instrument to evaluate team communication in the operating room: balancing measurement authenticity and reliability. Qual Saf Health Care. 2006;15:422–6.
14. Mishra A, Catchpole K, McCulloch P. The Oxford NOTECHS system: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care. 2009;18:104–8.
15. Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N. Observational teamwork assessment for surgery: content validation and tool refinement. J Am Coll Surg. 2011;212:234–43. e1–5
16. Sevdalis N, Forrest D, Undre S, Darzi A, Vincent C. Annoyances, disruptions, and interruptions in surgery: the disruptions in surgery index (DiSI). World J Surg. 2008;32:1643–50.
17. Onwuegbuzie AJ, Collins KMT. A typology of mixed methods sampling designs in social science research. Qual Rep. 2007;12:281–316.
18. Hu YY, Arriaga AF, Roth EM, et al. Protecting patients from an unsafe system: the etiology and recovery of intraoperative deviations in care. Ann Surg. 2012;256:203–10.

Evaluating Non-Technical Skills in Surgery

12

Jason C. Pradarelli, Steven Yule, and Douglas S. Smink

Abstract

Non-technical skills are increasingly recognized as essential to providing safe and effective surgical care. In this chapter, we first review the policy context for studying surgeons' non-technical skills. Next, we define the Non-Technical Skills for Surgeons (NOTSS) behavior rating system, including situation awareness, decision making, communication and teamwork, and leadership in the operating room. We review major domains of ongoing research on surgeons' non-technical skills and discuss measurement considerations for researchers planning to study non-technical skills. Finally, we identify promising areas for future health services researchers to develop successful careers studying non-technical skills in surgery.

12.1 Context for Studying Non-Technical Skills in Surgery

In the evolution of surgical science, surgeons' non-technical skills have only recently been recognized as an essential component of intraoperative performance [1]. Traditionally, efforts to improve surgical care—including pay-for-performance, public

J. C. Pradarelli (✉) · D. S. Smink
Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA

Ariadne Labs at Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health, Boston, MA, USA
e-mail: JPRADARELLI@PARTNERS.ORG

S. Yule
Department of Surgery, Brigham and Women's Hospital, Boston, MA, USA

Ariadne Labs at Brigham and Women's Hospital and the Harvard T.H. Chan School of Public Health, Boston, MA, USA

STRATUS Center for Medical Simulation, Brigham and Women's Hospital,
Boston, MA, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_12

reporting of outcomes, or reducing hospital readmissions—have focused on bettering care outside the operating room (OR). However, policy attempts to enact these strategies have yet to realize substantial improvements in surgical outcomes [2, 3].

Currently, health policy makers and researchers are turning their focus toward intraoperative performance for improving surgical care delivery. Examples such as surgical volume pledges reflect this emphasis and indicate a trend toward studying what happens inside the OR [4]. Intuitively, it makes sense that the more often surgeons and hospitals perform a certain procedure, the better their outcomes are. But why is this so? Despite instituting these logical policies, it remains elusive to identify exactly what happens in the OR that determines a surgical patient's outcome.

Initially, light was shed on surgeons' technical skills as major determinants of patient outcomes with surgery. In 2013, Birkmeyer et al. published a landmark study describing the association between bariatric surgeons' technical skill ratings and postoperative outcomes after laparoscopic gastric bypass [5]. In this study, higher ratings of surgical skill—assessed by analysis of laparoscopic video—were associated with lower rates of postoperative adverse events [5]. By analyzing videos of surgeons' laparoscopic instruments during an actual operation, an empiric relationship was established linking technical skill with surgical outcomes.

Despite uncovering this important skill-outcome relationship and demonstrating the learning potential of directly studying surgeons' actions in the OR, technical skills are not the only intraoperative source of variation in surgical outcomes. Prior research has shown that surgical errors are as likely to originate from behavior failures as from a lack of technical expertise [6]. Non-technical skills—such as situation awareness, decision making, communication and teamwork, leadership—are increasingly recognized as essential to attaining surgical excellence [1]. Deficiencies of non-technical skills have been associated with a higher risk of surgical complications and mortality through diagnostic failure [7], poor teamwork [8], or miscommunication [9]. Non-technical skills are essential for performing safe and effective surgery and have just begun to show their potential for impactful investigation and for establishing successful careers in surgical health services research.

12.2 Defining Non-Technical Skills for Surgeons

Non-technical skills can be defined as the cognitive and social skills that underpin knowledge and expertise in high demand workplaces [1, 10]. In surgery, they include the thinking skills and personal interactions that accompany surgical knowledge and technical competence as requisites for achieving surgical excellence [1]. The Non-Technical Skills for Surgeons (NOTSS) taxonomy described in this chapter comprise four major categories of skill that provide individual surgeons, surgical educators, and researchers with the vocabulary and parameters necessary for rating surgical performance in the OR. The NOTSS taxonomy arranges behaviors into four higher order categories—situation awareness, decision making, communication and teamwork, leadership—which are explained by several lower order elements of behavior (Table 12.1).

Table 12.1 Non-Technical Skills for Surgeons (NOTSS) skills taxonomy (adapted from Flin et al., 2016 [1] and the NOTSS Handbook v1.2 [11])

Category	Element
Situation awareness	Gathering information Understanding information Projecting and anticipating future state
Decision making	Considering options Selecting and communicating option Implementing and reviewing decisions
Communication and teamwork	Exchanging information Establishing a shared understanding Coordinating team activities
Leadership	Setting and maintaining standards Supporting others Coping with pressure

Together, situation awareness and decision making are considered the cognitive skills of NOTSS; communication and teamwork and leadership are the social skills. This arrangement allows the tool to guide observers to classify behaviors into greater or lesser detail depending on the granularity of the analysis required. In this chapter, an overview of the NOTSS taxonomy will be presented; detailed examples of specific behaviors for each category can be found in the NOTSS Handbook v1.2 [11].

12.2.1 Situation Awareness

Situation awareness in NOTSS is defined as “developing and maintaining a dynamic awareness of the situation in the operating room, based on assembling data from the environment (patient, team, time, displays, equipment), understanding what they mean, and thinking ahead about what may happen next” [11]. Situation awareness is considered the most critical non-technical skill because of its importance for the other three skills [1]. Specifically, accurate decision making, timely communication, and appropriate leadership all depend on the dynamic awareness of the situation that the surgeon possesses. Situation awareness comprises three distinct behavior elements: (i) gathering information, (ii) understanding and interpreting information, and (iii) projecting and anticipating future states based on this interpretation.

12.2.2 Decision Making

In NOTSS v1.2, decision making is defined as “skills for diagnosing a situation and reaching a judgement in order to choose an appropriate course of action” [11]. Decision making styles can differ based on the individual scenario, depending on the urgency of the situation, availability of existing guidelines or protocols, level of experience of the surgeon, and novelty of the situation. Theories are described elsewhere to differentiate between analytical [12], rule-based [12], recognition-primed [13], and creative styles of decision making [12]. The three behavior elements of decision making are (i) considering options, (ii) selecting and communicating an option to relevant personnel, and (iii) implementing and reviewing decisions as the situation evolves.

12.2.3 Communication and Teamwork

Communication and teamwork are defined as "skills for working in a team context to ensure that everyone has an acceptable shared picture of the situation and can complete tasks effectively" [11]. Across the OR team, having a shared mental model of the current and anticipated future states is an important concept underpinning the communication and teamwork in NOTSS. The three relevant elements of behavior for communication and teamwork are (i) exchanging information in a timely manner, (ii) establishing a shared understanding among team members, and (iii) coordinating team activities in a simultaneous, collaborative manner.

12.2.4 Leadership

In the final behavior category of NOTSS, leadership is defined as "leading the team and providing direction, demonstrating high standards of clinical practice and care, and being considerate about the needs of individual team members" [11]. Leadership is necessary in the OR whether the procedure is elective or emergent, with routine instruments or with new technology. The three leadership elements in NOTSS are (i) setting and maintaining standards, (ii) supporting others via emotional or physical help as needed, and (iii) coping with pressure.

12.3 Why Study Non-Technical Skills in Surgery?

Opportunities abound for health services researchers interested in studying non-technical skills in surgery. Although frameworks for describing these aspects of surgical performance date back to the early 2000s [10], studies describing outcomes related to them are only in their early stages. Initial studies have shown associations between non-technical skills and measures of patient safety, surgeons' medicolegal risk, and psychological safety of other OR team members.

12.3.1 Patient Safety

Inadequate non-technical skills have been associated with heightened risks to patient safety. For instance, researchers have identified that errors leading to bile duct injuries during laparoscopic cholecystectomy stem primarily from non-technical skill deficiencies like misperception and diagnostic failure, not from errors of technical skill [7]. Early studies have associated worse interdisciplinary communication behaviors with higher odds of major complications in a convenience sample of procedures in the Kaiser network [8]. Furthermore, a recent experimental study using video scenarios demonstrated a positive correlation between surgeons' non-technical skills and patient safety ratings among surgeons in the UK and the USA [14]. Currently, direct evidence linking non-technical skills on traditional patient outcomes such as complications and mortality is limited and represents a potential area for future investigation.

Indirectly, the successes of interventions to improve OR team communication have demonstrated the potential for non-technical skills to influence patient outcomes. In 2007 and 2008, eight hospitals in a diverse global setting enrolled in the World Health Organization's Safe Surgery Saves Lives program, which included implementing the Surgical Safety Checklist. The checklist was designed to improve team communication and consistency of surgical care processes. In aggregate at all participating hospitals, the rate of any postoperative complication decreased from 11.0% to 7.0% after implementation of the checklist; the rate of in-hospital postoperative deaths decreased from 1.5% to 0.8% [15]. When the Surgical Safety Checklist was introduced at scale across all hospitals in the South Carolina Hospital Association in 2015, hospitals who fully implemented the checklist saw a relative 22% reduction in postoperative mortality rates compared to those that did not complete the program [16]. While explicit team behaviors were not measured in these studies, the intended purpose of the team-based checklist suggests that non-technical skills were an important component of its success in reducing rates of adverse surgical events.

12.3.2 Medicolegal Risks

While patient safety remains the top priority in studying health care delivery, researchers have considered other important financial and psychological factors related to health care providers when examining the effects of non-technical skills in surgery. A review of institutional incident reports revealed that nearly half of the adverse events involving surgical patients could be attributed to communication breakdowns among the patient care team [6]. Surgical malpractice claims have been reviewed to analyze patterns in communication breakdowns [9]. Furthermore, researchers have identified correlations between poor communication and teamwork behaviors and higher numbers of malpractice claims filed against surgeons [17]. In mitigating high-risk scenarios, studies of simulated OR crises have shown that operative teams led by surgeons and anesthesiologists with better non-technical skills are faster to resolve low-frequency, high-acuity events like major hemorrhage and airway emergencies [18]. Poor use of non-technical skills has medicolegal and financial implications for patients, surgeons, surgical departments, and insurers that health services researchers can help to identify and improve.

12.3.3 Psychological Safety

Non-technical skills can also influence the psychological safety of the entire OR team, which has implications for understanding implementation efforts to improve surgical patient care. Psychological safety is defined as a shared belief held by members of a team that the team is safe for interpersonal risk-taking [19]. Survey research has produced evidence that the Surgical Safety Checklist is used most appropriately when surgeons and surgical departments demonstrate better teamwork and leadership behaviors [20]. When OR teams demonstrated examples of respect, support for team-based practices, and information-sharing behaviors, implementation of patient safety interventions such as the checklist was more likely to be successful [20]. Further emphasizing this point, surveys of OR personnel that used the Surgical Safety Checklist variably in South Carolina showed a positive correlation between postoperative mortality rates and the *perception* of safety after the team-based checklist intervention [21]. Researchers who study the diverse and complex influence of non-technical skills in surgery can help understand and design more effective patient safety initiatives to benefit patients, surgical care providers, and insurers alike.

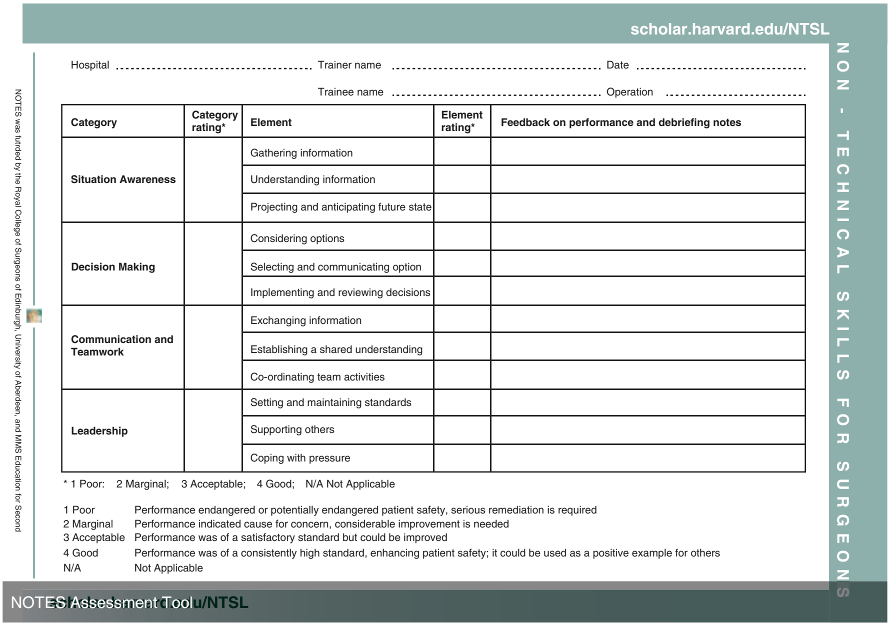
12.4 Measuring Non-Technical Skills in Surgery

To assess non-technical skills reproducibly, behavioral marker systems were developed in surgery both for individuals and for teams. The overall premise for these behavior rating tools is that a third-party observer uses a validated framework to assess the intraoperative behavior of individuals or teams for a defined period of time [1]. These behavioral marker systems generally comprise three components: (i) a skills taxonomy, (ii) behavioral indicators of levels of performance for each skill, and (iii) a rating scale [1]. They were developed in response to observational research and analyses of closed claims that highlighted the importance of non-technical skills in operative performance and patient safety. Observers are trained to use a specific framework to enhance the objectivity of judgments regarding effectiveness and appropriateness of the cognitive and social skills displayed during an operation. To learn how to observe and assess non-technical skills in surgery, professional development workshops are hosted by several surgical organizations, including the Royal College of Surgeons of Edinburgh [22] and the American College of Surgeons, among others [23]. A description of the major behavior rating tools available to researchers and educators can be found in Table 12.2.

The first behavior rating tools were developed to measure intraoperative performance of individuals, first for anesthesiologists and then for surgeons. The Anesthetists' Non-Technical Skills (ANTS) system was used to support training and assessment for anesthesiologists [24], which then was modified to rate anesthetic assistants [25]. The method for developing ANTS was adapted from the process used to develop similar rating tools in aviation to assess pilots' non-technical skills (e.g. NOTECHS). This method has since formed the template for developing behavioral marker systems for individual surgeons (NOTSS) [26] and scrub practitioners (SPLINTS) [27, 28]. As an example, the rating scale for NOTSS can be found in Fig. 12.1.

Table 12.2 Available behavior rating tools in surgery (adapted from Flin et al., 2016 [1])

Behavior rating tool	Description
NOTSS	Non-Technical Skills for Surgeons: [11, 26] Developed *de novo* with surgeons to observe and rate individual surgeons intraoperative performance
ANTS	Anesthetists' Non-Technical Skills: [24] Developed with anesthesiologists to observe and rate individual anesthesiologists' performance in the OR
SPLINTS	Scrub Practitioners' List of Intraoperative Non-Technical Skills: Developed with scrub nurses and surgeons to observe and rate individual scrub practitioners' performance in the OR
OTAS	Observational Teamwork Assessment for Surgery: [29] Teamwork assessment tool for three OR sub-teams, developed according to a theory of leadership
Oxford NOTECHS	Aviation tool for observing and rating entire teams, applied to OR teams [30]
Revised NOTECHS	Amended aviation tool for observing and rating entire teams, tailored to OR teams [31]
Oxford NOTECHS II	Amended aviation tool for observing and rating entire OR teams, revised to have better discriminatory ability between teams [32]
T-NOTECHS	Amended tool for assessing non-technical skills in trauma teams [34]


scholar.harvard.edu/NTSL

NON-TECHNICAL SKILLS FOR SURGEONS

Hospital .... Trainer name .... Date ....

Trainee name .... Operation ....

Category	Category rating*	Element	Element rating*	Feedback on performance and debriefing notes
Situation Awareness		Gathering information		
		Understanding information		
		Projecting and anticipating future state		
Decision Making		Considering options		
		Selecting and communicating option		
		Implementing and reviewing decisions		
Communication and Teamwork		Exchanging information		
		Establishing a shared understanding		
		Co-ordinating team activities		
Leadership		Setting and maintaining standards		
		Supporting others		
		Coping with pressure		

* 1 Poor; 2 Marginal; 3 Acceptable; 4 Good; N/A Not Applicable

1 Poor	Performance endangered or potentially endangered patient safety, serious remediation is required
2 Marginal	Performance indicated cause for concern, considerable improvement is needed
3 Acceptable	Performance was of a satisfactory standard but could be improved
4 Good	Performance was of a consistently high standard, enhancing patient safety; it could be used as a positive example for others
N/A	Not Applicable

NOTES was funded by the Royal College of Surgeons of Edinburgh, University of Aberdeen, and MMS Education for Second

NOTES Assessment Tool

scholar.harvard.edu/NTSL

Fig. 12.1 Non-Technical Skills for Surgeons (NOTSS) rating scale [11]

Categories of team behaviors for non-technical skills parallel many of those for individuals in the OR; as such, team behavior rating tools reflect a similar structure. The major difference for team behavioral marker systems is that they include modifying descriptors for behaviors of each discipline in the OR team (e.g. surgical, anesthetic, and nursing disciplines). The dominant team behavioral marker systems are Objective Teamwork Assessment in Surgery (OTAS) and surgical versions of NOTECHS. OTAS was developed from a theoretical model of teamwork and designed specifically for the OR environment [29], whereas NOTECHS was originally developed for pilots in civil aviation and subsequently applied to the OR where it showed good levels of reliability [1, 30]. The revised NOTECHS system added a category for communication behaviors and tailored the subcategory descriptors specifically for teams in an operative context [31]. NOTECHS II was developed to provide better discrimination between levels of performance for OR teams within the normal range [32]. In recent years, several behavioral marker systems have been developed to assess individual and team behaviors in health care settings outside of the OR [1], including postoperative care handoffs [33], trauma [34], critical care [35], and emergency departments [36].

For any of the behavioral marker systems described above, the assessment tool is only useful to researchers and educators if it produces valid and reliable results. That is, the assessments must measure the skills they are supposed to measure and be consistent with repeated measures across observers and over time. Subjectivity on the part of the rater must be minimized [1]. Furthermore, researchers should select assessment tools that can differentiate between levels of performance in a meaningful manner. In other words, the tool should also have good discriminatory ability between subjects, which motivated the revamping of the NOTECHS system to NOTECHS II [32]. To study non-technical skills in surgery effectively, researchers should understand the different types of validity evidence (e.g. criterion-related validity showing prediction of surgical outcomes) and reliability testing (e.g. inter-rater reliability using intra-class correlation coefficient) that support or limit implementation of a behavioral marker system [1]. In this early growth phase of non-technical skills as a scientific discipline, gathering validity and reliability data for behavioral marker systems offer potential for scientists to build the foundations for successful careers in surgical quality improvement and health services research.

12.5 Future Directions for Health Services Researchers Studying Non-Technical Skills

For researchers who desire to steer their careers toward studying non-technical skills in surgery, opportunities to enter and advance the field are countless. Growing areas of non-technical skill study currently include measures of patient safety, medicolegal and financial risks, and psychological safety of the OR team. These topics remain early in their scientific progression and offer angles to develop an existing discipline of non-technical skills in surgery. Interventions to improve surgeons' non-technical skills are also being investigated, although efforts focused on education, training, and implementation of non-technical skills are not covered in this chapter. In general, current evidence for individual surgeons' non-technical skills has focused on rating surgical trainees by expert raters. Future investigations could focus on practicing surgeons, real operative settings, and considerations of non-experts as raters. In this section, research questions that face the field of studying surgeons' non-technical skills will be highlighted for health services researchers.

Questions to be answered involve how to scale up measurement of non-technical skills in surgery [1]. For example, how many observers are required? Should they be expert surgeons or social scientists? Can surgical trainees rate non-technical skills accurately? Can lay people achieve the same level of measurement accuracy as trained raters using crowdsourcing technology? How much training is required to rate non-technical skills? Do raters need to observe in-person, or is video-based review as good or better? Opportunities exist to clarify the logistics and scalability of measuring non-technical skills in surgery.

The effects of non-technical skills on patient outcomes also needs to be studied more directly. Studies of surgical checklist implementation have reported effects on patient outcomes [15, 16]. The mechanisms for these effects are poorly characterized but are likely to be mediated by improved teamwork behaviors, increased sharing of mental models, and enhanced psychological safety and willingness to speak up. These non-technical skills can be observed and assessed directly using the behavioral marker systems described in this chapter. If non-technical skills can be shown to definitively predict patient outcomes with surgery, then it will be possible to identify which of these skills are most critical to emphasize from an educational and professional policy perspective. Linking patient outcomes directly with surgeons' non-technical skills would provide a convincing argument to standardize practices to reduce variation in surgical outcomes.

Beyond their effects on patient care, it is unknown how non-technical skills affect physicians' capacity to deal with cognitive and emotional stressors outside of the OR. Conceptually, it is appealing to link surgeons' non-technical skills and measures such as burnout, depression, and physician suicide rates. This would have important implications for health policy but has not yet been examined empirically. Non-technical skills could be investigated in relation to job satisfaction, academic productivity, or social status among surgical trainees. Surgical departments and national policymakers would benefit from a more comprehensive understanding of non-technical skills' influence on physician behavior.

Lastly, as the field of non-technical skills in surgery continues to grow and be implemented more broadly, refinements of the behavioral marker systems will be necessary. Just as team-based assessment tools have evolved and spread from the OR to external environments such as the trauma bay, critical care units, and emergency departments, the NOTSS system should also evolve and require updates as researchers reveal nuances of measuring non-technical skills in individual surgeons. Health services researchers have numerous opportunities to develop successful careers and make a significant impact on surgical care delivery by pursuing the study of non-technical skills in surgery.

References

1. Flin R, Youngson G, Yule S. Enhancing surgical performance: a primer in non-technical skills. Boca Raton, FL: CRC Press; 2016.
2. Osborne NH, Nicholas LH, Ryan AM, Thumma JR, Dimick JB. Association of hospital participation in a quality reporting program with surgical outcomes and expenditures for medicare beneficiaries. J Am Med Assoc. 2015;313(5):496–504.

3. Hawn M, Vick CC, Richman J, et al. Surgical site infection prevention: time to move beyond the surgical care improvement program. Ann Surg. 2011;254(3):494–501.
4. Urbach DR. Pledging to eliminate low volume surgery. N Engl J Med. 2015;373(15):1385–8.
5. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med. 2013;369(15):1434–42. https://doi.org/10.1056/NEJMsa1300625.
6. Gawande AA, Zinner MJ, Studdert DM. Analysis of errors reported by surgeons at three teaching hospitals. Surgery (United States). 2003;133:614–21.
7. Way LW, Stewart L, Gantert W, et al. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. Ann Surg. 2003;237(4):460–9.
8. Mazzocco K, Petitti DB, Fong KT, et al. Surgical team behaviors and patient outcomes. Am J Surg. 2009;197(5):678–85.
9. Greenberg CC, Regenbogen SE, Studdert DM, et al. Patterns of communication breakdowns resulting in injury to surgical patients. J Am Coll Surg. 2007;204(4):533–40.
10. Flin R, Martin L. Behavioural marker systems in aviation. Int J Aviat Psychol. 2001;11:95–118.
11. The Non-Technical Skills for Surgeons (NOTSS). System Handbook v1.2: Structuring observation, rating and feedback of surgeons' behaviours in the operating theatre. Aberdeen, Scotland: University of Aberdeen; 2006.
12. Pauley K, Flin R, Yule S, Youngson G. Surgeons' intraoperative decision making and risk management. Am J Surg. 2011;202(4):375–81. https://doi.org/10.1016/j.amjsurg.2010.11.009.
13. Klein G. A Recognition-Primed Decision (RPD) model of rapid decision making. New York: Ablex; 1993.
14. Yule S, Gupta A, Gazarian D, et al. Construct and criterion validity testing of the Non-Technical Skills for Surgeons (NOTSS) behaviour assessment tool using videos of simulated operations. Br J Surg. 2018;105(6):719–27.
15. Haynes AB, Weiser TG, Berry WR, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med. 2009;360(5):491–9. https://doi.org/10.1056/NEJMsa0810119.
16. Haynes AB, Edmondson L, Lipsitz SR, et al. Mortality trends after a voluntary checklist-based surgical safety collaborative. Ann Surg. 2017;266(6):923–9. https://doi.org/10.1097/SLA.0000000000002249.
17. Lagoo J, Berry WR, Miller K, et al. Multisource evaluation of surgeon behavior is associated with malpractice claims. Ann Surg. 2019;270(1):84–90. https://doi.org/10.1097/SLA.0000000000002742.
18. Doumouras AG, Hamidi M, Lung K, et al. Non-technical skills of surgeons and anaesthetists in simulated operating theatre crises. Br J Surg. 2017;104(8):1028–36.
19. Edmondson A. Psychological safety and learning behavior in work teams. Adm Sci Q. 1999;44(2):350–83.
20. Singer SJ, Molina G, Li Z, et al. Relationship between operating room teamwork, contextual factors, and safety checklist performance. J Am Coll Surg. 2016;223(4):568–580.e2. https://doi.org/10.1016/j.jamcollsurg.2016.07.006.
21. Molina G, Berry WR, Lipsitz SR, et al. Perception of safety of surgical practice among operating room personnel from survey data is associated with all-cause 30-day postoperative death rate in South Carolina. Ann Surg. 2017;266(4):658–66.
22. Non-Technical Skills for Surgeons (NOTSS). The Royal College of Surgeons of Edinburgh. https://www.rcsed.ac.uk/professional-support-development-resources/learning-resources/non-technical-skills-for-surgeons-notss. Published 2019. Accessed March 27, 2019.
23. Friends of NOTSS. NOTSS.org. https://www.notss.org/communitynts/. Published 2019. Accessed March 27, 2019.
24. Fletcher G, Flin R, McGeorge M, Glavin R, Maran N, Patey R. Anaesthetists' non-technical skills (ANTS): Evaluation of a behavioural marker system. Br J Anaesth. 2003;90(5):580–8. https://doi.org/10.1093/bja/aeg112.
25. Rutherford JS, Flin R, Irwin A, McFadyen AK. Evaluation of the prototype Anaesthetic Non-technical Skills for Anaesthetic Practitioners (ANTS-AP) system: a behavioural rating sys-

tem to assess the non-technical skills used by staff assisting the anaesthetist. Anaesthesia. 2015;70(8):907–14. https://doi.org/10.1111/anae.13127.

26. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. Med Educ. 2006;40(11):1098–104. https://doi.org/10.1111/j.1365-2929.2006.02610.x.
27. Mitchell L, Flin R, Yule S, Mitchell J, Coutts K, Youngson G. Development of a behavioural marker system for scrub practitioners' non-technical skills (SPLINTS system). J Eval Clin Pract. 2013;19(2):317–23. https://doi.org/10.1111/j.1365-2753.2012.01825.x.
28. Mitchell L, Flin R, Yule S, Mitchell J, Coutts K, Youngson G. Evaluation of the scrub practitioners' list of intraoperative non-technical skills system. Int J Nurs Stud. 2012;49(2):201–11.
29. Russ S, Hull L, Rout S, Vincent C, Darzi A, Sevdalis N. Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. Ann Surg. 2012;255(4):804–9. https://doi.org/10.1097/SLA.0b013e31824a9a02.
30. Mishra A, Catchpole K, Mcculloch P. The Oxford NOTECHS system: reliability and validity of a tool for measuring teamwork behaviour In the operating theatre. Qual Saf Heal Care. 2009;18(2):104–8. https://doi.org/10.1136/qshc.2007.024760.
31. Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent CA. Reliability of a revised NOTECHS scale for use in surgical teams. Am J Surg. 2008;196(2):184–90. https://doi.org/10.1016/j.amjsurg.2007.08.070.
32. Robertson ER, Hadi M, Morgan LJ, et al. Oxford NOTECHS II: a modified theatre team non-technical skills scoring system. PLoS One. 2014;9(3):1–8. https://doi.org/10.1371/journal.pone.0090320.
33. Nagpal K, Abboudi M, Fischler L, et al. Evaluation of postoperative handover using a tool to assess information transfer and teamwork. Ann Surg. 2011;253(4):831–7. https://doi.org/10.1097/SLA.0b013e318211d849.
34. Steinemann S, Berg B, Ditullio A, et al. Assessing teamwork in the trauma bay: introduction of a modified "NOTECHS" scale for trauma. Am J Surg. 2012;203(1):69–75. https://doi.org/10.1016/j.amjsurg.2011.08.004.
35. Frengley RW, Weller JM, Torrie J, et al. The effect of a simulation-based training intervention on the performance of established critical care unit teams. Crit Care Med. 2011;39(12):2605–11. https://doi.org/10.1097/CCM.0b013e3182282a98.
36. Flowerdew L, Gaunt A, Spedding J, et al. A multicentre observational study to evaluate a new tool to assess emergency physicians' non-technical skills. Emerg Med J. 2013;30(6):437–43. https://doi.org/10.1136/emermed-2012-201237.

Collaborative Quality Improvement 13

Margaret E. Smith and Jonathan F. Finks

Abstract

Implementing change at a system level requires a broad, comprehensive approach to quality improvement that engages multiple stakeholders, encourages a culture of knowledge sharing, and takes into account differences in local contexts. It is in this regard that collaborative quality improvement (CQI) efforts are most effective. CQI involves multi-institutional teams who share data through centralized, clinical registries and work collectively to identify best practices. These practices are subsequently implemented and evaluated in the local setting of participants' home institutions. The aim of these collaborative efforts is to improve care by reducing variation among hospitals and providers, minimizing the time lag between changes in knowledge and translation to changes in practice, and evaluating care strategies in real-world settings.

13.1 Introduction

13.1.1 Need for Improvement

Payers, policy makers, and professional societies have increasingly focused on improving surgical care across US hospitals. These efforts are largely driven by a heightened awareness that many patients receive care that is not evidence-based, are harmed by preventable medical errors and growing evidence of wide variation in the quality and cost of inpatient surgery in the US [1, 2]. Millions of patients receive surgical care in US annually, accounting for over $500 billion dollars in costs annually [3]. Complications from surgery are not only harmful to patients but

M. E. Smith (✉) · J. F. Finks
Department of Surgery, University of Michigan Health System, Ann Arbor, MI, USA
e-mail: smargare@umich.edu; jfinks@umich.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_13

also substantially increase the cost of health care, making them a prime target for quality improvement efforts. With some operations, avoidable complications may account for up to 20% of the total cost of inpatient care, with per patient costs exceeding $10,000 [4, 5]. In two widely influential reports, the Institute of Medicine made the case for failures in quality and urged a critical rethinking of our health care systems [6, 7].

13.1.2 Current Strategies in Quality Improvement – The Top Down Approach

In recent years, there have been a number of different efforts by payers and policy makers to promote quality improvement. Incentive-based models, or Pay for Performance (P4P) programs, aim to reward hospitals for adherence to specific evidence-based practices, such as the Surgical Care Improvement Project (SCIP) measures of perioperative antimicrobial use and antithrombotic prophylaxis to reduce postoperative infection and venous thromboembolism (VTE) [8, 9]. More punitive approaches include the non-payment policy of the Centers for Medicare and Medicaid Services (CMS) for complications such as catheter-associated urinary tract and bloodstream infections, [10, 11] and penalties for higher 30-day readmission rates [12]. Other initiatives, such as Center of Excellence models and public reporting of hospital performance data, have focused on steering patients toward high quality hospitals.

Despite the large-scale nature of many of these initiatives, their impact has been somewhat modest. For example, hospital adherence to Medicare's SCIP measures has not been shown to reduce rates of postoperative infection or VTE [8, 9, 13]. Medicare's Premier Hospital Quality Incentive Demonstration, a P4P initiative, was not found to reduce 30-day mortality with coronary artery bypass graft, myocardial infarction, congestive heart failure or pneumonia [14]. Similarly, an evaluation of Medicare's policy of nonpayment for catheter-associated bloodstream and urinary tract infections demonstrated no measurable effect on infection rates in US hospitals [10]. Large systematic reviews of both public reporting [15, 16] and P4P programs [17] have failed to demonstrate evidence of improved care. Finally, hospital accreditation programs, including Medicare's policy to restrict bariatric surgery coverage to hospitals designated as Centers of Excellence, has been shown to reduce complications; [18] however, evidence suggests that outcomes vary widely even across accredited centers [19].

There are several potential explanations for the limited success of these quality improvement (QI) initiatives. First, individual process measures are one small component of factors contributing to outcomes with surgical procedures. Other local factors, such as technical variation with operations, surgeon skill and judgment, operative environment, and organization dynamics, are likely to have a greater impact on patient outcomes. Yet, it is difficult to account for these factors with the administrative data used for most P4P and other QI programs. Furthermore, provider-specific measures are limited by small sample sizes and a lack of clinically rich data sources for adequate risk adjustment, potentially leading to false conclusion regarding hospital quality [20]. There is also the problem of unintended consequences. Public reporting, P4P programs, and non-payment policies may encourage providers to avoid sicker patients [21] and can lead to a decline in the reliability of the administrative data on which they are based, as hospitals modify their billing data to enhance apparent performance [22]. Finally, the one-size-fits-all approach of many of these efforts fails to account for institutional differences in resources and culture, both of which can impact the implementation of QI changes.

13.2 Collaborative Quality Improvement

13.2.1 Defining Collaborative Quality Improvement

Collaborative quality improvement (CQI) initiatives have emerged as an alternative approach to advance the quality of health care. These programs foster a culture of knowledge sharing and take into account differences in local contexts. CQI initiatives are typically regional in scope and involve multi-institutional teams who meet to share data through centralized, clinical registries and work collectively to identify best practices which are then rapidly implemented and evaluated in the local setting of their home institutions [23]. The aim of these collaborative efforts is to improve care by reducing variation among hospitals and providers, minimizing the time lag between changes in knowledge and translation to clinical practice changes, and evaluating care strategies in real-world settings [24].

QI collaboratives are generally centered on a robust clinical registry containing detailed information on patient demographics and comorbidities, as well as provider and hospital characteristics, processes of care and outcomes. Participating institutions receive performance data to allow for benchmarking to other programs. Data elements can be iteratively modified to ensure outcomes of interest are captured, and CQI participants meet at regular intervals to evaluate the data, identify best practices and develop targeted interventions focused on specific clinical problems [5].

Principles of evidence-based medicine, industrial quality management science and organizational theory are utilized by QI collaboratives to generate improvements in health care across multiple institutions. Multi-disciplinary groups from participating institutions identify a particular clinical problem, such as prevention of VTE after bariatric surgery, and best practices are developed from published evidence as well as through exchange of knowledge and experience that occurs at CQI meetings and other activities, such as local site visits and conference calls. Through an iterative process, practice changes are made and evaluated rapidly through frequent reporting of data with analysis and dissemination of results throughout the collaborative. This cycle of intervention, evaluation and adjustment allows for an accelerated process of quality improvement [25].

13.2.2 Advantages of Collaborative Quality Improvement

Behavior is strongly influenced by the social networks in which people participate and changing clinician behavior is more likely to succeed when part of a social process. Evidence suggests that clinicians are more likely to alter behavior when influenced by knowledge gained from peers and their own experience, than that obtained through traditional approaches such as lectures or the threat of legal or other hierarchical sanction. CQIs are therefore unique in that they achieve their results through social interaction and collective learning among a network of people focused on a particular area of interest. Additionally, collaborative decision-making improves the process of adapting strategies to local institutional contexts [26–29].

Another distinct advantage of QI collaboratives is the large sample size of their clinical registries. The statistical power provided from these registries allows for a more robust evaluation of the association between processes and outcomes and the impact of QI initiatives than would be possible with most other intervention studies, including randomized clinical trials (RCT) [5]. The size of QI collaboratives also allows investigators to conduct studies sufficiently powered to identify risk factors for infrequent complications, such as leak after colorectal resection [30]. Additionally, data from collaborative registries can be used to supplement RCTs and better analyze underrepresented subgroups in the trial [31]. Of course, the large size of collaboratives also ensures that QI initiatives and improvements reach a greater number of patients across an entire system or region [32].

13.2.3 Historical Success of CQI - MHA Keystone Center ICU Project

QI collaboratives have been used in health-care related fields for over two decades in disciplines as disparate as patient safety, health care disparities, chronic medical care, pediatrics, and primary care [33]. One of the most successful and well-known examples of CQI is the Michigan Health and Hospital Association (MHA) Keystone Center's Intensive Care Unit (ICU), which involved 67 hospitals and focused on reducing rates of catheter-related blood stream infections (CRBSI) in ICU's across Michigan [34].

The Keystone ICU project began with interventions designed to improve teamwork and communication and to enhance the patient safety culture. These steps included a daily goals sheet and a comprehensive unit-based safety program. These were followed up with a bundle of evidence-based technical interventions focused on reducing rates of CRBSI. Checklists were used to ensure adherence to infection control practices and ICU performance data was rapidly reported back at regular intervals to each hospital to allow for benchmarking. Each ICU had physician and nurse champions who were instructed in the science of safety, data collection and in the specific interventions.

The Keystone ICU project resulted in a significant and sustained decline in rates of CRBSI, with a similar degree of improvement at hospitals across the entire collaborative [35]. (Fig. 13.1) Several factors likely contributed to the remarkable success of the Keystone ICU project. First, the initiative paired a limited number of

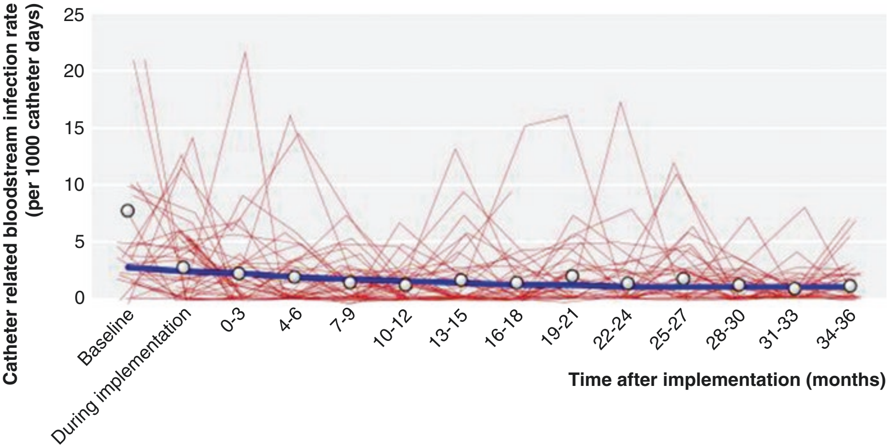


Fig. 13.1 Catheter related bloodstream infection as a function of time following a collaborative intervention in 103 intensive care units in Michigan. *Circles* represent mean infection per quarter; *thick blue line* represents estimated mean rate of infection; *thin red lines* represent changes in observed infection rates over time within a random sample of 50 intensive care units [35].

technical process changes with a broader program designed to influence provider behavior through improved teamwork and communication and an enhanced focus on patient safety. In addition, the agents of change were frontline clinicians within each ICU, thus ensuring that participating ICU's could provide input on the intervention and that the interventions would be optimally adapted for the local environment. Finally, standardization of data collection and the timely feedback of comparative performance data helped maintain team engagement and garner the support of hospital leadership [32].

13.3 Collaborative Quality Improvement in Surgery

Successful CQI initiatives have been developed in surgery; however, the focus is somewhat different than those in medical disciplines. In medicine, numerous evidence-based processes exist and are often compiled into consensus guidelines [36–38]. Improving adherence to published guidelines is a primary objective of many medicine-based collaboratives. The same is not true with surgical collaboratives, as evidence-based guidelines are generally lacking in surgery [20]. Therefore, surgical collaboratives more often focus on determining the drivers of patient outcomes and identifying best practices to optimize those outcomes [20, 39].

13.3.1 Northern New England Cardiovascular Disease Study Group

The first major surgical QI collaborative was the Northern New England Cardiovascular Disease Study Group (NNECDSG). Founded in 1987 as a response to government-mandated public reporting with coronary artery bypass graft (CABG) surgery, the NNECDSG was a voluntary consortium representing all of the hospitals in Maine, New Hampshire and Vermont that performed coronary artery bypass graft (CABG) surgery. Their approach to reducing mortality after CABG surgery provides a beautiful illustration of the way in which regional collaboration can dramatically affect patient care across multiple institutions and settings [40–42].

In 1991, the NNECDSG examined in-hospital mortality following CABG surgery in the region and found substantial variation among hospitals (a 2.5-fold difference between the best and worst hospitals) and surgeons (a 4.2-fold difference) that could not be explained by patient factors alone [43]. They concluded that these differences in mortality most likely represented variation in unmeasured aspects of patient care, leading to the group's first major intervention to reduce CABG mortality across the region.

The first component of the intervention provided continuous performance feedback to the participating institutions, allowing for ongoing self-assessment and benchmarking at each center. Extensive training courses in the techniques of continuous QI for both the collaborative leadership and the general members were also implemented. The third component involved a of round-robin site visits to all centers with visiting teams consisting of industrial engineers, surgeons, nurses and perfusion staff. These benchmarking visits allowed clinical teams from each hospital to learn from each other and ultimately resulted in practice changes that resulted in a 24% reduction in inpatient mortality following CABG, with significant improvement at all of the participating institutions and across all patient subgroups [40, 42].

Following this success, the collaborative launched an effort to identify the factors leading to mortality within 2 years of CABG surgery They found that low-output cardiac failure was the most common mode of death, but also for 80% of the difference in mortality between low-risk and high-risk surgeons [44]. Through further site visits and the inclusion of additional perioperative variables, four process variables associated with a reduced risk for mortality from low-output failure were identified: continuation of preoperative aspirin [45], use of the left internal mammary artery as a bypass conduit [46] avoidance of anemia while on cardiopulmonary bypass [47], and adequacy of beta-blockade-induced heart rate control before induction of anesthesia [48]. Based on these finding, individualized care protocols were instituted, mortality from low-output failure declined across the region from 1.24 to 0.72% [41].

13.3.2 Surgical Care and Outcomes Assessment Program

Another notable CQI project is the Surgical Care and Outcomes Assessment Program (SCOAP), developed in Washington State in 2003. The collaborative, largely funded by participating hospitals, collects data from over 60% of hospitals in the state and focuses on improving outcomes in general surgery, vascular surgery, and interventional radiology [49, 50].

As with other CQI programs, SCOAP has a standardized clinical data collection platform that contains information on patient characteristics, process measures and procedure-specific outcomes for all patients undergoing the selected procedures at participating hospitals. A strong emphasis was placed on tracking optimal processes of care, with some linked to evidence and others determined by consensus. These quality metrics include processes such as continuing beta-blocker use in the perioperative period, routine intraoperative leak testing after colorectal resection, using diagnostic imaging in patients with presumed appendicitis, and using appropriate neoadjuvant therapy for patients with rectal cancer. Adherence to these surgeon-determined process measures is reinforced by means of operating room checklists, preprinted order sets, e-newsletters and regional meetings. Efforts to correct underperformance occur largely through education and peer support/pressure, often with peer-led interventions focusing on sharing best practices and creating behavior change around quality metrics.

Since its launch, SCOAP has registered a number of achievements through targeted interventions such as a significant reduction in the rate of negative appendectomies (Fig. 13.2) [51] and adverse events following elective colorectal resection (Fig. 13.3). When compared to non-SCOAP institutions, hospitals participating in SCOAP have significantly reduced the costs associated with appendectomy, colorectal and bariatric operations [50]. (Fig. 13.4).

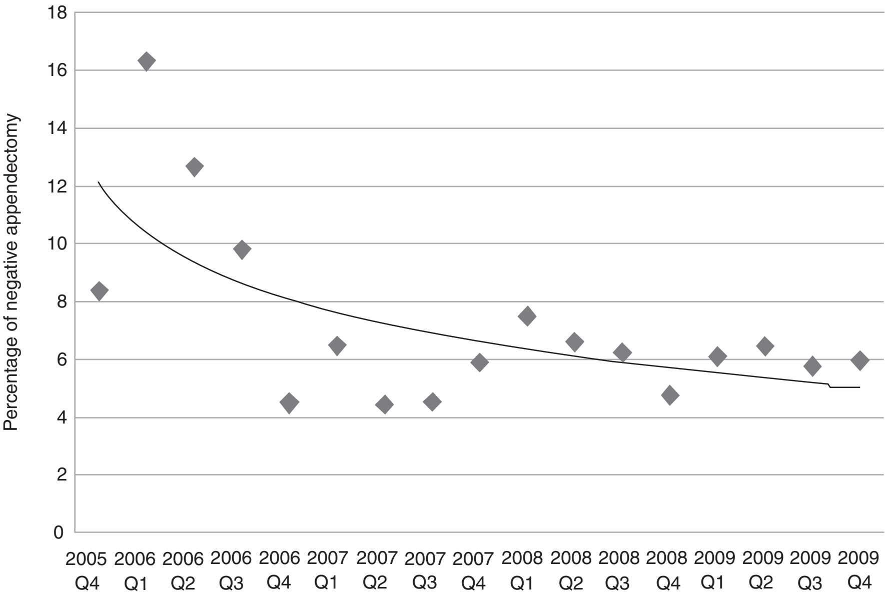


Fig. 13.2 Negative appendectomy rates, by calendar quarters, among hospitals participating in Washington State's Surgical Care and Outcomes Assessment Program [50]

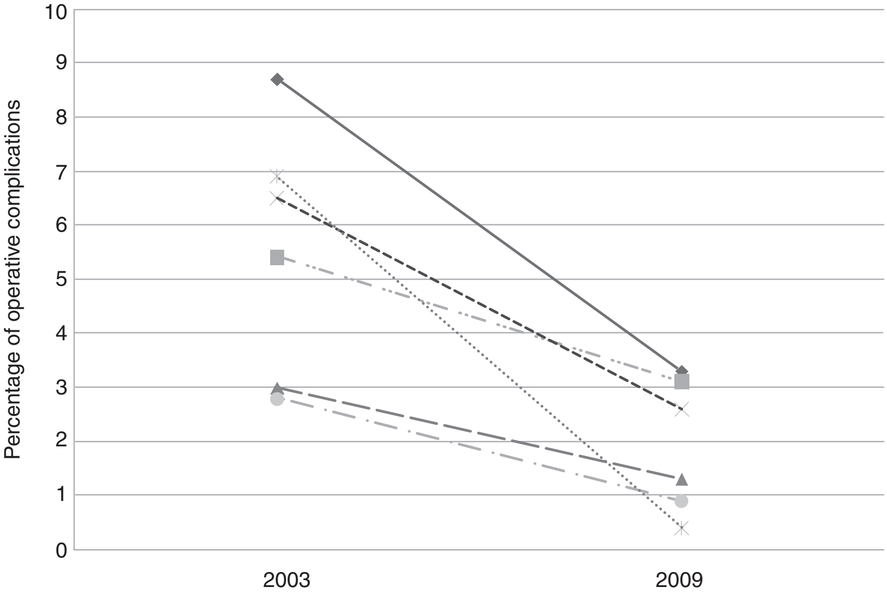


Fig. 13.3 Rates of operative complications in elective colorectal operations in sites (n = 6) that eventually joined the Surgical Care and Outcomes Assessment Program [50]

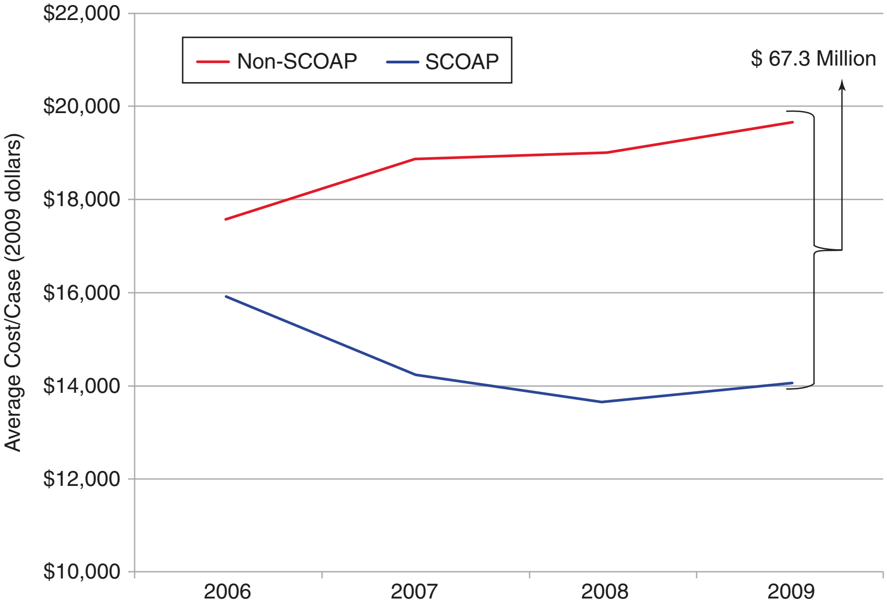


Fig. 13.4 Average cost per case for appendectomy, colorectal and bariatric operations by calendar year, among hospitals participating in the Surgical Care and Outcomes Assessment Program [50]

13.3.3 Payer-Funded Collaborative Quality Improvement—Pay for Participation

A major challenge for CQI efforts is funding. The costs for starting and maintaining large multicenter CQI programs are substantial, particularly with regard to data collection, creating a significant financial burden that is potentially prohibitive, especially for smaller hospitals. At the same time, complications are very expensive and their cost is borne largely by payers [52]. In states with at least one dominant payer; therefore, there is a strong business case to be made for payer-supported CQI programs, since even a small reduction in complications can result in substantial cost savings for the payer [53]. This model of quality improvement has been in place in Michigan for over a decade.

Since 2004, Blue Cross and Blue Shield of Michigan/Blue Care Network (BCBSM/BCN) has partnered with Michigan providers and hospitals to support statewide registry-based CQI programs in a number of different disciplines Value Partnership Program (VPP) [5, 54]. The insurer invests over $30 million annually to fund 17 programs across the state, collectively encompassing the care of over 200,000 patients each year. The CQI focuses primarily on clinical conditions that are common and associated with high episode cost as well as procedures that are technically complex, rapidly evolving and associated with wide variation in hospital practice and outcomes. Current collaboratives include general surgery, vascular surgery, cardiothoracic surgery, bariatric surgery, trauma, prostate cancer, joint replacement, spine surgery and more.

In this Pay for Participation model, the cost for administering the VPP comes largely in the form of payments to hospitals, most of which participate in several different CQI programs. In exchange for these supplemental payments, hospitals are expected to submit timely, accurate data to the coordinating center and allow regular site visits from data auditors. Each hospital is also required to send a physician champion and program coordinator to the quarterly meetings held by each collaborative and is expected to actively engage in regional quality improvement interventions [5].

The coordinating center for each collaborative maintains a clinical registry containing high quality clinical outcomes data, including information on patient characteristics necessary for risk adjustment, procedure-specific processes of care and relevant outcomes. The data are prospectively collected by trained abstractors using standardized definitions and are externally audited annually to ensure accuracy and completeness. Hospitals and surgeons are provided with timely feedback on their performance, benchmarked against other providers in the collaborative. That performance data is not publicly reported or released to the payer. Rather, these data are used to drive QI initiatives that are implemented at all participating hospitals under the direction of local program coordinators [20].

13.3.4 Michigan Surgical Quality Collaborative

Since their inception in 2004, the Michigan CQI programs have resulted in improvements across a wide range of clinical conditions. The largest CQI is the Michigan

Surgery Quality Collaborative (MSQC), which includes general and vascular surgery procedures. The MSQC primarily focuses on aspects of perioperative care, including specific practices designed to reduce VTE and surgical site infections. In a study designed to evaluate the added value of the CQI model, hospitals participating in the MSQC were compared to non-Michigan hospitals participating in the American College of Surgeons' National Surgical Quality Improvement Program (NSQIP). In the period between 2005 and 2009, risk-adjusted complication rates at MSQC hospitals fell from 13.1% to 10.5% ($p < 0.001$), while the complication rate at non-Michigan NSQIP hospitals remained relatively flat between 2005 and 2008 with a modest decline in 2009 (Fig. 13.5). The 2.6% decline in morbidity observed in the MSQC hospitals represents approximately 2500 fewer patients with surgical complications annually, or an annual cost savings of roughly $20 million, far more than the $5 million annual cost of administering the MSQC [5].

13.3.5 Michigan Bariatric Surgery Collaborative

In some instances, the quality improvement and cost savings have come not only from reducing rates of adverse outcomes but also from reducing unnecessary procedures. The Michigan Bariatric Surgery Collaborative (MBSC), launched in 2006, collects data from 42 hospitals (over 95% of those performing bariatric surgery) and nearly 6000 patients annually.

One of the first MBSC initiatives grew from the observation that almost 10% of patients undergoing bariatric surgery had an IVC filter placed preoperatively to prevent VTE, and that IVC filter used varied widely from 0% to 35% across the state. Analysis of the data from the MBSC revealed that IVC filter use not only failed to

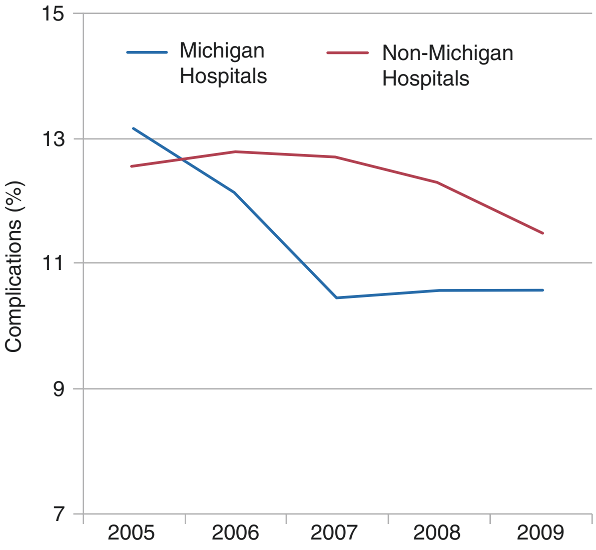


Fig. 13.5 Risk-adjusted morbidity with general and vascular surgery: Hospitals in Michigan versus hospitals outside of Michigan, 2005–2009 (Source: Michigan Surgical Quality Collaborative and National Surgical Quality Improvement Program registries, 2005–2009 [5]

reduce the risk for VTE, but itself was a source of complications, such as filter migration [55, 56]. Feedback of this data with discussions at tri-annual CQI meetings prompted a QI initiative leading to a collaborative-wide effort to discontinue IVC filter use. Simultaneously, evaluation of risk factors for VTE resulted in development of statewide guidelines for risk-stratified VTE prophylaxis. Within a year of implementation across the collaborative, IVC filter use had dropped to less than 2% of patients. With an estimated average cost of $13,000 for IVC filter placement, the decline of this unnecessary procedure resulted in an estimated annual cost savings of $four million and was accompanied by a significant decline in VTE-related deaths. Additionally, between 2007 and 2009, overall mortality rates after bariatric surgery in Michigan hospitals declined significantly faster than in non-Michigan hospitals. (p = .045) [5] (Fig. 13.6).

The trust fostered over years of collaboration resulted in MBSC surgeons voluntarily submitting surgical videos for evaluation of technical skill. Analysis of the videos demonstrated that surgeon skill, as measured by peer bariatric surgeons, was highly associated with morbidity and mortality after gastric bypass. (Fig. 13.7) [55] As a result, a video-based peer-to-peer coaching pilot project was implemented within the MBSC. Through 1-hour sessions at CQI meetings, coaching pairs identified performance goals, evaluated current performance and developed action plans to help lower skilled surgeons achieve designated goals [57]. Videos have also been used to identify the best technical practices based on their association with clinical outcomes [58].

Similar successful QI interventions have been implemented across the other Michigan CQI programs. For example, implementation of practice guidelines and the use of bedside tools for risk assessment with percutaneous coronary interventions led to reductions in contrast-induced nephropathy, transfusions, strokes and vascular complications associated with this procedure [59]. Furthermore, a series of specific, focused interventions in cardiac surgery resulted in a reduction in the use of two expensive therapies: intra-aortic balloon pumps and prolonged mechanical

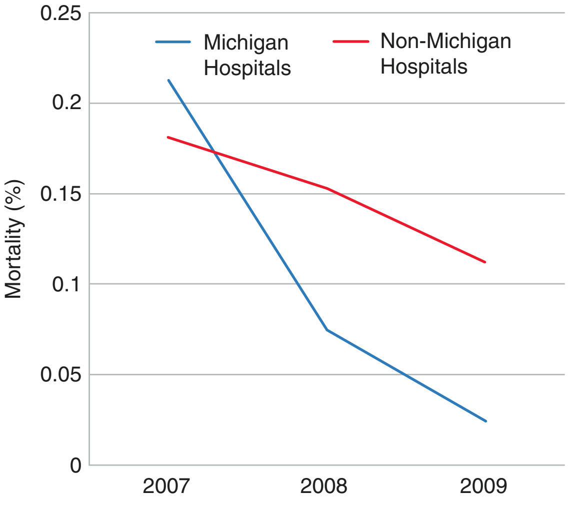


Fig. 13.6 Thirty-day mortality after bariatric surgery: Hospitals in Michigan versus hospitals outside of Michigan, 2007–2009 (Source: Michigan Bariatric Surgery Collaborative and National Surgical Quality Improvement Program registries, 2007–2009 [5]

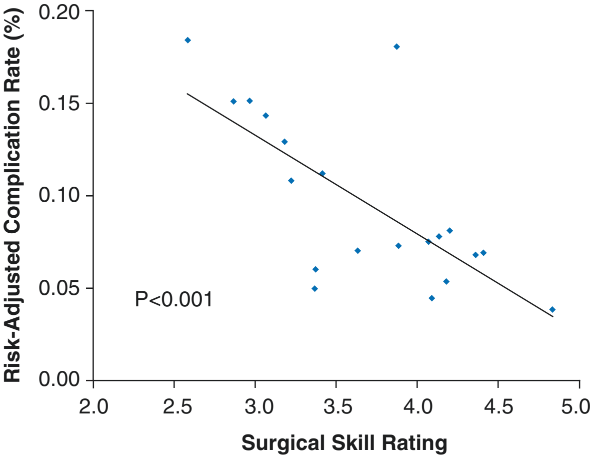


Fig. 13.7 Association of summary peer rating of technical skills and risk-adjusted complication rates following Roux-en-Y gastric bypass [55].

ventilation [60]. Finally, a QI initiative centered on comparative performance feedback and dissemination of practice guidelines in urology resulted in improved adherence with recommended staging practices and a decline in the use of expensive bone and computerized tomography scans for surveillance of prostate cancer in low and intermediate risk tumors [61]. Based on successes in Michigan, similar payer-funded collaboratives have started in at least 4 other states [54].

13.4 Opportunities and Challenges to Regional Collaborative Quality Improvement

Partnerships between surgeons, hospitals and payers can be a win for all. Large payers, including CMS, have the capital resources to provide infrastructure to support collaboration among surgeons and hospitals, as well as the resources necessary for interventions involving large numbers of patients and hospitals. Furthermore, payers may have the political influence with hospitals to ensure broad participation in CQI efforts [20, 62]. Payers reap the benefits from this partnership model in terms of a significant reduction in costs from adverse events and unnecessary tests and procedures. Hospitals receive compensation for their participation. Surgeons and other health care providers benefit from the professional satisfaction that comes from collaborative learning and interaction with colleagues. Most important of all, patients receive better care.

Despite these benefits, CQI efforts face a number of challenges. Competition between regional centers can hinder the sense of community that is fundamental to a successful collaborative. The process of engagement and development of mutual trust takes time, considerable effort and strong, effective clinical leadership. Additionally, sustained quality improvement requires a high level of commitment from participating centers as well as surgeons, and local barriers can be difficult to overcome. Finally, CQI efforts rely on a clinically rich and accurate patient registry, which is expensive to maintain. Without payer support, the cost for data entry falls to the hospitals, which may have limited ability to pay for it [26, 50].

13.5 Keys to Success with Collaborative Quality Improvement

Although the hallmark of QI collaboratives is a bottom-up, participatory approach to decision making, a major contributor to success with CQI efforts is strong central leadership. Collaboratives represent a clinical community with members often from diverse professional backgrounds. It is the leader's responsibility to ensure the cohesiveness of the community and coordinate the group's efforts [26]. The leader must possess expertise in epidemiology and the science of quality improvement, while also being able to communicate a shared vision and sense of purpose among participants [32, 63, 64]. For a number of reasons, leaders of QI collaboratives must be viewed as credible, authoritative and worthy of trust by the members of the collaborative. Members need to trust that the leadership is not driven by other objectives (research, commercial, political, etc.), and leaders need the support of the group to challenge practices once thought to be routine [26]. Taking on the routine use of preoperative IVC filters in bariatric surgery patients exemplifies just such a challenge for the leadership of the MBSC.

Another key role of leadership within collaboratives is to develop consensus and foster a sense of community. The expectations and goals of individuals within a collaborative (e.g. physicians, nurses, administrators, etc.) may be quite disparate. Failure to obtain consensus regarding the goals and objectives of the collaborative can result in declining morale among some groups and a reduction in coordinated, effective action. Dedicated time for inclusive debate on objectives for the collaborative and on individual interventions will enhance the sense of community and encourage individual groups to marshal their own resources in support of the collective interests of the collaborative. As clinicians are far more likely to change their behavior if they participated in designing the intervention, the task of ensuring that all voices are heard falls on the shoulders of leaders within the collaborative [26].

Finally, one of the most important determinants of success of a CQI initiative is an understanding of local contextual factors that may influence adoption of a particular intervention. The existing culture, relationships and resources within an individual organization will affect the outcome of a given strategy or approach. To some extent, the impact of these factors will be mitigated through the process of building consensus around development of the intervention, as local issues will help inform this process. Some centers, however, may require additional resources or support, such as extra educational materials, peer site visits and/or team training interventions for centers that are falling behind. There may also be a role for the CQI leadership to assist clinical champions in influencing their organization's leadership to support and implement QI initiatives. At very least, open discussion of local barriers and enablers will assist in refining care over time [26].

13.6 Conclusion

Collaborative quality improvement is an efficient and powerful tool for sharing knowledge and advancing innovation through a process of collective learning. Access to rich clinical data from a large patient sample drawn from multiple institutions allows for problem identification, robust assessment of the relationship between processes and outcomes, intervention development and implementation, and rapid evaluation and iterative refinement of those changes across multiple centers. CQI also allows interventions to reach large numbers of patients at once and creates opportunities for individual centers to improve more rapidly than they could on their own. Given that interventions are designed through participatory discussions, they are more likely to be adopted and are often more adaptable to local contexts. Finally, partnerships with payers, government agencies and national societies may prove critical to the long-term success of these initiatives.

A number of areas will require further investigation if CQI efforts are to reach their potential for dramatic and long-lasting improvement. Determining which organizational attributes promote successful implementation of CQI efforts is necessary to ensure sustainability of improvements made through CQI-driven interventions. Additionally, CQI efforts should be carefully compared with other QI efforts to establish the arenas where CQI initiatives are likely to be most effective. Overall, CQI programs epitomize what the National Academy of Medicine described as a "learning health system", [65] and to improve care and reduce cost on a large scale, there is likely no better investment.

References

1. Birkmeyer JD, Gust C, Dimick JB, et al. Hospital quality and the cost of inpatient surgery in the United States. Ann Surg. 2012;255:1–5.
2. Wakeam E, Molina G, Shah N, et al. Variation in the cost of 5 common operations in the United States. Surgery. 2017;162:592–604.
3. Munoz E, Munoz W 3rd, Wise L. National and surgical health care expenditures, 2005-2025. Ann Surg. 2010;251:195–200.
4. Dimick JB, Chen SL, Taheri PA, et al. Hospital costs associated with surgical complications: a report from the private-sector National Surgical Quality Improvement Program. J Am Coll Surg. 2004;199:531–7.
5. Share DA, Campbell DA, Birkmeyer N, et al. How a regional collaborative of hospitals and physicians in Michigan cut costs and improved the quality of care. Health Aff (Millwood). 2011;30:636–45.
6. Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, D.C.: National Academy Press; 2001.
7. Kohn LT, Corrigan JM, Donaldson MS. Kohn LT, Corrigan JM, Donaldson MS (Institute of Medicine) To err is human: building a safer health system. Washington, D.C: National Academy Press; 2000.
8. Chang V, Blackwell RH, Markossian T, et al. Discordance between surgical care improvement project adherence and postoperative outcomes: implications for new joint commission standards. J Surg Res. 2017;212:205–13.
9. Stulberg JJ, Delaney CP, Neuhauser DV, et al. Adherence to surgical care improvement project measures and the association with postoperative infections. JAMA. 2010;303:2479–85.

10. Lee GM, Kleinman K, Soumerai SB, et al. Effect of nonpayment for preventable infections in U.S. hospitals. N Engl J Med. 2012;367:1428–37.
11. Rajaram R, Chung JW, Kinnier CV, et al. Hospital characteristics associated with penalties in the centers for Medicare & Medicaid Services Hospital-Acquired Condition Reduction Program. JAMA. 2015;314:375–83.
12. Joynt KE, Jha AK. Thirty-day readmissions--truth and consequences. N Engl J Med. 2012;366:1366–9.
13. Hawn MT, Vick CC, Richman J, et al. Surgical site infection prevention: time to move beyond the surgical care improvement program. Ann Surg. 2011;254:494–9. discussion 499–501
14. Jha AK, Joynt KE, Orav EJ, Epstein AM. The long-term effect of premier pay for performance on patient outcom es. N Engl J Med. 2012;366:1606–15.
15. Fung CH, Lim YW, Mattke S, et al. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med. 2008;148:111–23.
16. Ketelaar NA, Faber MJ, Flottorp S, et al. Public release of performance data in changing the behaviour of healthcare consumers, professionals or organisations. Cochrane Database Syst Rev. 2011:CD004538.
17. Houle SKMF, Jackevicius CA, Chuck AW, Tsuyuki RT. Does performance-based remuneration for individual health care practitioners affect patient care? A systematic review. Ann Intern Med. 2012;157(12):889–99. https://doi.org/10.7326/0003-4819-157-12-201212180-00009.
18. Morton J. The first metabolic and bariatric surgery accreditation and quality improvement program quality initiative: decreasing readmissions through opportunities provided. Surg Obes Relat Dis. 2014;10:377–8.
19. Ibrahim AM, Ghaferi AA, Thumma JR, Dimick JB. Variation in outcomes at bariatric surgery centers of excellence. JAMA Surg. 2017;152:629–36.
20. Birkmeyer NJ, Birkmeyer JD. Strategies for improving surgical quality--should payers reward excellence or effort? N Engl J Med. 2006;354:864–70.
21. Werner RM, Asch DA. The unintended consequences of publicly reporting quality information. JAMA. 2005;293:1239–44.
22. Farmer SA, Black B, Bonow RO. Tension between quality measurement, public quality reporting, and pay for performance. JAMA. 2013;309:349–50.
23. Eppstein MJ, Horbar JD, Buzas JS, Kauffman SA. Searching the clinical fitness landscape. PLoS One. 2012;7:e49901.
24. Kilo CM. A framework for collaborative improvement lessons from the Institute for Healthcare Improvement's Breakthrough Series. Qual Manag Health Care. 1998;6:1–13.
25. Benn J, Burnett S, Parand A, et al. Factors predicting change in hospital safety climate and capability in a multi-site patient safety collaborative: a longitudinal survey study. BMJ Qual Saf. 2012;21:559–68.
26. Aveling EL, Martin G, Armstrong N, et al. Quality improvement through clinical communities: eight lessons for practice. J Health Organ Manag. 2012;26:158–74.
27. Parboosingh JT. Physician communities of practice: where learning and practice are inseparable. J Contin Educ Heal Prof. 2002;22:230–6.
28. Shaw EK, Chase SM, Howard J, et al. More black box to explore: how quality improvement collaboratives shape practice change. J Am Board Fam Med. 2012;25:149–57.
29. Stoopendaal A, Bal R. Conferences, tablecloths and cupboards: how to understand the situatedness of quality improvements in long-term care. Soc Sci Med. 2013;78:78–85.
30. Matthews JB. Risky business? Collaborative databases and quality improvement. Arch Surg. 2012;147:605–6.
31. Gurm HS, Smith DE, Berwanger O, et al. Contemporary use and effectiveness of N-acetylcysteine in preventing contrast-induced nephropathy among patients undergoing percutaneous coronary intervention. JACC Cardiovasc Interv. 2012;5:98–104.
32. Watson SR, Scales DC. Improving intensive care unit quality using collaborative networks. Crit Care Clin. 2013;29:77–89.
33. Livingood W, Marshall N, Peden A, et al. Health districts as quality improvement collaboratives and multijurisdictional entities. J Public Health Manag Pract. 2012;18:561–70.

34. Pronovost P, Needham D, Berenholtz S, et al. An intervention to decrease catheter-related bloodstream infections in the ICU. N Engl J Med. 2006;355:2725–32.
35. Pronovost PJ, Goeschel CA, Colantuoni E, et al. Sustaining reductions in catheter related bloodstream infections in Michigan intensive care units: observational study. BMJ. 2010;340:c309.
36. Carlhed R, Bojestig M, Peterson A, et al. Improved clinical outcome after acute myocardial infarction in hospitals participating in a Swedish quality improvement initiative. Circ Cardiovasc Qual Outcomes. 2009;2:458–64.
37. Crandall WV, Margolis PA, Kappelman MD, et al. Improved outcomes in a quality improvement collaborative for pediatric inflammatory bowel disease. Pediatrics. 2012;129:e1030–41.
38. Powell AA, Nugent S, Ordin DL, et al. Evaluation of a VHA collaborative to improve follow-up after a positive colorectal cancer screening test. Med Care. 2011;49:897–903.
39. Cross RR, Harahsheh AS, McCarter R, Martin GR. Identified mortality risk factors associated with presentation, initial hospitalisation, and interstage period for the Norwood operation in a multi-Centre registry: a report from the national pediatric cardiology-quality improvement collaborative. Cardiol Young. 2014;24:253–62.
40. Likosky DS, Nugent WC, Ross CS. Improving outcomes of cardiac surgery through cooperative efforts: the northern new England experience. Semin Cardiothorac Vasc Anesth. 2005;9:119–21.
41. Nugent WC. Building and supporting sustainable improvement in cardiac surgery: the northern New England experience. Semin Cardiothorac Vasc Anesth. 2005;9:115–8.
42. O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. The northern New England cardiovascular disease study group. JAMA. 1996;275:841–6.
43. O'Connor GT, Plume SK, Olmstead EM, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. The northern New England cardiovascular disease study group. JAMA. 1991;266:803–9.
44. O'Connor GT, Birkmeyer JD, Dacey LJ, et al. Results of a regional study of modes of death associated with coronary artery bypass grafting. Northern New England Cardiovascular Disease Study Group. Ann Thorac Surg. 1998;66:1323–8.
45. Dacey LJ, Munoz JJ, Johnson ER, et al. Effect of preoperative aspirin use on mortality in coronary artery bypass grafting patients. Ann Thorac Surg. 2000;70:1986–90.
46. Leavitt BJ, O'Connor GT, Olmstead EM, et al. Use of the internal mammary artery graft and in-hospital mortality and other adverse outcomes associated with coronary artery bypass surgery. Circulation. 2001;103:507–12.
47. DeFoe GR, Ross CS, Olmstead EM, et al. Lowest hematocrit on bypass and adverse outcomes associated with coronary artery bypass grafting. Northern New England cardiovascular disease study group. Ann Thorac Surg. 2001;71:769–76.
48. Fillinger MP, Surgenor SD, Hartman GS, et al. The association between heart rate and in-hospital mortality after coronary artery bypass graft surgery. Anesth Analg. 2002;95:1483–8. table of contents
49. Flum DR, Fisher N, Thompson J, et al. Washington State's approach to variability in surgical processes/outcomes: surgical clinical outcomes assessment program (SCOAP). Surgery. 2005;138:821–8.
50. Kwon S, Florence M, Grigas P, et al. Creating a learning healthcare system in surgery: Washington State's surgical care and outcomes assessment program (SCOAP) at 5 years. Surgery. 2012;151:146–52.
51. Cuschieri J, Florence M, Flum DR, et al. Negative appendectomy and imaging accuracy in the Washington state surgical care and outcomes assessment program. Ann Surg. 2008;248:557–63.
52. Dimick JB, Weeks WB, Karia RJ, et al. Who pays for poor surgical quality? Building a business case for quality improvement. J Am Coll Surg. 2006;202:933–7.
53. Englesbe MJ, Dimick JB, Sonnenday CJ, et al. The Michigan surgical quality collaborative: will a statewide quality improvement initiative pay for itself? Ann Surg. 2007;246:1100–3.
54. Campbell DA Jr, Krapohl GL, Englesbe MJ. Conceptualizing partnerships between private payers and Medicare for quality improvement initiatives. JAMA Surg. 2018;153:4–5.

55. Birkmeyer NJ, Finks JF, English WJ, et al. Risks and benefits of prophylactic inferior vena cava filters in patients undergoing bariatric surgery. J Hosp Med. 2013;8:173–7.
56. Birkmeyer NJ, Share D, Baser O, et al. Preoperative placement of inferior vena cava filters and outcomes after gastric bypass surgery. Ann Surg. 2010;252:313–8.
57. Greenberg CC, Dombrowski J, Dimick JB. Video-based surgical coaching: an emerging approach to performance improvement. JAMA Surg. 2016;151:282–3.
58. Varban OA, Niemann A, Stricklen A, et al. Far from standardized: using surgical videos to identify variation in technique for laparoscopic sleeve gastrectomy. J Laparoendosc Adv Surg Tech A. 2017;27:761–7.
59. Moscucci M, Rogers EK, Montoye C, et al. Association of a continuous quality improvement initiative with practice and outcome variations of contemporary percutaneous coronary interventions. Circulation. 2006;113:814–22.
60. Prager RL, Armenti FR, Bassett JS, et al. Cardiac surgeons and the quality movement: the Michigan experience. Semin Thorac Cardiovasc Surg. 2009;21:20–7.
61. Miller DC, Murtagh DS, Suh RS, et al. Regional collaboration to improve radiographic staging practices among men with early stage prostate cancer. J Urol. 2011;186:844–9.
62. Scales DC. Partnering with health care payers to advance the science of quality improvement: lessons from the field. Am J Respir Crit Care Med. 2011;184:987–8.
63. Harris Y, Kwon L, Berrian A, Calvo A. Redesigning the system from the bottom up: lessons learned from a decade of federal quality improvement collaboratives. J Health Care Poor Underserved. 2012;23:11–20.
64. Palmer C, Bycroft J, Healey K, et al. Can formal collaborative methodologies improve quality in primary health care in New Zealand? Insights from the EQUIPPED Auckland collaborative. J Prim Health Care. 2012;4:328–36.
65. Ferguson TB Jr. The Institute of Medicine committee report "best care at lower cost: the path to continuously learning health care". Circ Cardiovasc Qual Outcomes. 2012;5:e93–4.

Machine Learning and Artificial Intelligence

14

David F. Schneider

Abstract

Interest in artificial intelligence for use in healthcare and health services research is growing as the amount of data available is ever increasing. In this chapter, we define the terminology surrounding artificial intelligence. Machine learning methods are the building blocks for artificial intelligence, and we provide an overview of selected methods useful in healthcare and health services research. Cutting edge "deep learning" hold particular promise for image analysis and natural language processing. We review two examples to illustrate the features of deep learning that make it a powerful tool for research and clinical applications.

14.1 Introduction

Why has artificial intelligence generated so much interest for those working in medicine and health services research? While current events may have prompted some of the interest, the application of artificial intelligence to human medicine also grew out of the need to process ever-increasing amounts of information. Traditional statistical modeling performs quite well when the user knows everything. That is, the investigator knows all the known predictor variables and all the interactions between these variables. In today's medical environment, providers encounter an ever-increasing number of predictor variables, and with new information, it is not always clear a priori which features are important. Furthermore, the interactions between all these variables are not obvious. Machine learning is well suited for such a large number of features and infinite number of interactions [1].

D. F. Schneider (✉)
Department of Surgery, University of Wisconsin School of Medicine and Public Health,
Madison, WI, USA
e-mail: schneiderd@surgery.wisc.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_14

Furthermore, artificial intelligence can continuously improve as more data and more features become available. As the number of features increases, human intuition becomes less accurate, and we succumb to a number of biases [2]. Not every task is well suited for the complex algorithms of artificial intelligence. Yet, as the amount and variety of information increases, so too will the need for artificial intelligence tools that can help us process, select, and model the data to make better decisions. This chapter is intended to serve as a very brief introduction to the methods of artificial intelligence, with a special focus on methods well suited surgery and health services research.

14.2 Clarification of Terms

> *Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.*
> *—Dan Ariely, Duke University*

The terminology surrounding "artificial intelligence" is often misused, leading to confusion or even fear, especially as these terms apply to healthcare and health services research. Before discussing machine learning and artificial intelligence, it is important to clarify these terms.

Artificial intelligence is a field of computer science that aims to make computers reason and act more like humans [3]. The lay public assumes the computer's skills will far exceed human capabilities in terms of intelligence, speed, and scale. In truth, current work in artificial intelligence aims to produce results equivalent to human capability. With equivalent results, computers will likely exceed human capability in terms of speed and scale, but it remains unclear whether computers can achieve superior intelligence.

The primary methodology behind artificial intelligence is machine learning. **Machine learning** is really a collection of methods for inferring predictive models from sets of training instances. Stated more simply, these are methods for training a computer to predict "unknowns" from a set of "knowns." There are many broad categories of machine learning methods. First, computer scientists distinguish supervised from unsupervised machine learning methods. **Unsupervised machine learning** discovers patterns from a dataset that has not been classified, labeled, or categorized. Commonly, unsupervised machine learning methods cluster the cases in a dataset by their similarity or differences of their features (variables) [1, 3, 4].

Most applications of machine learning, especially those employed for artificial intelligence, rely on supervised machine learning. In **supervised machine learning** the dataset is labeled or classified, typically by the outcome of interest. For example, if one were to "train" a machine learning algorithm to recognize sepsis from a set of patient vitals, the first step would be to label each patient, or case, as septic vs not septic. In this chapter, we will focus on supervised machine learning. Numerous types of machine learning methods exist, and each type is characterized by the underlying mathematical function or routine employed. Examples include rule-based, tree-based, Bayesian, neural networks, and support vector machine methods [4, 5]. In this chapter, we will provide a brief overview of selected machine learning commonly used in medical research.

Natural language processing refers to sets of instructions or algorithms that allow computers to recognize and interpret human language, and machine learning is one approach for accomplishing this task. Therefore, natural language processing is one specific application of machine learning. Other natural language processing approaches take a symbolic, grammatical, or rule-based approach [6, 7].

When machine learning and natural language processing are discussed, the term "big data" often gets used (or misused). **Big Data** refers to large and complex datasets prohibited from being processed with common or traditional database management tools and traditional data processing applications. "Traditional" database management software includes Microsoft Excel or Access while traditional data processing software includes R or STATA. There are four dimensions of Big Data that define it: volume, variety, veracity, and velocity [3]. Hence, big data is not just about size. Although the volume is large, the data features are complex and heterogeneous (e.g, text documents, structured numerical data, and categorical variables). Big Data changes, updates, or multiplies very quickly (velocity). Often, data scientists working with big data will utilize machine learning and natural language processing, but these particular tools are not a requirement. Analyzing SEER Medicare data to describe trends in imaging utilization for patients with pancreatic cancer is *not* big data. However, processing data from a consortium of 100 hospitals' continuously updated electronic health records with genomic, text, structured data, and data from wearable technology *is* big data.

Finally, all of the terms discussed above constitute some of the tools and science in **biomedical informatics** - the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, driven by efforts to improve human health. **Clinical informatics** is the practice of biomedical informatics in healthcare including medical, dental, and nursing as applied to patients and promoting human health [8, 9].

To describe the tools of machine learning and artificial intelligence, we also need to define some basic terms because some of these terms have different meanings in the computer science realm. A **case** is an instance of a dataset. In health services research, a case is usually a patient. The patient is described by many variables, or **features**, such as their age, gender, co-morbidities, weight, medications, and laboratory values. A **method** refers to a routine or algorithm. Hence, machine learning and artificial intelligence are really a collection of different methods. In this chapter we use method and **algorithm** interchangeably.

In the remainder of this chapter, we will first discuss the general workflow for machine learning. Then, we review some of the common machine learning methods used in medical applications of artificial intelligence. Finally, we review two specific use cases of artificial intelligence in medicine: computer vision and natural language processing.

14.3 General Workflow

Machine learning requires lots of data to develop an accurate model. Providing lots of examples in the training dataset ensures that all variations or "edge cases" are represented. As most health services researchers realize, data comes in many different formats, and the first step toward analyzing all this data is to clean it and format it such that your statistical software can work with it. Machine learning is no different. In the case of a classification problem, the requisite expert should label the cases according to the output of interest (e.g., diseased vs healthy, or phenotype A vs phenotype B). Once the data is cleaned, formatted, and labeled correctly, other pre-processing may include imputation for missing data, feature selection, and data transformation. Many features depend on one another or contain redundant information. Selecting meaningful features and transforming them often leads to better comprehensibility of the learned algorithm and its outcome(s) [1, 4, 10]. For example, continuous laboratory data might be converted into a binary feature of "normal" vs "abnormal."

The next step is to separate a training set from a testing set. An investigator uses the training set to test different machine learning methods and optimize their parameters all to identify the most accurate method. Next, the investigator runs the selected method on a separate, held out test set of data to formally assess and report its accuracy. In this way, the selected machine learning algorithm is evaluated on data it has never "seen" previously. Therefore, it is best to separate the test set from the training set very early in the workflow, and maintain strict standards for its access and use [3, 4, 11].

Aside from a held out test set, another way to evaluate a machine learning algorithm is with k-fold cross validation. Cross validation is simply a resampling procedure where k indicates the number of groups the data set is split into for evaluation. For example, in ten-fold cross validation, the data is split into 10 random groups, where one serves as the testing set and the remaining nine form the training set. The procedure is repeated 10 times. Cross validation is popular for evaluating machine learning because its generalizable and less biased as the evaluation really occurs multiple times with different testing sets. The disadvantage of cross validation is that the training and testing sets all come from the same pool of data. One might use both cross validation and a held out testing set if cross validation is used in developing and tuning a machine learning algorithm and then formally evaluated with a separate held out testing set [4, 11].

Once the investigator identifies a training set, the next step is algorithm selection. Unlike other statistical approaches, the investigator will often run several different machine learning experiments to find the most accurate method [10]. One major consideration of the methods tested is the outcome variable. Broadly speaking, some machine learning methods can handle binary or categorical outcomes while others deal with continuous outcomes. Although accuracy becomes a primary driver for the specific algorithm selected, the investigator should also consider the audience or user of the algorithm and the desired presentation of their work. Certain methods like neural networks are opaque in terms of which features are utilized and the underlying relationships between these features. Other machine learning approaches like tree-based methods or Bayesian networks provide a visual output that displays the relationship between different features. Often clinicians are uncomfortable with more opaque methods, so if the machine learning method is for clinical decision support, one might choose a method with visual outputs even if the accuracy is inferior to something more opaque like a neural network.

After selecting the most accurate method, its parameters and hyperparameters are tuned to increase the accuracy even further. A parameter is a configuration variable that is internal to the method, but its value is estimated from your training data. Generally speaking, parameters are not adjusted by the investigator. Examples of parameters are the number of vectors in a support vector machine, the coefficients in a logistic regression, or the weights in a neural network. Hyperparameters, by contrast, are configured by the investigator and cannot be estimated from the training data [4, 12]. Examples include how the computer searches through the dataset for certain values or features (grid search vs random search) or the learning rate for training a neural network.

Finally, the selected method is evaluated with a held out testing set or cross validation as discussed above. The entire workflow is shown in Fig. 14.1. The arrows

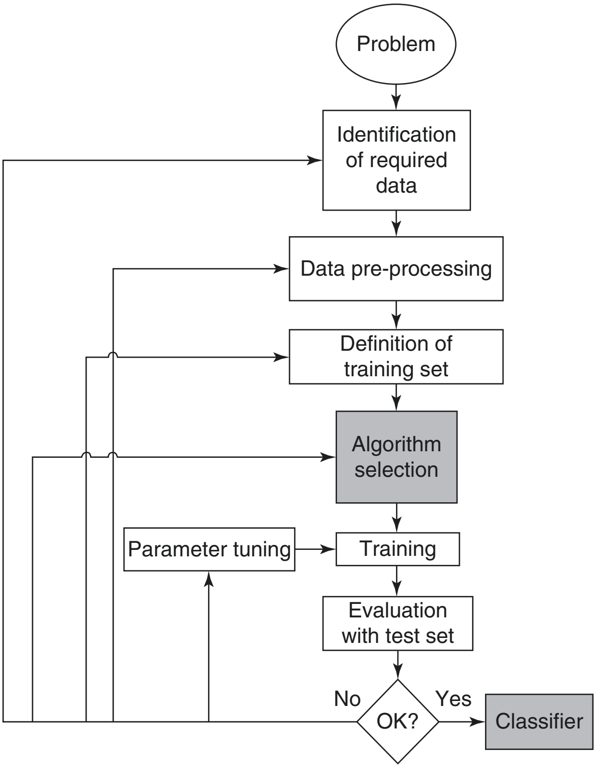


Fig. 14.1 A generalized machine learning workflow. Adapted from Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. Informatica. 2007, 31: 249–268 with permission

from the bottom decision node indicate that the entire process is iterative with continued adjustment or continued testing with different methods depending on the accuracy and acceptability to the end user(s). Combining more than one learning method through ensembles can improve the predictive accuracy (see "Ensembles" below) [10].

14.4 Selected Machine Learning Methods

Here, we will provide a brief review of selected machine learning methods commonly used in health services research. This guide is intended to serve as a high level overview of each method, its underlying statistical mechanism, and the particular advantages or disadvantages of using the method as it relates to clinical medicine.

14.4.1 Tree-Based Methods

Tree-based methods use recursive partitioning to classify cases based on their feature values. The output is a tree where each node represents a question or decision, and the branches lead to terminal nodes that indicate the outcome classification. The investigator can adjust stopping rules or "prune" the tree by specifying the number of layers and branches to make the tree less complex and easier to display. Clinicians will find tree-based methods easy to comprehend because they are similar to clinical pathways and approximate their own clinical reasoning for differential diagnosis or choosing a treatment strategy. Shown in Fig. 14.2 is an example of a tree-based method called CART [13–16].

14.4.2 Bayesian Networks

Bayesian Networks also provide a visual output. Specifically, Bayesian networks specify a graphical model for probability relationships among a set of features. The graphical model is a directed acyclic graph (DAG) where each variable, or node, represents multiple possible states and the probability associated with each state. Arcs or arrows connect the nodes represent conditional dependence relationships and influence the probability of the node to which it is connected. As the name and use of probabilities implies, the underlying statistics are based on Bayes Theorem. While the DAG is not as intuitive as a tree, the underlying use of probabilities is familiar to most clinicians. Furthermore, when using a Bayesian network for predicting an outcome classification, one can also report the probability of the case truly belonging to that outcome [17–19]. Figure 14.3a shows a naïve Bayesian Network in that all features are conditionally related to the same outcome class and not to one another [20]. However, Bayesian networks can become much more complex when there are conditional dependence relationships (arcs) *between* the features *and* the outcome (Fig. 14.3b) [17].

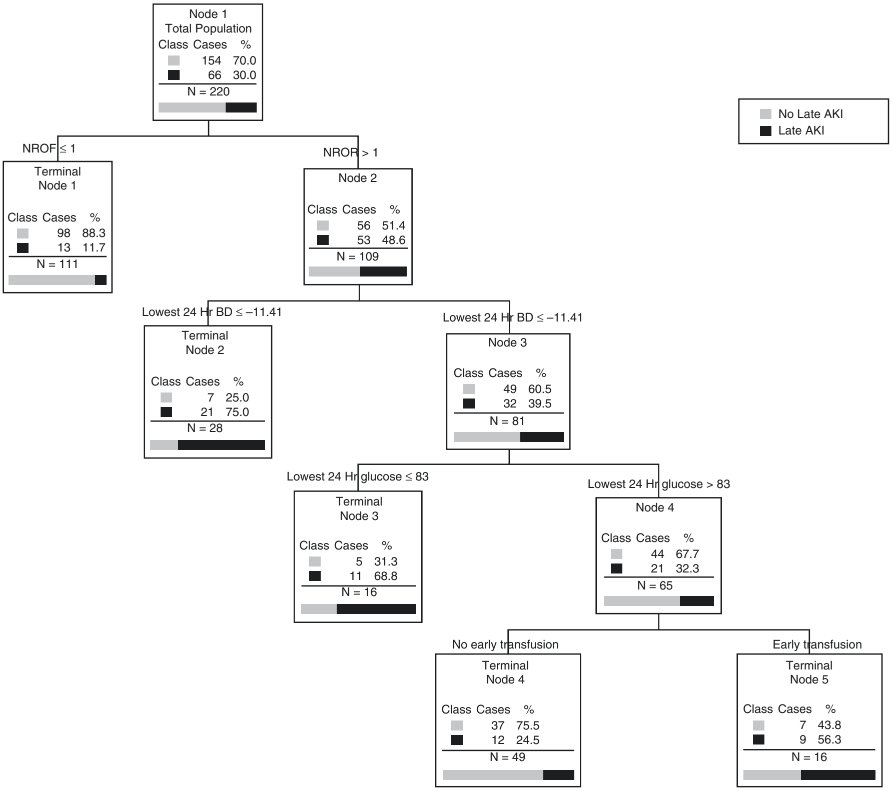


Fig. 14.2 A sample CART Analysis. The classification and regression tree shown predicts late acute kidney injury in burn patients. Figure reprinted with permission Schneider DF, et al. Predicting acute kidney injury in burn patients: A CART analysis. *J Burn Care Res*, 2012; 33:242–251

14.4.3 Rule-Based Methods

Rule-based methods do not provide a visual output, but clinicians will easily relate to their "if, then" structure. Simple logical rules often require the investigator to discretize continuous variables (e.g, "if the calcium is greater than 10 mg/dL, then the patient has hypercalcemia"). Rule quality values resolve conflicts when multiple rules apply to a specific case. Like tree-based methods, the number of rules can be grown or pruned by user-specified constraints [4, 21, 22].

14.4.4 Support Vector Machines (SVMs)

SVMs are more abstract than the methods discussed thus far. Each case is represented in a multi-dimensional vector space with the vectors as numerical representations of the features. The underlying mechanism revolves around fitting a hyperplane or

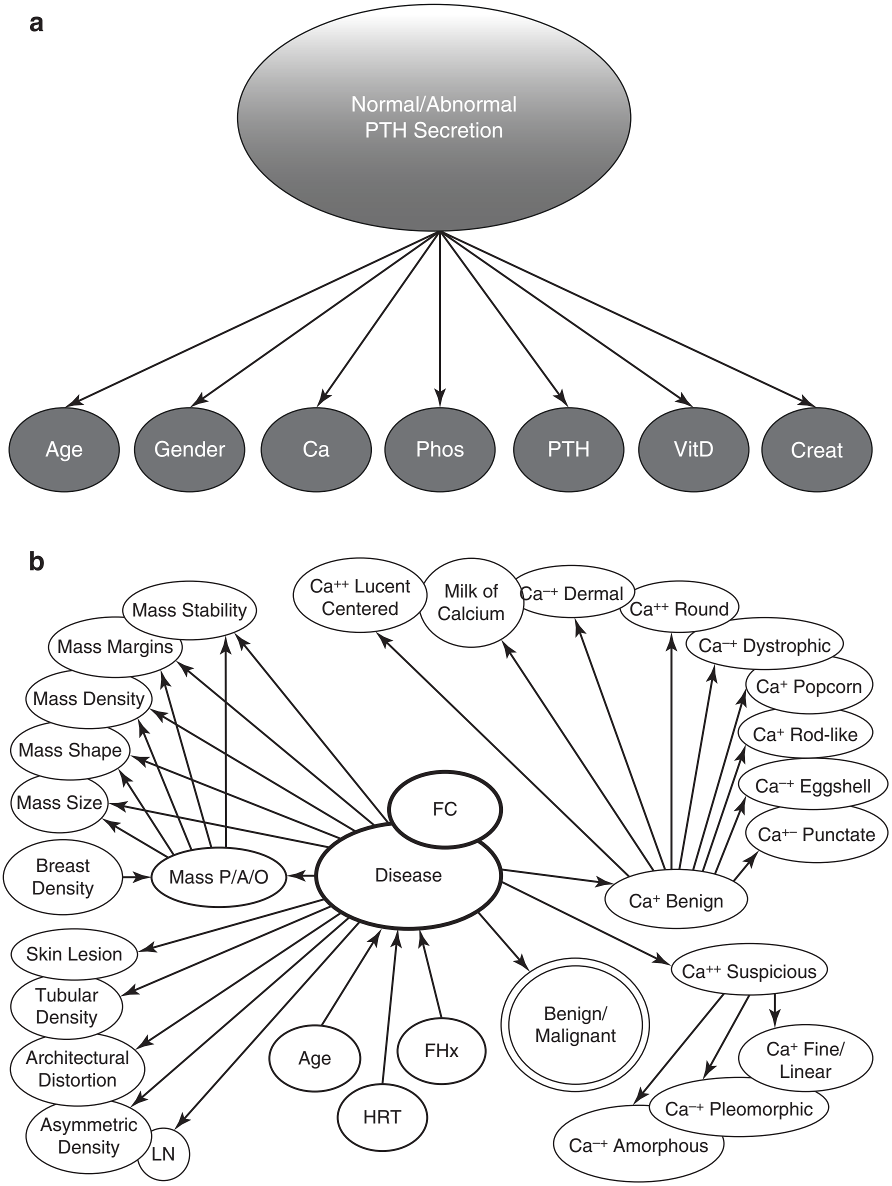


Fig. 14.3 Sample Bayesian Networks. Shown is a naïve Bayesian network to identify patients with primary hyperparathyroidism. All features are conditionally dependent on the outcome and not to one another (**a**). By contrast, a more complex Baysian network to predict breast cancer risk from mammographic microcalcifications contains arcs between the features as well as the outcome of interest (**b**). As seen in Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, Schneider DF. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. Surgery. 2017;161 [4]:1113–21 (with permission) (A) and Burnside ES, et al. Bayesian Network to Predict Breast Cancer Risk of Mammographic Microcalcifications and Reduce Number of Benign Biopsy Results: Initial Experience. *Radiology*; 2006:240:666 (B) (with permission)

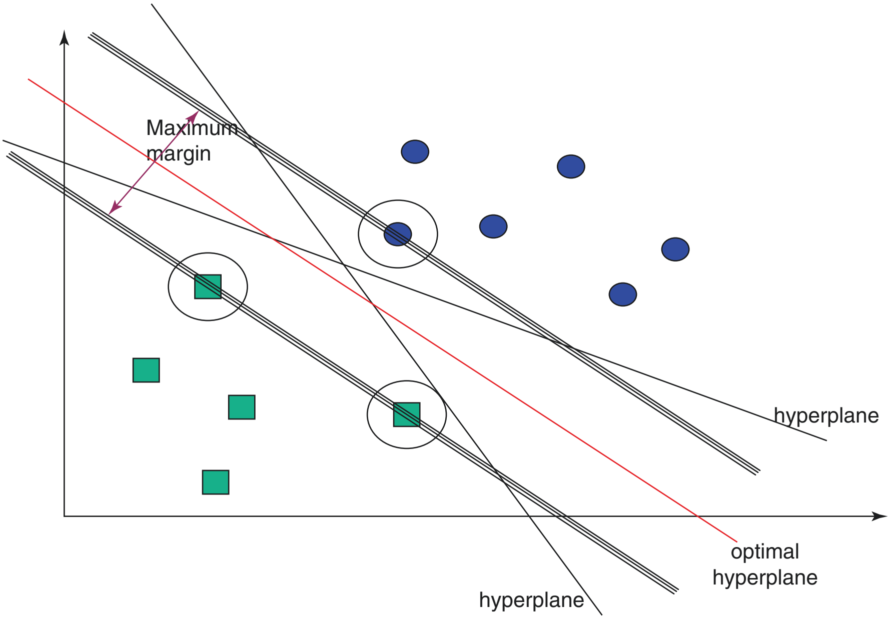


Fig. 14.4 Support Vector Machine. A hypothetical support vector machine and its hyperplane are shown. The blue dots are cases from one outcome class while the green dots are the other outcome class. From Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. Informatica. 2007, 31: 249–268 with permission

boundary between the outcome class(es). The concept of the hyperplane is demonstrated in Fig. 14.4. The algorithm seeks to create the largest possible distance between the hyperplane and the cases on either side of the hyperplane. Once the hyperplane is identified, the cases that lie on its margin(s) are the support vector points, and the line is defined mathematically using only these support vectors (Fig. 14.4). Therefore, the complexity of the SVM model is unaffected by the number of cases or the number of features in the dataset. For these reasons, SVMs are well suited for learning tasks where the number of features is large relative to the number of cases. For other learning methods, one requires many more cases as the number of features increases. A good rule of thumb is 10 times the number of cases as there are features. SVMs, however, are uniquely suited for datasets that are wider than tall – a good point to remember when selecting an algorithm [10, 23, 24]. In clinical medicine, we often encounter datasets with fewer number of cases despite collecting a lot of feature variables – think rare diseases, sparse enrollment, or the early years for a registry.

14.4.5 Neural Networks

Neural networks or multilayered perceptrons are the least transparent method for clinicians to comprehend. As the name implies, the concept loosely mimics the brain in that there are “neurons” segregated into three layers: the input layer, a hidden layer, and the output layer. The layers connect to one another (edges). The neurons and edges have a weight that adjusts the strength of each connection as the learning proceeds. In medicine, neural networks are used for computer vision or image recognition. State-of-the-art "deep learning" uses many hidden layers to improve accuracy for image recognition tasks [25, 26]. We will discuss this topic with two examples in the following section.

14.4.6 Ensembles

An "ensemble" uses more than one machine learning method to improve the predictive performance beyond what any constituent method could achieve by itself. Sometimes, one of methods performs a specific task, such as feature selection, and is then combined with a second method for classification. Certain methods truly function as an ensemble. For example, a random forest is really a multitude of decision trees [27]. The random forest combines randomly drawn decision trees with **bagging**, or bootstrap aggregating. Each tree then "votes" for the final classification. By contrast, **boosting** also combines models of the same type, but each new model emphasizes or weights the misclassified cases from the previous model such that it learns from the previous model's mistakes [4, 20, 27].

14.5 Deep Learning: Specific Cases

14.5.1 Computer Vision

Broadly, computer vision refers to the use of artificial intelligence to obtain information from images. In medicine, most applications of computer vision use "deep learning" to process medical photos or radiology images [28]. For example, a convolutional neural network could distinguish malignant vs benign appearing skin lesions with accuracy exceeding that of board-certified dermatologists [29]. Using a training set of over 100,000 retinal images and two separate testing sets, a team that included scientists from Google detected referable diabetic retinopathy with very high sensitivity or specificity. The majority decision of seven board-certified opthalmologists determined the ground truth for these images [25].

Both examples above used a convolutional neural network. In the case of the diabetic retinopathy study, the neural network was originally built for ImageNet, a contest where researchers trained a neural network to recognize images in 1000 different categories for everyday objects such as cats, cars, and different foods. The convolutional neural network uses a specific function to combine nearby pixels from each image and aggregates them into local features. Local features, in turn, are aggregated into higher level global features. As we discussed in the "General Workflow" section, the data were first preprocessed. Here, this refers to normalizing the images, and resizing the diameter of the fundus so that all images were standardized. The parameters, or mathematical components underpinning the neural network are initially set to random values. For each image, the label (in this case, a severity grade) given by the network is compared to that of the expert opthalmologists. The network adjusts its parameters to decrease the error on that particular image. Through a process called back-propagation, the network changes every parameter through the network proportional to how quickly it reduces the discrepancy between the predicted and actual outcome label(s). The process is repeated for each image in the training set such that the network "learns" to grade the severity of diabetic retinopathy from the set of pixels (25).

The power of this type of neural network is that a single network was trained to make multiple predictions including: whether the image processed represented moderate retinopathy, severe retinopathy, referable macular edema, and whether the image was fully gradable. Another powerful tool illustrated by this example is "transfer learning." The neural net was initially trained for the ImageNet competition – its parameters were learned on ImageNet pictures and then used to bootstrap a similar network for a different task, diabetic retinopathy. Hence, data from one domain can apply to a completely different domain. Transfer learning makes progress even when abundant training data is lacking since you can initialize a neural network on a separate, more robust dataset [30, 31].

14.5.2 Natural Language Processing

Standard natural language processing toolkits such as MedLEE and cTAKEs exist, and these software systems function by breaking the text document into sentences and phrases (parsing) and then matching the phrases to a dictionary of known medical terms such as SNOMED [32–35]. Such software works well for entity recognition (e.g, finding progress notes that discuss "smoking cessation" or identifying CT scan reports that mention "pulmonary emobolus"). However, deep learning is capable of more complex tasks such as identifying patient language indicative of alcohol relapse or classifying patient instructions by their readability score.

Deep learning for more complex natural language processing tasks first converts the document into a set of vectors [26]. Semantic vector space models represent each word, phrase, or document with a geometric vector or word embedding and have been used for tasks such as translation, question answering by digital assistants, and named entity recognition. Traditionally used methods for identifying word embeddings include the Word2Vec method which uses a neural net's internal representation of the word. That is, Word2Vec converts a document into a set of feature vectors for words in that document. While Word2Vec is just a two-layer neural net (not deep learning), it converts a document into a numerical form that deep nets can process for the prediction task [36, 37]. A well-trained set of word vectors will place words with similar meaning close to one another in the vector space. Furthermore, words and concepts that are similar to each other correspond to vectors that point in similar directions; words with analogous meanings are encoded sensibly (the difference vector between "man" and "woman" is the same as that between the vectors representing "boy" and "girl" and to that representing the difference between "king" and "queen").

One promising use case of deep learning is to analyze lay patient language in order to identify disease or health-related quality of life issues not commonly discussed in the clinical setting. For example, a preliminary study using a mobile app for patients with alcohol addiction trained an algorithm to recognize messages from patients at high risk for relapse [38, 39]. Human experts labeled messages indicative of a high relapse risk. Here, a novel type of transfer learning was utilized such that the neural network could be initialized with a larger corpus of text (e.g, phone conversations) and then adapted to a different domain (text messages in an alcoholism support group app). One could imagine this type of technology used to analyze clinical conversations or social media platforms, alerting providers about patients at risk for depression, substance abuse, or sexual assault [38–40].

14.6 Conclusions

As many of the examples in this chapter illustrate, the tools of artificial intelligence are proving quite accurate for certain tasks. However, do these tasks truly constitute human intelligence? Although these examples may be categorized as artificial intelligence, they are really examples of computers performing a singular, well-defined task. Notice that we did not provide any examples of computers operating without direction, reasoning and working with a patient to decide on an appropriate treatment course, or rejecting certain results for lack of clinical or logical sense. One important limitation of machine learning (and therefore, artificial intelligence) is the data itself. The performance of any algorithm is only as good as the training data. Any biases or inaccuracies in the training data will then occur in the outcome predictions. Capturing, improving, and auditing our datasets will ensure that we can train accurate machine learning models.

If physicians are not in danger of losing their jobs to computers, then we must learn to use the tools of artificial intelligence to help us perform our jobs – either as a second opinion, automating certain tasks, or extending our reach where the proper experts are lacking. The expert physician of the future will be savvy in assessing, interpreting, and communicating the methods and results of artificial intelligence algorithms. Certain tasks will still require a physician, and patients will still want to discuss their condition and potential treatments with an expert human being. The fundamental theorem of informatics is that machine + doctor is greater than either one alone, and this is a good way to approach artificial intelligence in surgery and health services research.

References

1. Waljee AK, Higgins PD. Machine learning in medicine: a primer for physicians. Am J Gastroenterol. 2010;105(6):1224–6.
2. Lyman GH, Balducci L. Overestimation of test effects in clinical judgment. J Cancer Educ. 1993;8(4):297–307.

3. Obermeyer Z, Emanuel EJ. Predicting the future— Big Data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.
4. Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco, CA: Elsevier; 2005.
5. Hindman M. Building better models: prediction, replication, and machine learning in the social sciences. Ann Am Acad Pol Soc Sci. 2015;659(1):48–62.
6. Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform. 2003;36(4–5):334–41.
7. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
8. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, et al. AMIA board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assoc. 2012;19(6):931–8.
9. Friedman CP. What informatics is and isn't. J Am Med Inform Assoc. 2013;20(2):224–6.
10. Kotsiantis SB. Supervised machine learning: a review of classification techniques. Informatica. 2007;31:249–68.
11. Chen JH, Asch SM. Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507–9.
12. Tobler JB, Molla MN, Nuwaysir EF, Green RD, Shavlik JW. Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. Bioinformatics. 2002;18(Suppl 1):S164–71.
13. Chipman HA, George EI, McCulloch RE. Bayesian CART model search. J Am Stat Assoc. 1998;93:935–60.
14. Biggs D, de Ville B, Suen B. A method of choosing multiway partitions for classification and decision trees. J Appl Stat. 1991;18:49–62.
15. Schneider DF, Dobrowolsky A, Shakir IA, Sinacore JM, Mosier MJ, Gamelli RL. Predicting acute kidney injury among burn patients in the 21st century: a classification and regression tree analysis. J Burn Care Res. 2012;33(2):242–51.
16. Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, et al., editors. Recursive deep models for semantic compositionality over a sentiment tree-bank. 2013 Conference on Empirical Methods in Natural Language Processing; 2013
17. Burnside ES, Rubin DL, Fine JP, Shachter RD, Sisney GA, Leung WK. Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience. Radiology. 2006;240(3):666–73.
18. Kahn CE Jr, Roberts LM, Wang K, Jenks D, Haddawy P. Preliminary investigation of a Bayesian network for mammographic diagnosis of breast cancer. Proc Annu Symp Comput Appl Med Care. 1995:208–12.
19. Liu YI, Kamaya A, Desser TS, Rubin DLA. Bayesian classifier for differentiating benign versus malignant thyroid nodules using sonographic features. AMIA Annu Symp Proc. 2008:419–23.
20. Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, et al. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. Surgery. 2017;161(4):1113–21.
21. D'Avolio LW, Nguyen TM, Goryachev S, Fiore LD. Automated concept-level information extraction to reduce the need for custom software and rules development. J Am Med Inform Assoc. 2011;18(5):607–13.
22. Imbus JR, Randle RW, Pitt SC, Sippel RS, Schneider DF. Machine learning to identify multigland disease in primary hyperparathyroidism. J Surg Res. 2017;219:173–9.
23. Dhawan A, Wenzel B, George S, Gussak I, Bojovic B, Panescu D. Detection of acute myocardial infarction from serial ECG using multilayer support vector machine. Conf Proc IEEE Eng Med Biol Soc. 2012;2012:2704–7.
24. Kim S. Weighted K-means support vector machine for cancer prediction. Springerplus. 2016;5(1):1162.

25. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10.
26. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
27. Ahn H, Moon H, Fazzari MJ, Lim N, Kodell RL. Classification by ensembles from random partitions of high-dimensional data. Comput Stat Data Anal. 2007;51:6166–79.
28. Shiraishi J, Li Q, Appelbaum D, Doi K. Computer-aided diagnosis and artificial intelligence in clinical imaging. Semin Nucl Med. 2011;41(6):449–62.
29. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
30. Do CB, Ng AY, editors. Transfer learning for text classification. Neural information processing systems (NIPS). Vancouver, British Columbia, Canada: Neural Information Processing Systems Foundation; 2005.
31. Raina R, Ng AY, Koller D editors, Constructing Informative Priors using Transfer Learning. 23rd International Conference on Machine Learning; 2006; Pittsburgh, PA.
32. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med. 2005;29(5):434–9.
33. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology. 2002;224(1):157–63.
34. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18(5):544–51.
35. Zhou L, Tao Y, Cimino JJ, Chen ES, Liu H, Lussier YA, et al. Terminology model discovery using natural language processing and visualization techniques. J Biomed Inform. 2006;39(6):626–36.
36. Mikolov T, Chen K, Corrado G, Dean D, editors. Efficient estimation of word representations in vector space. International Conference on Learning Representations Workshop; 2013; Scottsdale, AZ, 2013.
37. Pennington J, Socher R, Manning CD, editors. GloVe: Global vectors for word representation. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014); 2014; Doha, Qatar.
38. Sarma P, Liang Y, Sethares W. Domain adapted word embeddings for improved sentiment classification. The 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018); July 15–20, 2018; Melbourne, Australia 2018.
39. Sarma P, Sethares W. Simple algorithms for sentiment analysis on sentiment rich, data poor domains. 27th International Conference on Computational Linguistics (COLING 2018); August 2018; Santa Fe, New Mexico 2018.
40. Kim Y, editor Convolutional neural networks for sentence classification. Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 October 25–29, 2014; Doha, Qatar: Association for Computational Linguistics.

Part III

Tools of the Trade

Large Databases Used for Outcomes Research

15

Kyle H. Sheetz, Terry Shih, and Justin B. Dimick

Abstract

Health services researchers often focus on population-based assessment of the effectiveness of health care interventions, evaluation of broad-based delivery system reforms, or variation in use of services across regions. These analyses not only require large sample size, but also diverse practice settings. There are two main sources of secondary datasets, administrative databases and clinical registries, each with their advantages and disadvantages. Administrative databases, such as the national Medicare Provider Analysis and Review (MEDPAR), are primarily compiled for billing purposes. Other sources of data such as large clinical registries such as the Society of Thoracic Surgeons (STS) National Database and the American College of Surgeons National Safety Quality Improvement Program (ACS NSQIP) were created to facilitate quality improvement. The purpose of this chapter is to provide an overview of secondary databases.

15.1 Introduction

Health services researchers often focus on population-based assessment of the effectiveness of health care interventions, evaluation of broad-based delivery system reforms, or variation in use of services across regions. These analyses not only

K. H. Sheetz (✉) · J. B. Dimick
Department of Surgery, University of Michigan Health System, Ann Arbor, MI, USA
e-mail: ksheetz@med.umich.edu

T. Shih
Department of Surgery, University of Michigan Health System, Ann Arbor, MI, USA

Department of Surgery, Oregon Health Science University, Portland, OR, USA

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_15

require large sample size, but also diverse practice settings. Because randomized controlled trials are not technically feasible for all these settings due to cost and sample size considerations, we often turn to large existing databases. These databases have the advantage of a large sample size to answer clinical questions about infrequently performed procedures (e.g. Whipple or esophagectomy) or to answer large population-based questions.

There are two main sources of secondary datasets, administrative databases and clinical registries, each with their advantages and disadvantages. Administrative databases, such as the national Medicare Provider Analysis and Review (MEDPAR), are primarily compiled for billing purposes. Other sources of data such as large clinical registries such as the Society of Thoracic Surgeons (STS) National Database and the American College of Surgeons National Safety Quality Improvement Program (ACS NSQIP) were created to facilitate quality improvement. Both administrative databases and clinical registries have inherent strengths and limitations when used for research that will be discussed further in this chapter.

The purpose of this chapter is to provide an overview of secondary databases. We will begin by discussing the types of research questions for which large databases are frequently used with examples of landmark studies. We will then differentiate between administrative databases and clinical registries with a discussion of the advantages and disadvantages as well as situations in which one may be better than the other. Finally, we will give a brief overview of frequently used databases.

15.2 Common Research Questions that Require Large Database Analysis

15.2.1 Studying Rare Diagnoses, Procedures, or Complications

Health services researchers can leverage the advantages of large secondary databases to answer many clinical questions (Table 15.1). One natural advantage is the large sample size of these databases, allowing for meaningful conclusions to be drawn for rare diagnoses, procedures, or complications. For example, Sheffield et al. used statewide Medicare data in Texas to explore the rate of common bile duct injury during cholecystectomy [1]. As the rate of this complication is quite rare (0.3–0.5%), single institution studies of this complication are often inadequately powered to draw any significant inferences. Previous studies performed with both administrative and clinical registry data have illustrated a significant controversy regarding the role of intraoperative cholangiography in the prevention of common bile duct injury during cholecystectomy. Using an instrumental variable analysis (a technique described in Chap. 16) to adjust for unmeasured confounding, Sheffield et al. demonstrated no statistically significant association between intraoperative cholangiography and common duct injury. This study was also able to link patient-level data to both hospital and surgeon-level characteristics, allowing for exploration of how factors at multiple levels influence patient outcomes.

Table 15.1 Common research questions that utilize large databases

Type of question	Example	Description
Examination of rare conditions, procedures, complications	Sheffield et al. Association between cholecystectomy with vs without intraoperative cholangiography and risk of common duct injury. JAMA 2013	Used Texas Medicare data to explore a rare complication to demonstrate no statistically significant association between intraoperative cholangiography and common duct injury
Trends over time in utilization or outcomes	Santry et al. Trends in bariatric surgical procedures. *JAMA* 2005	Used the Nationwide inpatient sample to demonstrate a national increasing trend in the use of bariatric surgical procedures
Regional variations in utilization	Dartmouth atlas of health care	Used national Medicare data to publish numerous studies exploring large regional variation in healthcare spending and outcomes
"Real world" efficacy vs. effectiveness	Wennberg et al.: Variation in carotid endarterectomy mortality in the Medicare population. *JAMA* 1998	Used national Medicare data to demonstrate carotid endarectomy mortality was significantly higher than reported in randomized controlled trials
Relationship of hospital or surgeon factors on patient outcomes	Birkmeyer et al. Hospital volume and surgical mortality in the United States. *NEJM* 2002	Used national Medicare data to demonstrate inverse relationship between hospital volume and surgical mortality among 14 major cardiovascular and cancer operations
Relationship between mortality and complications	Ghaferi et al.: Variation in hospital mortality associated with inpatient surgery. *NEJM* 2009	Used NSQIP database to demonstrate low mortality hospitals in general and vascular surgery had similar complication rates, but superior ability to rescue patients from complications when compared to high mortality hospitals
Policy evaluation	Dimick et al. Bariatric surgery complications before vs. after implementation of a national policy restricting coverage to centers of excellence. *JAMA* 2013	Used the state inpatient database to demonstrate no benefit in complication rates or reoperation rates after enactment of policy to restrict bariatric surgery to designated centers of excellence

15.2.2 Defining Temporal Trends or Regional Differences in Utilization

Secondary databases are also often used to explore the utilization rates of specific surgeries to discern temporal trends or regional differences. In 2005, Santry et al. used the Nationwide Inpatient Sample (NIS) to discern a significant increase trend in the use of bariatric surgical procedures over the years 1998–2002 [2]. Regional variation in rates of procedures has been described as early as 1970. The Dartmouth Atlas of Health Care has used national Medicare data to publish numerous studies exploring large regional variations in healthcare spending and outcomes [3].

15.2.3 Examining Surgical Outcomes in the "Real World": Efficacy vs. Effectiveness

Another natural application for the use of secondary databases is in the design of "real world" comparative effectiveness studies as opposed to efficacy studies. Efficacy studies are randomized controlled trials, generally regarded as the gold standard to practicing evidence-based medicine. These clinical trials are often performed in a narrow patient population with strict inclusion and exclusion criteria. Efficacy trials tend to include only the most experienced surgeons and, therefore, may not capture how treatment effects vary across a heterogenous group of providers. They are also performed under ideal clinical conditions with close follow-up and ample hospital resources. These stringent criteria and ideal conditions can reduce type 1 error (incorrectly rejecting the null hypothesis) in randomized controlled trials. However, they threaten the external validity of the results. The large survival benefit of an intervention may not hold up in the real world, where patients and practice conditions are a lot messier. Because secondary databases tend to be population-based, these databases allow for assessment of outcomes in the real world. Wennberg et al. demonstrated this efficacy vs. effectiveness distinction examining carotid endarterectomy mortality using national Medicare data [4]. They report the mortality among Medicare beneficiaries undergoing carotid endarterectomy, a large heterogeneous cohort, was appreciably higher at low volume hospitals (2.5%) compared to the higher volume hospitals (1.4%) participating in two large well-designed randomized controlled trials: the North American Symptomatic Carotid Endarterectomy Trial (NASCET) and the Asymptomatic Carotid Atherosclerosis Study (ACAS).

15.2.4 Studying Outcomes Across Diverse Practice Settings

Health services researchers also use secondary databases to explore variation in outcomes across diverse practices settings. Studies can be designed to explore the effect of hospital volume or patient demographics (age, gender, race, socioeconomic status) on outcomes. Birkmeyer et al. used national Medicare data to definitively establish the inverse relationship between hospital volume and surgical mortality among six major cardiovascular procedures and eight major cancer resections [5]. Most case reports on complex operations like pancreatic resection are from high-volume institutions. Lower-volume centers would not be represented in the literature without large database studies. This study has since inspired a multitude of other studies exploring the volume-outcome relationship in many other procedures.

Studies using secondary datasets have also examined surgeon and/or hospital characteristics and their impact on patient outcomes. Ghaferi et al. used the NSQIP database to explore the variation in hospital mortality associated with inpatient general and vascular surgery [6]. To study variations in mortality and other patient outcomes, a broad sample of hospitals is required. Without secondary data, it would not be feasible to compare outcomes across hospitals and draw meaningful inferences. Using a large clinical registry, Ghaferi et al. were able to compare low-mortality and high-mortality hospitals and discovered similar complication rates. The distinction in low-mortality hospitals was their superior ability to prevent mortality in patients experiencing complications.

15.2.5 Evaluating Health Care Policy

As discussed above, there are many instances in which performing a randomized controlled trial for an intervention would be impractical, and using a large secondary database is the only option to assess outcomes. One example of this is in the evaluation of the impact of large-scale policy changes such as Medicare's decision to restrict coverage of bariatric surgery to hospitals designated as centers of excellence. With such a large-scale policy decision, it would be infeasible to conduct a randomized controlled trial to evaluate this policy. Dimick et al. used the State Inpatient Database (SID) from 12 states to compare center of excellence hospitals and non-center of excellence hospitals using a difference-in-differences analytic approach incorporating existing time trends to demonstrate no benefit in rates of complications or reoperation after the enactment of this policy [7]. The difference-in-differences analytic approach is discussed in detail in Chap. 4.

15.3 Administrative Databases vs. Clinical Registries

15.3.1 Strengths and Weaknesses of Administrative Data

Administrative databases offer several advantages in health services research (Table 15.2). These data sets are population based, allowing for the examination of time trends or regional differences as well as real world outcomes, and have a large sample size, allowing for the study of rare conditions, as discussed above. These datasets are also relatively inexpensive and readily available. Therefore, administrative data are a good source of preliminary data for a grant. Another benefit of administrative data is the ability to link patients across episodes of care and institutions. For example, this allows for more accurate tracking of readmissions than clinical

Table 15.2 Strengths and weaknesses of administrative databases and clinical registries

Administrative databases	
Strength	**Weakness**
Population-based	Lack clinical detail
Readily available	Inaccurate/variable coding
Linkable across episodes	Lags in availability
Large sample size	Collected for billing purposes
Clinical registries	
Strength	**Weakness**
Clinical granularity	Lags in availability
Large sample size	Relies on abstractors for reliability
	Resource-intensive

data, which often relies on hospitals to keep track of their own readmissions. This may be easy enough if a patient is readmitted to the same surgeon or service, but more complicated if a patient is readmitted to a different provider or service within the same hospital, or a completely different hospital altogether.

The primary weaknesses in administrative data lie in the inaccuracy and variability in coding. Administrative databases were developed primarily for billing purposes. As a result, primary procedure and diagnosis codes, demographics, length-of-stay and outcomes such as mortality and readmission are recorded with good accuracy. However, administrative data often lack clinical granularity beyond these data points. Patient comorbidities, often used to adequately risk-adjust patients, rely on the varying quality of secondary diagnoses coding. Furthermore, the coding of complications, an important outcome measure in surgery, has also been criticized for clinical inaccuracy. The surgical health services researcher must have effective strategies to address these weaknesses as discussed below.

15.3.2 Identifying Comorbidities in Administrative Data

Popular methods to perform comorbidity risk adjustment include the Charlson Index and the Elixhauser method [8, 9]. The Charlson Index assigns a specific point value for certain comorbidities to predict 10-year survival. It has been adapted for use with both the International Classification of Disease, ninth and tenth Revisions, Clinical Modification (ICD-9-CM, ICD-10-CM) codes by various groups, with the most popular proposed by Deyo et al. [10]. Studies have shown that the incidence of coded comorbidities are similar between ICD-9-CM and ICD-10-CM. Thus, it is reasonable to perform risk-adjustment of outcomes from claims data that include both coding modifications. The Elixhauser method uses a list of 30 comorbidities that can be identified by ICD-9/10-CM codes. These comorbidities are not simplified by an index, as each comorbidity can affect outcomes differently among different patient groups. Rather, each comorbidity is used as an indicator variable in logistic regression when performing risk adjustment. Both methods for risk adjustment are widely used and have been previously validated for use with administrative data. However, limitations still exist in using these methods. Though both the Elixhauser method and Charlson Index can discern the presence of comorbidities, the severity of comorbidities is not discernable. Also, in situations where confounding by indication based on clinical severity are important, administrative data may not adequately identify patients with higher clinical severity. Administrative data will not yield the appropriate clinical detail to perform rigorous comparative effectiveness studies without a more advanced method for causal inference.

15.3.3 Identifying Complications in Administrative Data

The Complications Screening Program developed by Iezzoni et al. is commonly used to identify complications from administrative data [11]. Using specific

ICD-9/10-CM codes, the following postoperative complications can be identified: pulmonary failure, pneumonia, myocardial infarction, deep venous thrombosis/pulmonary embolism, acute renal failure, hemorrhage, surgical site infection, and gastrointestinal hemorrhage. This method has also been previously validated by chart review and shown to be in good agreement. Again, however, we caution the reader in using these methods in studies for which other complications not targeted in the Complications Screening Program (e.g. anastomotic leak, urinary tract infection, central line associated bloodstream infection) may be important outcome measures. Furthermore, the severity of these complications cannot be ascertained with administrative data.

15.3.4 Clinical Registries

In contrast, clinical registries are developed for expressed research purposes or quality improvement. As a result, they contain more clinical granularity than administrative data. Severity of disease, intricacy of procedure, and complexity of post-operative care are all examples of clinical detail that can be obtained in clinical registries not found in administrative databases. For example, through the Society of Thoracic Surgeons Adult Cardiac Surgery Database a researcher can find information on preoperative ejection fraction, cardiopulmonary bypass time, and postoperative transfusions: all data that would not be possible through claims data.

There are also disadvantages to clinical registries though. Participation in clinical registries often require full-time data abstractors at each site that review the medical record and enter data into a large data warehouse. Though research has suggested participation in clinical registries may lead to quality improvement with a return on investment, hospital administrators may not be able or willing to support the resource-intensive costs required for participation. Additionally, clinical registry data are never perfect. Some outcomes, such as long-term out-of-hospital survival or readmissions may not be adequately captured in clinical registries, depending on how these outcomes are reported and verified. Health services researchers must also realize that although clinical registry data improve greatly on the granularity of clinical detail, there are still limitations in the data that can be collected. Observational comparative effectiveness studies performed with clinical registries may still be biased by confounding factors that remain unobserved.

15.4 Example Datasets

15.4.1 Administrative Data

Large administrative databases are produced and maintained by many sources. Several of the most commonly used databases are listed below (Table 15.3).

Table 15.3 Examples of administrative databases

Database	Description
Centers for medicare and medicaid services	
Medicare provider analysis and review (MEDPAR)	Medicare part A claims for inpatient hospitals and skilled nursing facilities
Part B claims data	Medicare part B claims for physician fees and hospital outpatient care
Surveillance, epidemiology and end results (SEER)-Medicare linked data	Clinical and demographic data from population-based cancer registries linked to Medicare claims data
Veterans affairs database	
Patient treatment file (PTF)	VA inpatient hospital discharge claims
National Patient Care Database (NPCD)	VA outpatient visits at VA-based clinics
Healthcare cost and utilization project	
Nationwide inpatient sample (NIS)	20% stratified sample of hospital discharge data from all-payers
Kids' inpatient database (KID)	Sample of all pediatric inpatient discharges from 44 states
State inpatient database (SID)	State-specific inpatient discharges from 45 states. Individual state databases available for purchase
State ambulatory surgery database (SASD)	State-specific data from 30 states for same-day procedures
Marketscan	
Commercial claims and encounter database (CCAE)	Claims submitted to >100 health plans that contract with large private employers, public agencies, and public organizations in the United States
Provider-level data for linkage	
American Medical Association (AMA) physician Masterfile	Provider-level database containing information on education, training, and professional certification
American Hospital Association (AHA) annual survey database	Hospital-level database containing information on hospital demographics, organizational structure, facilities and services, utilization, expenses and staffing

15.4.1.1 Medicare

Perhaps the most widely used database for surgical outcomes research is the Medicare Provider Analysis and Review (MEDPAR) file [12]. It contains Medicare Part A claims for services provided to fee-for-service beneficiaries admitted to Medicare certified inpatient hospitals and skilled nursing facilities. The Center for Medicare and Medicaid Services also maintains research files for Medicare Part B claims submitted by physicians or hospitals for outpatient care. The Chronic Condition Warehouse (CCW) is a 5% sample of Medicare patients that provides claims across the care continuum and can be used to answer questions that require a variety of claims files.

15.4.1.2 SEER-Medicare

Medicare claims data can be augmented by linkage to clinical registries. Commonly, the Medicare data can be linked to the Social Security Death Index to assess long-term survival. Also, Medicare data has commonly been linked with the Surveillance, Epidemiology and End Results (SEER) registry to obtain clinical and demographic data from population-based cancer registries from 18 SEER regions, representing approximately 28% of the US population [13]. The linked SEER-Medicare database can be used to examine variation in cancer-directed surgery and long-term outcomes after cancer surgery.

15.4.1.3 Veterans Affairs Hospitals

The federal government also gathers data for patients receiving care by the Department of Veterans Affairs. The Patient Treatment File (PTF) contains hospital discharge abstracts for inpatient care while the National Patient Care Database (NPCD) contains outpatient visits at VA-based clinics [14].

15.4.1.4 Healthcare Cost and Utilization Project

The Healthcare Cost and Utilization Project (HCUP) is a family of useful healthcare databases developed through a federal-state-industry partnership sponsored by the Agency for Healthcare and Research Quality (AHRQ) [15]. It is the largest collection of longitudinal hospital care data with all-payer, encounter-level information beginning in 1988. The Nationwide Inpatient Sample (NIS) is a large, national database containing hospital discharge data for all payers, though does not contain 100% of all discharges. It is a 20% stratified sample of discharges from all US non-federal hospitals that participate in HCUP. In 2012, HCUP changed the sampling methodology to capture a 20% sample of representative discharges rather than drawing cases from a sample of hospitals. According to HCUP, this provides more representative sample of discharges and reduces sampling errors that came from the prior hospital-level methodology. State-specific databases are available through the State Inpatient Databases (SID). They contain all inpatient discharge abstracts in 46 states. The resultant databases represent approximately 97% of annual discharges. A number of states make SID files available for purchase through HCUP. These databases are maintained in a uniform format, allowing for easy comparison of data between different states to examine geographic variation in utilization, access, charges, and outcomes. The State Ambulatory Surgery Databases (SASD) capture data from 30 participating states for same day procedures. The Kids' Inpatient Database (KID), similar to the NIS, is an all-payer inpatient care database for children ≤20 years of age from 44 states in the US. Unlike the NIS, the KID does not involve sampling of hospitals. Instead, the KID is a sample of pediatric patients from all hospitals in the sampling frame. For the sampling, pediatric discharges in all participating states are stratified by uncomplicated in-hospital birth, complicated in-hospital birth, and all other pediatric cases.

15.4.1.5 AMA Masterfile and AHA Annual Survey

Provider-level and hospital-level information can be obtained through the American Medical Association (AMA) and American Hospital Association (AHA). These can then be linked to existing HCUP and Medicare data to add provider-level variables in data sets. The AMA physician masterfile contains information about education, training, and professional certification for nearly all physicians in the US [16]. The AHA Annual Survey Database includes data from >6000 hospitals with detailed information regarding hospital demographics, organizational structure, facilities and services, utilization, expenses, and staffing [17].

15.4.1.6 Marketscan

The Marketscan Commercial Claims and Encounter (CCAE) database is compiled from claims submitted to more than 100 health plans that contract with large private employers, public agencies and public organizations in the United States [18]. The database is available for purchase through Truven Health Analytics. Health plan types that are included in this database are employer-sponsored, private, fee-for-service, and capitated insurance to employees and covered dependents. This longitudinal database tracks all patient-level inpatient and outpatient claims for as long as employees remain with their employers.

15.4.2 Clinical Registries

Clinical registries were designed for clinical research and quality improvement. Many of these arose in response to the increasing use of administrative data which many viewed as suboptimal. With improved data accuracy and clinical granularity, these registries are ideal for questions that require in-depth data of disease severity, comorbidities, detail of operation, and complexity of post-operative care (Table 15.4).

15.4.2.1 STS National Database

The Society of Thoracic Surgeons (STS) established a National Database in 1989 as an initiative for quality improvement and patient safety [19]. The database is divided

Table 15.4 Examples of clinical registries

Database	Description
Society of Thoracic Surgeons (STS) National Database	National database divided into three components: Adult cardiac, general thoracic, congenital heart surgery with >90% participation by cardiothoracic surgery programs in United States
American college of surgeons National Surgical Quality Improvement Program (ACS-NSQIP)	Nationally validated, risk-adjusted, outcomes-based program to measure and improve surgical quality across surgical subspecialties
National Trauma Data Bank (NTDB)	National database for >400 level I–IV trauma centers across the United States maintained by the American College of Surgeons with trauma patient data including injury profiles, injury severity score, and mechanism of injury
National Cancer Data Base (NCDB)	National database for >1500 commission on Cancer-accredited cancer programs in the United States with detailed clinical, pathological and demographic data on approximately 70% of all incident cancer cases

into three components: Adult Cardiac, General Thoracic, and Congenital Heart Surgery with Anesthesiology participation within the Congenital Heart Surgery Database. The Adult Cardiac Database is the largest, with over 1000 participating sites and data for over 4.9 million surgical procedures. Analyses suggest that the STS Database has enrollment from more than 90% of cardiothoracic surgery programs in the United States. The STS National Database has been used to derive Risk Calculators for seven different individual procedures. Recently, the STS has developed a composite star rating for hospital or group quality for isolated coronary artery bypass grafting and isolated aortic valve replacement, and the organization is encouraging its members to allow these star ratings to be available to the public. Requests for data are reviewed by the STS Access & Publications Task Force five times per year through an online data request form.

15.4.2.2 ACS NSQIP

National Surgical Quality Improvement Program (NSQIP) began in the Department of Veterans Affairs and was brought into the private sector by the American College of Surgeons (ACS) in 2004 [20]. It is the first nationally validated, risk-adjusted, outcomes-based program to measure surgical quality across surgical specialties in the private sector. Participation by member hospitals requires a Surgical Clinical Reviewer (SCR) that collects clinical variables including preoperative risk factors, intraoperative variables, and 30-day postoperative mortality and morbidity outcomes for patients undergoing major and minor surgical procedures. The ACS NSQIP database has become a valuable tool for participating institutes for quality improvement and clinical studies. All ACS NSQIP participants may access the database by requesting the Participant Use Data File (PUF) through the ACS NSQIP website. This file contains Health Insurance Portability and Accountability Act (HIPAA) compliant patient-level, aggregate data, and does not identify hospitals, healthcare providers, or patients. The PUF is provided at no additional cost to employees of ACS NSQIP participating hospitals.

Many system-wide and regional collaboratives participate in the ACS NSQIP database. However, some collaboratives maintain their own databases that may be used for research purposes. These include the Michigan Bariatric Surgical Collaborative (MBSC) [21] and the Michigan Surgical Quality Collaborative (MSQC) [22]. Finding regional collaboratives in your area may provide a unique opportunity to collect and analyze detail-rich data that may not exist in other databases.

15.4.2.3 National Trauma Data Bank

Created by the ACS to serve as the principal national repository for trauma center registry data, the National Trauma Data Bank (NTDB) is composed of de-identified HIPAA compliant data from >400 level I-IV trauma centers across the United States. It includes information on patient demographics, vital signs, diagnoses, Injury Severity Scores (ISS), injury profiles, mechanism of injury (based on ICD-9-CM codes), procedures, complications, and in-hospital mortality [23]. To gain access to NTDB data, researchers must submit requests through an online application process.

15.4.2.4 National Cancer Data Base

The National Cancer Data Base (NCDB) is jointly managed by the American College of Surgeons' Commission on Cancer (COC) and the American Cancer Society. Created in 1988, it is a nationwide database for more than 1500 COC-accredited cancer programs in the United States [24]. The NCDB contains detailed clinical, pathological, and demographic data on approximately 70% of all US incident cancer cases. The NCDB Participate Use Data File (PUF) provides HIPAA complaint de-identified patient level data that does not identify hospitals or healthcare providers. Investigators at COC-accredited cancer programs must apply for access through an online application process.

15.4.2.5 Clinical Quality Collaboratives

In recent years, the number of states forming surgical quality collaboratives have increased significantly. States like Illinois, Tennessee, Florida, and Michigan all have smaller local collaboratives that not only maintain data registries, but also interact directly with payers and local providers [25, 26]. The data collection platforms resemble those of larger clinical registries, like NSQIP. However, because of their size, many collect more granular information on patient satisfaction, postoperative opioid prescribing practices, or even intraoperative processes of care. Some collaboratives are also involved in the implementation and evaluation of novel surgeon coaching programs for which larger national data is not currently available [27].

15.5 Conclusion

This chapter has provided an overview of secondary data sources available to health services researchers and the scope of questions these databases can explore. Administrative data and clinical registries are both available with their respective strengths and weaknesses. Administrative data are relatively inexpensive and readily available for use, though they lack clinical granularity. Clinical registries improve greatly on the level of clinical detail available, however, they may be costly and do not exist for all diseases and populations. We have reviewed several available secondary databases, but ultimately the choice of data needs to be tailored to the specific research question and analytical strategies discussed elsewhere in this book need to be employed to ensure that sound conclusions are drawn.

References

1. Sheffield KM, Riall TS, Han Y, Kuo YF, Townsend CM Jr, Goodwin JS. Association between cholecystectomy with vs without intraoperative cholangiography and risk of common duct injury. JAMA. 2013;310(8):812–20.
2. Santry HP, Gillen DL, Lauderdale DS. Trends in bariatric surgical procedures. JAMA. 2005;294(15):1909–17.
3. The Dartmouth Atlas of Health Care. http://www.dartmouthatlas.org.

4. Wennberg DE, Lucas FL, Birkmeyer JD, Bredenberg CE, Fisher ES. Variation in carotid endarterectomy mortality in the Medicare population: trial hospitals, volume, and patient characteristics. JAMA. 1998;279(16):1278–81.
5. Birkmeyer JD, Siewers AE, Finlayson EVA, et al. Hospital volume and surgical mortality in the United States. N Engl J Med. 2002;346(15):1128–37.
6. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. N Engl J Med. 2009;361(14):1368–75.
7. Dimick JB, Nicholas LH, Ryan AM, Thumma JR, Birkmeyer JD. Bariatric surgery complications before vs after implementation of a national policy restricting coverage to centers of excellence. JAMA. 2013;309(8):792–9.
8. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373–83.
9. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. 1998;36(1):8–27.
10. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol. 1992;45(6):613–9.
11. Iezzoni L, Daley J, Heeren T, et al. Identifying complications of care using administrative data. Med Care. 1994;32(7):700–15.
12. Medical Provider Analysis and Review (MEDPAR) File. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/MEDPAR.html.
13. SEER-Medicare Linked Database. http://healthservices.cancer.gov/seermedicare/.
14. VA Utilization Files. http://www.herc.research.va.gov/data/util.asp.
15. Healthcare Cost and Utilization Project (HCUP). http://www.ahrq.gov/research/data/hcup/index.html.
16. AMA Physician Masterfile. http://www.ama-assn.org/ama/pub/about-ama/physician-data-resources/physician-masterfile.page?
17. AHA Data and Directories. http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml.
18. MarketScan Research Databases. http://www.truvenhealth.com/your_healthcare_focus/pharmaceutical_and_medical_device/data_databases_and_online_tools.aspx.
19. STS National Database. http://www.sts.org/national-database.
20. Welcome to ACS NSQIP. http://site.acsnsqip.org.
21. Michigan Bariatric Surgery Collaborative. https://michiganbsc.org/Registry/.
22. Michigan Surgical Quality Collaborative. http://www.msqc.org/index.php.
23. National Trauma Data Bank. http://www.facs.org/trauma/ntdb/index.html.
24. National Cancer Data Base. http://www.facs.org/cancer/ncdb/index.html.
25. Daley BJ, Cecil W, Clarke PC, Cofer JB, Guillamondegui OD. How slow is too slow? Correlation of operative time to complications: an analysis from the Tennessee surgical quality collaborative. J Am Coll Surg. 2015;220(4):550–8.
26. Campbell DA Jr, Englesbe MJ, Kubus JJ, et al. Accelerating the pace of surgical quality improvement: the power of hospital collaboration. Arch Surg. 2010;145(10):985–91.
27. Quinn CM, Bilimoria KY, Chung JW, Ko CY, Cohen ME, Stulberg JJ. Creating individual surgeon performance assessments in a statewide hospital surgical quality improvement collaborative. J Am Coll Surg. 2018;227(3):303–12. e303

16 Methods for Enhancing Causal Inference in Observational Studies

Kyle H. Sheetz and Hari Nathan

Abstract

When making health care decisions, patients, their providers, and health care policymakers need evidence on the benefits and harms of different treatment options. Many questions in medicine are not amenable to randomized controlled trials (RCT). The use of observational data, such as insurance claims data, tumor registry data, quality collaborative databases, and clinical registries, to evaluate the comparative effectiveness of various treatment strategies is an attractive alternative. However, causal inference is more difficult in observational studies than in RCTs because patients are not randomly assigned to treatment groups. The objectives of this chapter are to discuss the challenges with using observational data to assess treatment effectiveness and to review methods for enhancing causal inference. Investigators need to identify potential threats to the validity of their results including selection bias, confounding, and measurement bias. Studies that do not account for such threats to validity can produce biased effect estimates that contribute to inappropriate treatment and policy decisions. In this chapter, we focus on careful study design and encourage intimate knowledge of the observational dataset, especially when using administrative data. Finally, we review statistical methods including multivariate analysis, propensity score analysis, and instrumental variable analysis that may be used to adjust for bias and strengthen causal inference. Several clinical examples are provided throughout to demonstrate threats to validity and the statistical methods used to address them.

Based on prior chapter written by Kristin M. Sheffield Ph.D. and Taylor S. Riall M.D., Ph.D.

K. H. Sheetz (✉) · H. Nathan
Department of Surgery, University of Michigan Medical School, Ann Arbor, MI, USA
e-mail: ksheetz@med.umich.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_16

Abbreviations

FEV1	Forced expiratory volume
INR	International normalized ratio
IV	Instrumental variable
MI	Myocardial infarction
NSQIP	National Surgical Quality Improvement Program
RCT	Randomized controlled trial
SEER	Surveillance epidemiology and end results

16.1 Using Observational Data for Health Services and Comparative Effectiveness Research

When making health care decisions, patients, their providers, and health care policymakers need evidence on the benefits and harms of different treatment options. Randomized controlled trials (RCTs) are considered the most valid methodology for evaluating treatment effects, because they balance not only measured but also unmeasured confounders. However, there are practical barriers that prevent RCTs from being performed to answer many important clinical questions. Due to the cost of trials, lack of equipoise, or logistical challenges of recruiting adequate number of patients, many medications, treatments, tests, surgical procedures, and health care delivery methods have not been and cannot be evaluated in RCTs. High-risk groups, such as older and sicker adults, or minorities, are typically not well represented in RCTs of medical interventions and procedures. Similarly, rare conditions are not easily studied in a RCT.

While RCTs indicate how a treatment performs in a controlled trial setting (known as treatment efficacy), they have limited generalizability to community populations and settings. It is important to evaluate effectiveness—or how a treatment performs when used more generally outside the context of a clinical study, by physicians treating diverse patients (many of whom may not have met trial eligibility criteria) in everyday practice—to assess whether demonstrated efficacy in a randomized trial translates to real-world application. For example, a treatment might lack effectiveness in the community because of broadening of indications, comorbidities that interfere with treatment, poorer treatment adherence, or differences in the age and health status of the treated population.

Increasingly, investigators are using non-randomized studies to evaluate the effectiveness and safety of medical interventions in real-world community practice settings, either where RCTs are not feasible, or to motivate expenditure of resources for RCTs. Insurance claims data (e.g., Medicare), tumor registry data (e.g., Surveillance, Epidemiology, and End Results, or SEER), hospital discharge data (National Inpatient Sample), complications registry data (e.g., National Surgical Quality Improvement Program, or NSQIP), and other observational datasets, which may not be collected for research purposes, can be used to conduct comparative effectiveness research. For example, observational datasets have been used to evaluate the effectiveness of emergent cardiac catheterization in patients with acute myocardial infarction [1], laparoscopic vs. open appendectomy [2], antecolic vs. retrocolic gastrojejunostomy in the prevention of internal hernia during gastric bypass [3], long-term survival after endovascular and open repair [4, 5], and observation vs. active treatment (surgery or radiation) for early stage prostate cancer [6, 7].

16.2 Causal Inference with Observational Data

Causal inference is more difficult in observational studies than in RCTs because patients are not randomly assigned to treatment groups. In everyday practice, treatment decisions are influenced by clinician judgment and preference, patient characteristics, processes of care, access to care, and the presence of multiple concurrent interventions or problems. This process of individualized treatment based on nonrandom clinical judgment about risk/benefit creates systematic differences between treatment groups. If the reasons for receiving a given treatment are associated with predictors of the outcome of interest, then measures of treatment effectiveness may be biased.

Furthermore, in observational studies the information that patients and providers use to select treatment is often inaccessible to the investigator. This is especially true when investigators use administrative data, which are not collected for research purposes. Measured factors (e.g., age or race) that influence treatment decisions can be accounted for by inclusion in multivariate models. However, unmeasured factors (e.g., frailty, functional status, or social determinants of health) that are correlated with both the treatment and the outcome will lead to biased estimates of the treatment effect. In observational studies, it can be very difficult to determine whether differences in outcomes between treatment groups should be attributed to a treatment effect or to unmeasured patient differences.

16.2.1 Threats to Validity

Prior to conducting an observational study, investigators need to identify potential threats to the validity of their results. Below, we review the most common issues for comparative effectiveness research.

16.2.1.1 Selection Bias

Selection bias refers to two processes: the selection of patients into the study sample and the allocation of patients into treatment groups. In the first process, selection bias occurs when the selection of patients into the study produces a sample that is not representative of the population of interest. In the second process, selection bias occurs when non-random factors that influence treatment lead to allocation of patients into treatment groups that are systematically different. Treatment selection can be influenced by patient factors such as demographics, severity of illness, functional status, comorbid conditions, exposure to the health care system, socioeconomic status, and concurrent treatment, and other factors associated with the provider, health care system, and environment. Selection bias is common in observational studies evaluating the effectiveness of surgical procedures because patients who undergo surgery are selected for good underlying health, while patients who are not fit enough for surgery get included in the “no surgery” group.

16.2.1.2 Confounding

Confounding occurs when variables that determine treatment selection are also related to the outcome. There are several sources of selection bias or confounding in observational studies of treatment effectiveness [8]. Confounding by indication or disease severity occurs when the sickest patients are more likely to receive a new drug or type of therapy, such as a new thrombolytic for myocardial infarction or antifibrinolytic agents for severe hemorrhage [8]. As a result, the intervention is associated with increased risk estimates and it appears to *cause* the outcome it is intended to prevent. Another source of confounding is selective prescribing or treatment discontinuation in very sick patients [8]. Patients who are disabled, frail, cognitively impaired, or in otherwise poor health may be less likely to receive treatment (particularly preventive medication) and more likely to have poor health outcomes, which exaggerates the estimated treatment effect. For example, patients with serious postoperative complications after cancer surgery may never be fit enough to receive adjuvant treatment. Conversely, the healthy user/adherer bias occurs because patients who initiate a preventive medication or who adhere to treatment may be more likely to engage in other healthy behaviors and seek out preventive healthcare services [8]. This can exaggerate the estimated treatment effect as well as produce spurious associations between the treatment and other health outcomes.

16.2.1.3 Measurement Bias

Measurement bias involves systematic error in measuring the exposure, outcome, or covariates in a study. Measurement error is a major concern for observational studies, particularly those using administrative data such as Medicare claims. The likelihood of measurement error differs based on the type of intervention, outcome, or covariates being measured. For example, it is fairly straightforward to define and identify surgical procedures in Medicare claims data, and costly procedures tend to be accurately coded in billing data. In contrast, it can be very difficult to identify medication use, define an exposure period, and classify patients as treated or untreated. An outcome such as survival is less likely to have measurement error than outcomes such as postoperative complications or incident disease. Similarly, comorbid conditions or risk factors such as smoking may be more difficult to measure. This is particularly true with claims data because diagnosis codes are subject to considerable error and the use of a particular diagnosis code on a claim depends on the presence of the condition, a provider’s use of the code, and the presence of other, more serious conditions. One way to estimate the validity of the exposure and outcome variables in an observational study is to compare them with a gold standard such as patient self-report or the medical record. It is important to recognize, however, that measurement errors do not necessarily invalidate the results of observational studies. First, it is important to assess whether a *differential* measurement error (i.e., a true bias) exists, such that the magnitude of the measurement error varies between treatment groups. Second, it is important to assess whether such measurement bias is likely to exaggerate estimated treatment effects or bias them towards the null.

16.2.2 Unmeasured Confounding

Even very rich datasets such as medical records lack complete information on factors influencing selection of treatment. Perhaps the best example of the prognostic strength of missing variables is self-rated health. In most cohort studies, self-rated health is the strongest predictor of survival (after age and gender). More importantly, self-rated health remains a strong predictor in studies that include a rich array of medical, physiologic, social, and psychological variables, such as the Cardiovascular Health Survey. This means that there is a factor known by the patient and easily accessible by the physician ("How are you feeling?") that clearly influences prognosis, that would likely influence treatment choice, and which is invisible in almost all comparative research studies using observational data.

Causal inference relies on the assumption that the study has adequately adjusted for unmeasured confounding; however, there is no way to test that the assumption is correct, making causal inference risky in observational studies. Investigators must do their best to identify, measure, and adjust for all potential confounders. Studies that do not use the appropriate methodology to account for observed and unobserved sources of bias and confounding produce biased effect estimates that can contribute to inappropriate treatment and policy decisions. In the next section, we discuss methods of controlling for bias in observational studies, including steps investigators can take during the design and analysis phases to minimize unmeasured confounding.

16.3 Controlling for Bias in Observational Studies

Careful study design and research methods are key for causal inference with observational data. No amount of sophisticated statistical analysis can compensate for poor study design. A helpful exercise to conduct when designing an observational study is to describe the randomized experiment the investigator would like to—but cannot—conduct, and attempt to design an observational study that emulates the experiment. Research investigators also should collaborate with statisticians, methodologists, and clinicians with relevant subject-matter knowledge during the study design and analysis process. These collaborators can provide expert input to identify issues with the research question and approach. They also can help to identify confounding variables—determinants of treatment that are also independent outcome predictors—and other potential sources of bias, and determine the expected strength and direction of the anticipated bias.

16.3.1 Study Design

The ideal way to minimize bias in observational studies of treatment effectiveness is to collect comprehensive patient, treatment, and outcome data suggested by relevant clinicians and methodologists. This is ideal for primary research studies; however, it is not an option for secondary data analysis on existing observational data sets. Investigators who use existing datasets cannot control how patients were identified and selected, and the analysis is limited to available variables and the way they were measured at the time of data collection. Therefore, it is critical for investigators using secondary data to consider the comprehensive list of potential patient, provider, and process of care factors as the investigator considers potential challenges to causal inference in order to evaluate the feasibility of answering the research question with the available data. We review several research practices for secondary data analysis that will help to improve causal inference.

Prior to designing a study and analyzing the data, investigators must familiarize themselves with the dataset they will be using, including how the sample was selected, how the data were collected, what variables are included and how they were defined, and the potential limitations. For example, hospital discharge datasets such as the National Inpatient Sample represent hospital discharges and not individual persons. Patients with multiple hospitalizations will be counted multiple times, and the dataset does not contain unique patient identifiers that allow follow-up after discharge. Additionally, the sampling methodology should be considered in using such a dataset for population-level inferences. Administrative claims data such as Medicare data were not collected for research purposes and do not contain direct clinical information. Rather, clinical information has to be inferred from diagnosis and procedure claims. In addition, diagnosis codes listed on claims were designed for reimbursement rather than surveillance purposes, and conditions may be included based on reimbursement rather than clinical importance.

Investigators need to explicitly define the intervention or treatment of interest. A well-defined causal effect is necessary for meaningful causal inference. For many interventions, there are a number of ways to define or measure exposure, which could lead to very different estimates of effectiveness. Investigators using administrative claims data have to infer receipt of treatment based on claims for services and often have to develop surrogate measures of an intervention. This requires the investigator to make assumptions, and he/she must consider how results may be affected.

Another prerequisite for causal inference is a well-characterized target population. Investigators need to explicitly define the subset of the population in which the effect is being estimated and the population to whom the results may be generalized. The investigator should carefully select the study cohort and construct the treatment comparison groups and carefully define the observation period in which outcomes will be monitored. Cohort selection criteria should be specified to construct a ‘clean’ patient sample. For example, in a study evaluating overuse of cardiac stress testing before elective noncardiac surgery, the cohort was restricted to patients with *no* active cardiac conditions or clinical risk factors [9]. Cardiac stress testing was clearly not indicated in such patients; therefore, investigators could label testing as overuse. When defining an observation period, investigators must consider the length of time that is appropriate for the research question and the study outcome. For example, 2-year survival would be an adequate amount of time to assess the effectiveness of interventions for pancreatic cancer, but not for breast or prostate cancer. Finally, investigators must determine the extent to which the potential confounders identified by the research team are observable, measurable, or proxied by existing variables in the observational dataset.

16.3.2 Statistical Techniques

There are a number of statistical methods aimed at strengthening causal inference in observational studies of the comparative effectiveness of different treatments. Table 16.1 shows the most common statistical methods used to adjust for bias. Below, we briefly discuss the statistical methods with regard to their contributions to causal inference for observational data. A detailed description of these methods is beyond the scope of this chapter.

16.3.2.1 Multivariate Regression

Multivariable regression is the conventional method of data analysis in observational studies. Regression models may take many forms, depending on the distribution of the response variable and structure of the dataset. The most commonly used regression models include linear regression for continuous outcomes (e.g., the effect of age on systolic blood pressure), logistic regression for categorical outcomes (e.g., the effect of intraoperative cholangiography on bile duct injury), Cox proportional hazards models for time-to-event outcomes (e.g., effect of adjuvant chemotherapy on survival), and Poisson regression for count data (e.g., the effect of INR level on ischemic stroke rates).

Regression analysis is used to disentangle the effect of the relationship of interest from the contribution of the covariates that may affect the outcome. Regression can control for differences between treatment groups by providing estimates of the

Table 16.1 Statistical methods to reduce confounding in observational studies

Statistical method	Purpose/use
Multivariable regression	Estimate conditional expectation of dependent variable given one or more independent variables
Propensity score analysis (stratification, matching, inverse probability weighting, regression adjustment)	Reduce imbalance in treatment and control groups based on observed variables
Instrumental variable analysis	Adjust for unobserved confounding

treatment effect when the other covariates are held fixed. However, in order to control for a covariate, it must be measurable in the observational dataset; therefore, multivariable regression analysis is generally unable to control for the effects of unmeasured confounders.

16.3.2.2 Stratification or Restriction Prior to Multivariable Regression

Stratification may be used as a method to adjust for a measurable prognostic factor that differs systematically between treatment groups, that is, a potential confounder. Patients are grouped into strata of the prognostic variable, and the treatment effect is estimated by comparing treated and untreated patients within each stratum. This method yields effect measures for each stratum of the prognostic variable, known as conditional effect measures. They do not indicate the average treatment effect in the entire population. Sometimes investigators estimate the treatment effect in only some of the strata defined by the prognostic factor, a form of stratification known as restriction.

Stratification and restriction create subgroups that are more homogeneous, sometimes enabling the investigator to identify the presence of confounding. For example, a study assessing the short-term outcomes of incidental appendectomy during open cholecystectomy used restriction to evaluate the consistency and plausibility of their results [10]. Table 16.2 shows the unadjusted and adjusted associations between incidental appendectomy and adverse outcomes in the overall cohort and in restricted subgroups. Unadjusted comparisons showed paradoxical reductions in mortality and length of stay associated with incidental appendectomy. Multivariable models adjusting for potential confounders, such as comorbidity and nonelective surgery, showed increased risk of nonfatal complications with incidental appendectomy but no differences in mortality or length of stay. The investigators believed that unmeasured differences between the appendectomy and no appendectomy

Table 16.2 Outcomes of patients undergoing open cholecystectomy with vs. without incidental appendectomy for the overall patient cohort and low-risk subgroups

	In-hospital death or (95% CI)	Complications or (95% CI)	Length of hospital stay adjusted differences (95% CI)
Overall cohort			
Unadjusted	0.37 (0.23, 0.57)	1.07 (0.98, 1.17)	−0.46 (−0.38, −0.54)
Adjusted[a]	0.98 (0.62, 1.56)	1.53 (1.39, 1.68)	0.05 (−0.02, 0.12)
Low-risk subgroups, adjusted outcomes [a]			
Age ≤ 70 and elective surgery	2.65 (1.25, 5.64)	1.49 (1.32, 1.69)	0.12 (0.05, 0.19)
Elective surgery and no comorbidity	2.20 (0.95, 5.10)	1.53 (1.35, 1.73)	0.11 (0.04, 0.18)

Note: *OR* indicates odds ratio, *CI* confidence interval

[a] Analyses adjusted for patients' age, sex, primary diagnosis, comorbidity, and admission category, hospital teaching status and bed size, and year of surgery

groups were more likely to exist in high risk patients, confounding the estimates for the overall sample. After restricting the analysis to subgroups of patients with low surgical risk, incidental appendectomy was consistently associated with a small but definite increase in adverse postoperative outcomes.

There are ways to manage a stratified analysis to ensure that the stratification itself does not bias the results. It is often sufficient to perform the analysis in the stratified fashion first (i.e., specifying multiple equations). One can repeat the analysis in a single model with the stratifying variable as an additional covariate. If estimates of effect are different with these two approaches, it may be worth additional exploration to determine why. This could include the addition of interaction terms or even a review of the raw data itself to ensure that it is consistent across values of the stratifying variable.

16.3.2.3 Propensity Score Analysis

A propensity score is the conditional probability that a patient will be assigned to a treatment group given a set of covariates, for example, the probability that a patient will undergo incidental appendectomy given his age, health status, primary diagnosis, and other factors. Propensity scores are generated using a logistic regression analysis modeling the receipt of treatment, and each patient is assigned a propensity score based on his or her individual characteristics. It is appropriate to use propensity score analysis when a large number of variables influence treatment choice. Propensity score analysis enables the investigator to balance treatment groups according to distributions of measured covariates. An implicit assumption of propensity score analysis is that balancing the *observed* patient characteristics minimizes the potential bias from *unobserved* patient characteristics.

There are four general strategies for balancing patient characteristics using propensity scores: stratifying patients into groups on the basis of the propensity score percentiles; matching patients with similar propensity scores across treatment groups (allowing for some variation, or “calipers”); covariate adjustment using the propensity score in multivariable analyses (the least favored approach); and weighting observations based on propensity score, also known as inverse probability of treatment weighting. To determine whether the propensity score model has adequately balanced the treatment groups, an investigator can compare the distributions of measured covariates between treatment groups in the propensity score matched sample, or within strata of the propensity score, or within the weighted sample [11]. Once satisfied that balance has been achieved, the investigator can then directly estimate the effect of treatment on the outcome in the matched, stratified, or weighted sample. If the investigator is using the covariate adjustment using the propensity score approach, then a regression model relating the outcome to treatment status and the propensity score must be specified. Studies comparing these four propensity score strategies have demonstrated that propensity score matching is the most effective at removing systematic differences in baseline characteristics between treated and untreated patients [11].

Let’s review an example from the literature of a study that used propensity score analysis to compare lung cancer-specific survival between patients who underwent either wedge resection or segmentectomy. This study used SEER registry data to identify 3525 patients with stage IA non-small cell lung cancer [12]. A logistic regression model was used to estimate propensity scores for patients undergoing segmentectomy based on age, sex, race/ethnicity, marital status, and tumor characteristics. Baseline characteristics were balanced across the two treatment groups after adjusting for the estimated propensity scores. The investigators used three propensity score methods to estimate the association between segmentectomy and survival: adjusting for the propensity score in a Cox regression analysis; stratifying by propensity score quintiles and estimating a Cox model within the five strata; and matching based on propensity scores and using a Cox model to compare survival between matched groups. Table 16.3, above, shows the results for each method. Segmentectomy was associated with significant improvement in survival in all models, though propensity score matching resulted in slightly stronger associations.

There are few situations where propensity score analysis is advantageous over multivariable regression. When a study outcome is rare, investigators are limited in the number of covariates that may be included in the regression model (a general rule states that there should be at least 10 outcome events for every covariate). Propensity score analysis allows the investigators to adjust for more covariates than it is possible to include in a conventional multivariate model.

It is also possible to use propensity scores to match hospitals, rather than patients. For example, a study evaluating the effects of Accountable Care Organizations (ACO) on surgical outcomes compared hospital participating in ACOs to a matched cohort of similar hospitals [13]. In this case, hospitals are matched based on various characteristics that model their probability of participating in an ACO. The advantage here is that it allows for more intuitive comparisons between similar hospitals. If ACO hospitals tended to be larger or have greater resources, for example, comparing them to a similar subset of control hospitals makes the findings more robust.

16.3.2.4 Instrumental Variable (IV) Analysis

Instrumental variable (IV) analysis provides a powerful means to eliminate confounding arising from both measured and unmeasured factors in observational studies of treatment effectiveness. The IV method employs a variable called an instrument (hereafter, IV) that is predictive of treatment but has no effect on the outcome, except through its influence on treatment assignment. The most familiar illustration of an IV is random assignment to treatment groups in RCTs. This random assignment is unrelated to patient characteristics and has no direct effect on the

Table 16.3 Results of Cox models comparing lung cancer-specific survival of patients treated with segmentectomy vs. wedge resection, by propensity score method[a]

Propensity score method	Lung cancer survival HR (95% CI)
Adjustment for propensity score as covariate	0.76 (0.61, 0.94)
Stratifying by propensity score quintiles	0.76 (0.61, 0.94)
Matching based on propensity scores	0.72 (0.60, 0.86)

Note: *HR* indicates hazard ratio, *CI* confidence interval

[a]All Cox models adjusted for number of lymph nodes evaluated during surgery

study outcome. A good IV creates an allocation that is similar to randomization. As a result, the comparison groups can be expected to have similar distributions of measured and unmeasured characteristics. In IV analysis, patients are compared based on their *likelihood* of receiving treatment according to the IV, rather than their *actual* receipt of treatment. That is, one would report the difference in outcomes between patients who were likely to receive treatment based on the IV and those who were unlikely to receive treatment based on the IV. This is similar to an intention-to-treat estimator in randomized trials, where one might report the difference in outcomes between assigned treatment arms. The interpretation of an IV analysis is with respect to the "marginal patient"—an individual who would be considered a candidate for either treatment (e.g. when comparing two operations).

Table 16.4 shows some examples of IVs that have been used in studies evaluating outcomes of cancer therapy.

Finally, let's review an example from the literature that used geographic variation in treatment utilization as an IV [14]. This study used Medicare for patient undergoing colectomy procedures between 2008 and 2010. The IV was the regional use of laparoscopic colectomy at the HRR level in the year prior to the patient's operation. The question for this study was whether selection bias was confounding the relative safety benefits of laparoscopic vs. open colectomy.

The authors compared results across several risk-adjustment methods. Table 16.5 shows odds ratios for complications and mortality from the conventional risk-adjustment analysis and from the IV analysis [14]. The results are attenuated in the

Table 16.4 Instrumental variables used in outcomes research studies

Category of instrument	Examples
Availability of a key medical resource	Number of oncologists per capita in the hospital referral region (HRR); distance to nearest radiology facility; travel distance to surgical provider
Geographic variation in intensity of treatment utilization or provider practice patterns	Regional use of a particular operative technique (e.g. robotic surgery)
Economic incentives to provider and/or cost to patient of alternative treatments	Medicare's average physician fees for breast conserving surgery and mastectomy
Secular trends and/or changes in treatment patterns	Year of surgery
Patients' distance to a particular treatment	Differential distance between the facility where the patient is treated and the next closest facility that offers the treatment/resource in question

Table 16.5 Comparison of outcomes following laparoscopic vs. open colectomy using logistic regression and instrumental variable methods

Risk-adjustment method	Logistic regression OR (95% CI)	Instrumental variable OR (95% CI)
Complications, 30d	0.55 (0.53–0.56)	0.82 (0.79–0.85)
Mortality, 30d	0.38 (0.35–0.40)	0.75 (0.70–0.75)

IV analysis, suggesting that the treatment effect is partially attenuated because this model can account for unmeasured differences in patient characteristics that may influence the study outcomes.

16.3.3 Limits of Advanced Statistical Techniques

The application of advanced statistical methods can leave investigators and practitioners with a false sense of security in the results. Propensity score analysis will leave residual confounding if there is imbalance across treatment groups in unobserved factors that influence health. IV analysis can eliminate confounding arising from observed and unobserved factors. However, IV analysis requires the identification of a strong and logically justifiable instrument, which can be very difficult. Without question, the biggest limitation to IV analyses is the identification of suitable instrumental variables. To critically evaluate their results, investigators can compare the magnitude and direction of the predicted treatment "effect" across several different adjustment methods.

16.4 Conclusions

Bias and confounding are major issues in studies that assess treatment effectiveness based on observational data, making causal inference difficult. Investigators must conduct a rigorous assessment of threats to the validity of their findings and estimate the strength and direction of suspected bias. Advanced statistical methods are available to adjust for confounding and improve causal inference. However, investigators must carefully consider the limitations of their data because sometimes confounding cannot be overcome with statistical methods, and some comparative effectiveness questions cannot be answered with currently available observational data. Researchers must carefully map the boundaries of comparative effectiveness research using observational data. There is important information to be learned from observational studies, especially population-based cohorts that include patients/providers who are unlikely to participate in randomized clinical trials.

References

1. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA. 2007;297(3):278–85.
2. Hemmila MR, Birkmeyer NJ, Arbabi S, Osborne NH, Wahl WL, Dimick JB. Introduction to propensity scores a case study on the comparative effectiveness of laparoscopic vs. open appendectomy. Arch Surg. 2010;145(10):939–45.
3. Steele KE, Prokopowicz GP, Magnuson T, Lidor A, Schweitzer M. Laparoscopic antecolic Roux-En-Y gastric bypass with closure of internal defects leads to fewer internal hernias than the retrocolic approach. Surg Endosc Other Interv Tech. 2008;22(9):2056–61.

4. Jackson RS, Chang DC, Freischlag JA. Comparison of long-term survival after open vs endovascular repair of intact abdominal aortic aneurysm among medicare beneficiaries. JAMA. 2012;307(15):1621–8.
5. Lee HG, Clair DG, Ouriel K. Ten-year comparison of all-cause mortality after endovascular or open repair of abdominal aortic aneurysms: a propensity score analysis. World J Surg. 2013;37(3):680–7.
6. Wong YN, Mitra N, Hudes G, Localio R, Schwartz JS, Wan F, et al. Survival associated with treatment vs observation of localized prostate cancer in elderly men. JAMA. 2006;296(22):2683–93.
7. Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. Cancer. 2008;112(11):2456–66.
8. Brookhart MA, Sturmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care. 2010;48(6 Suppl):S114–20.
9. Sheffield KM, McAdams PS, Benarroch-Gampel J, Goodwin JS, Boyd CA, Zhang D, et al. Overuse of preoperative cardiac stress testing in medicare patients undergoing elective noncardiac surgery. Ann Surg. 2013;257(1):73–80.
10. Wen SW, Hernandez R, Naylor CD. Pitfalls in nonrandomized outcomes studies. The case of incidental appendectomy with open cholecystectomy. JAMA. 1995;274(21):1687–91.
11. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424.
12. Smith CB, Swanson SJ, Mhango G, Wisnivesky JP. Survival after segmentectomy and wedge resection in stage I non-small-cell lung cancer. J Thorac Oncol. 2013;8(1):73–8.
13. Nathan H, Thumma JR, Ryan AM, Dimick JB. Early impact of medicare accountable care organizations on inpatient surgical spending. Ann Surg. 2019;269(2):191–6.
14. Sheetz KH, Norton EC, Birkmeyer JD, Dimick JB. Provider experience and the comparative safety of laparoscopic and open colectomy. Health Serv Res. 2017;52(1):56–73.

Landmark Papers to Recommend to Readers

Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. Cancer. 2008;112(11):2456–66.

Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA. 2007;297(3):278–85.

Wen SW, Hernandez R, Naylor CD. Pitfalls in nonrandomized outcomes studies. The case of incidental appendectomy with open cholecystectomy. JAMA. 1995;274(21):1687–91.

17 Systematic Review and Meta-Analysis: A Clinical Exercise

Natalie N. Merchant and Melinda Maggard-Gibbons

Abstract

The growth of new clinical knowledge continues to challenge how surgery is practiced, and several types of literature reviews attempt to consolidate this expansion of information. Systematic reviews and meta-analysis are common methodologies that integrate findings on the same subject collected from different studies. Unlike a systematic review, a meta-analysis arrives at a conclusion backed by quantitative analysis. This chapter provides an overview of the principles, application, and limitations of these methods, which is fundamental to interpreting and critiquing their results.

17.1 Introduction

As the amount of clinical data expands, rigorous summaries of the literature are increasingly critical, as they have the ability to challenge current treatment paradigms. Systematic reviews and meta-analyses are two approaches to analyze and summarize results from various studies on the same topic [1–4]. They combine results from individual studies, either by reviewing the data together in one place or pooling the data, to determine if there is a meaningful difference that exists between

N. N. Merchant
David Geffen School of Medicine, University of California, Los Angeles, CA, USA

M. Maggard-Gibbons (✉)
Department of Surgery, David Geffen School of Medicine at University of California, Los Angeles, CA, USA

Department of Surgery, Olive View-UCLA Medical Center, Sylmar, CA, USA

Department of Surgery, Greater Los Angeles VA Medical Center, Los Angeles, CA, USA
e-mail: mmaggard@mednet.ucla.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_17

two treatments or study arms. The results can resolve controversies within the literature and ultimately shape and change clinical practice. This overview introduces the rationale behind performing meta-analyses, including systematic reviews, and discusses their strengths and limitations.

The first rule of a systematic review or meta-analysis is that it should be a clinical exercise. The selection of a thoughtful and clinically impactful question for a specific patient population is the foundation for a meaningful review. The statistical analysis for either is simply a mechanical process that can be performed on any group of data, regardless of whether or not it is appropriate to do so. As such, the question that guides a review is critically important. The judgment of "what makes sense" in terms of the appropriate studies to include, and what not to include, determines the ultimate value of the conclusions of these reviews. For example, if one had an inquiry about the outcomes of laparoscopic versus open cholecystectomy, there are thousands of RCTs and clinical papers that are readily available and could certainly produce a review. However, it isn't until the clinical inquiry is refined (based on the inclusion and exclusion criteria) that the literature search becomes meaningful and the statistical analysis and interpretation of that data becomes clinically useful and impactful. The refinement of the clinical inquiry and filtering of the literature is the guiding the clinical exercise essential.

17.2 Non-Systematic Reviews Versus Systematic Reviews

Both non-systematic and systematic reviews summarize results of the evidence in the literature, but they typically do not include the quantitative statistical pooling like is done for a meta-analysis. A non-systematic review condenses studies in narrative form and gives generalized conclusions; however, this type of review lacks structure in study selection, collection of data, and interpretation of findings. Non-systematic reviews are a convenience sample of the studies that the authors subjectively deemed important to discuss, which, by definition, limits the conclusions that can be made. Without a structured method to identify and select the appropriate articles, there is little way to reproduce the findings; a feature that is critical for the validity of a review. Since a rigorous set of selection criteria are not employed, the studies selected may not represent all available evidence, and the data, or the interpretation of the data, may unintentionally support the reviewer's bias. Additionally, a non-systematic review does not quantitatively combine or weigh the data, which can produce misleading interpretations [4]. Finally, non-systematic reviews tend to lack transparency in the addressing of heterogeneity in the study populations, interventions, or outcomes. Thus, non-systematic reviews provide more of a snapshot of a sample of studies on a particular topic and don't support robust clinical conclusions [3, 5].

A systematic review follows specific defined steps which include defining a research question, establishing inclusion/exclusion criteria, reproducible data abstraction, and quality assessment of the included studies [3, 5]. Systematic reviews tend to have less bias in study design and selection of studies and have the ability to offer more objective conclusions. Although systematic reviews neither formally combine (i.e., pool) data nor weight the relative contribution of each study based on sample size, they can provide a more objective and clinically significant summary of results on a common topic.

17.3 Meta-Analysis

A meta-analysis follows a rigorous methodology: it defines a clinical inquiry, establishes appropriate inclusion criteria, and quantitatively pools the data. The goal of a meta-analysis is to create an objective conclusion from all relevant and available evidence [3, 5]. Unlike systematic reviews, one of the main strengths of a meta-analysis is the quantification of the results from multiple studies by weighting and pooling the contribution of each study in order to aid in the interpretation of study results and guide clinical conclusions.

Meta-analyses can offer conclusions where data are conflicting or inconclusive. While randomized controlled trials (RCT)s are the gold standard of evidence-based medicine, it is common for the results to vary between the individual trials. Meta-analysis can help synthesize the results for atypical findings, e.g. when they show (1) no effect due to small sample size, (2) varying direction of effect, or (3) conflicting results regarding effect versus no significant effect.

For example, a meta-analysis by the Cochrane Collaboration (2004) compared laparoscopic versus open appendectomy [6]. The investigators identified a number of RCTs; none of which definitively established the benefits. However, the meta-analysis of those RCTs found that the laparoscopic approach was associated with lower wound infection rate, decreased incidence of intra-abdominal abscess, less post-operative pain, and shorter duration of hospital stay.

A meta-analysis helped demonstrate the benefit of radiation for early breast cancer; a conclusion that could not have been determined previously because each RCT had a small sample size. The pooled analysis demonstrated that radiation (after lumpectomy) had three times lower local recurrence rate than surgery alone and no difference in mortality [7]. Meta-analysis can combine the data from studies with small sample sizes when large definitive trials are not available and can determine definite conclusions when controversy between studies exist, which contributes to the strength and generalizability of the findings. A technique that utilizes weighted averages for a combination of large and small trials is discussed in the Methodology section and includes an example if its application.

17.3.1 Methodology and Potential Impact of Meta-Analysis

Meta-analysis follows a standardized approach similar in many ways to that used for primary data collection research [3, 4, 8]. The steps include: (1) define the research question, (2) establish study selection criteria, (3) perform a literature search, (4) abstract study variable and outcomes, and (5) data analysis and presentation of results.

The first step defines a specific research question and a formal protocol detailing objectives and hypotheses. This point cannot be over-emphasized: this first step forms the backbone of the analysis. The research question should address the type of patients, intervention, comparison group, and clinical outcomes of interest [9] as it is the foundation for which the literature, analysis and conclusions are based.

Of note, some institutions can be designated as either an Evidence Based Practice Centers (EPC) by the Agency for Healthcare Research and Quality (AHRQ) or an Evidence-based Synthesis Program (ESP) by the Department of Veterans Affairs (VA). EPCs select an expert panel of researchers to synthesize evidence for the facilitation of evidence-based practices to improve health care organization and delivery for the Medicare and Medicaid populations [10]. Whereas ESPs conduct research in a similar way for the healthcare of Veterans. Each expert panel analyzes the evidence generated from 2 to 3 specific clinical questions over a given period to generate reports for clinicians and policy makers. These reports facilitate improvement of clinical practice and delivery of care in those specific populations.

The second step defines the inclusion and exclusion criteria for the studies relevant to the research question. The criteria to be considered are the type and quality of study design (e.g. randomized controlled trial, observational cohort study), patients (e.g. age, gender, presence of medical conditions), data publication (e.g. abstract only or non-published data), language (e.g. inclusion of non-English studies), and time period [3, 4]. For example, there may be an evolution or change in clinical management over time. The breadth of criteria specificity can impact results, as broad inclusion criteria tend to increase heterogeneity among studies, while narrow inclusion criteria lead to limited subgroup analysis [11].

Meta-analysis can be performed using both observational or RCT data. Ideally, limiting the meta-analysis to only RCT data produces results with a higher level of scientific evidence. Randomized data will be less likely to have significant selection bias or other confounding factors. Pooling non-randomized data has many limitations that must be considered in the final assessment of the results.

If the literature search captures a limited number of RCTs, addition of observational studies may be considered. The inclusion of observational studies must be explicitly justified—why there is reasonable case for low bias between the study and control arms, for example. The amount of potential bias can be evaluated using a variety of tools, such as the ROBINS-I [12]. Assessment of observational studies should determine the level of quality and selection for analysis (based on the patient population in question). If the observational studies support meta-analysis, the significance and clinical impact of the studies' outcomes should be analyzed. If observational studies do not support a meta-analysis, they can be valuable in narrative format and the findings are interpreted to only supplement the RCT results. Depending on the quality and specificity of an observational study, its outcomes may be a rich source of supplementary data.

Third, a literature search is performed to obtain all relevant studies that meet the inclusion criteria, using Medline, the Cochrane Library, Current Contents, and EMBASE databases. Utilizing multiple databases helps ensure that pertinent publications are not omitted; non-English written articles should be included and translated when appropriate [3, 4, 8]. Scanning the bibliographies of the retrieved articles (referred to as reference mining) and asking experts in the field will identify additional publications [3]. Articles with duplicate or previously published data should be excluded.

The fourth step involves abstraction of study features, characteristics of the patient population, and outcomes onto a standard data collection form [8]. The quality of studies should be evaluated with regards to randomization, blinding, and explanation for dropouts and withdrawals, which addresses internal validity (minimization of bias) and external validity (generalizability) [13]. To maintain accuracy, two independent researchers should extract the data, and the degree of agreement between the reviewers should be calculated (e.g. kappa statistic) [8]. A formal process for how discrepancies are resolved must be established and reported. Blinding researchers to the study authors and other identifying information may decrease the chance of bias, but this is not routinely done.

A systematic and transparent way of rating the quality of the evidence is crucial for determining the strength of an article's recommendations and ultimately, the clinical importance. There are several ways to rate the quality of evidence including the GRADE guidelines and Cochrane Collaboration risk of bias tool. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) guidelines require a clear clinical question and specification of outcomes, categorizes four levels of quality (randomized control trials as highest) with various factors increasing or decreasing level of quality, and values simple summaries [14]. The Cochrane Collaboration's method focuses on assessing the bias of a study based on generalizability and outcome of a study, quality of reporting, appraisal of clearly designated the domains analyzed by the authors [15]. The most appropriate method should be determined and used consistently for all articles in the review.

The fifth step of a meta-analysis involves data analysis and presentation of results. Data must be translated to a common scale to allow for comparison. For example, a recent meta-analysis on bariatric surgery found that the majority of articles reported preoperative weight in kilograms, while some articles reported preoperative weight in pounds [16] Transforming data into a common scale allows for maximal inclusion of data and accuracy of analysis.

The type of analysis depends on whether the outcome variable is continuous (e.g., length of hospital stay) or dichotomous (e.g., adverse event occurrence). For continuous endpoints, the mean difference between two groups (e.g., control and treatment groups) is recorded [8].

If the endpoint is dichotomous (effect versus no effect), the odds ratio (OR) or relative risk (RR) is calculated [8]. OR is defined as the ratio of events to non-events, and the odds ratio is defined as the odds in one group (e.g. treatment group) divided by the odds in a second group (e.g., control group). OR greater than one ($OR > 1$) means that the event is more likely in the treatment group and therefore the treatment group is favored if the "event" is desirable (e.g. survival). Odds ratios can be harder to interpret. Risk is defined as the number of patients with an event divided by the total number of patients, and the risk ratio is defined as the risk in one group (e.g., treatment group) divided by the risk in a second group (e.g., control group). RR less than one (RR < 1) favors the treatment group if the "event" is not desirable (e.g., reoperation). RR tends to be an easier concept to understand when considering clinical outcomes—it is the risk of an event between two groups. Of note, odds ratios do not translate directly to relative risks, especially as the effect size increases. If an odds ratio is assumed incorrectly to be a relative risk, effect size can be overestimated.

In general, odds ratios are reported for retrospective (case–control) studies and RR are reported for prospective cohort studies. Review Manager (RevMan) is a type of software that allows researchers to enter the data for the included studies and generates pooled data results and graphic presentation diagrams [17].

When reporting a meta-analysis, the combined study effect (i.e., difference between the study arms) is presented graphically along with results of individual studies, as demonstrated by the forest plot example in Fig. 17.1 [18] In this example, the first column lists the six studies included in the pooled analysis on wound protectors at reducing surgical site infections. The number of infections in each study is displayed—for each "arm"—with and without the wound protector. The vertical line represents the point (i.e., OR of 1) where there is no difference in event

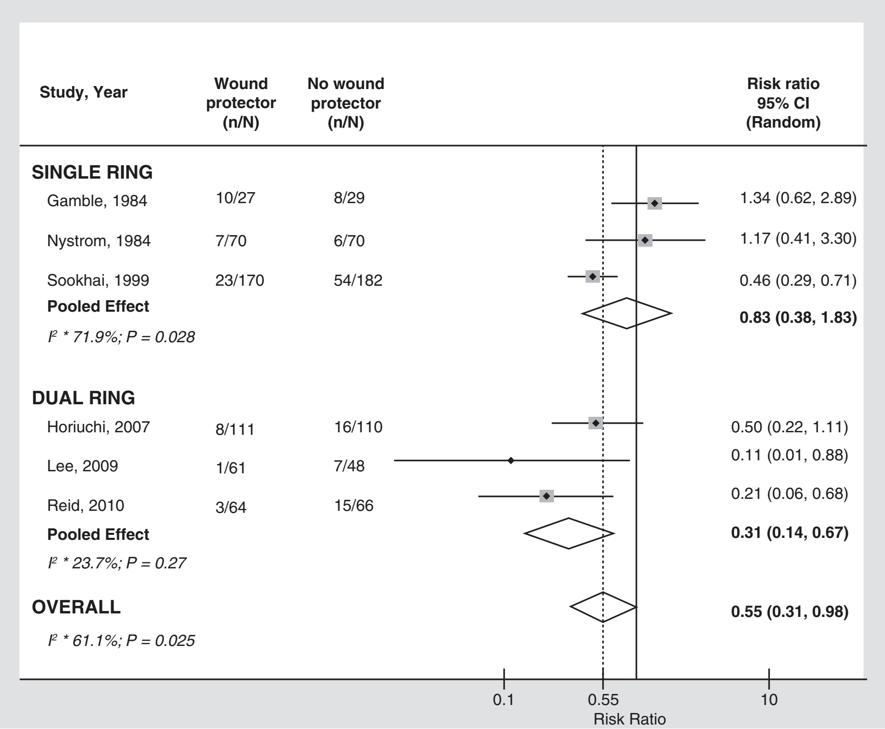


Fig. 17.1 Forest Plot depicting pooled random effects meta-analysis and subgroup estimates according to dual versus single ring structure of wound protector. Edwards JP, Ho AL, Tee MC, Dixon E, Ball CG. Wound protectors reduce surgical site infection: a meta-analysis of randomized controlled trials. Ann Surg. 2012 Jul;256(1):53–9

rate. The OR for each study is represented with a square while the 95% confidence interval (CI) is depicted as a horizontal line. If the CI includes 1, then there is no statistically significant difference between the procedures. Under the weight column, a percentage is provided quantifying each study's contribution, and the corresponding OR and 95% CI are reported. The diamond-shaped symbol represents the pooled analysis results where the midpoint corresponds to the pooled estimate and the horizontal spread the pooled 95% CI. If the horizontal spread does not meet the vertical line in this type of forest plot, it can be concluded that there is a statistically significant difference between the treatments. However, if the horizontal spread of the diamond crosses the vertical OR line, it can be determined that there would be no statistical difference. The forest plot also provides information about heterogeneity through the plots of individual study effect and the chi-square test. The test for overall effect determines the statistical significance of the data included in the meta-analysis by generating the z-value along with the p-value [2]. If the CIs of the studies do not overlap, it suggests substantial variation that is not accounted for by chance. If the chi-square test's p-value is less than 0.10 or the chi-square value is greater than the degrees of freedom, then study heterogeneity is likely present [2].

For meta-analyses that include RCTs, certain absolute measures, such as the risk difference (RD), also called the absolute risk reduction, (ARR) and number of patients needed to treat (NNT) can be calculated. The RD is defined as the risk in the treatment group minus the risk in the control group, which quantifies the absolute change in risk due to the treatment. In general, a negative risk difference favors the treatment group. The NNT is defined as the inverse of the risk difference and is the number of patients that need to be treated with the intervention to prevent one event. In the case where the risk difference is positive (does not favor the treatment group), the inverse will provide the number needed to harm (NNH) [2].

For example, a meta-analysis comparing RCTs of stent treatments for infragenicular vessels in chronic lower limb ischemia found that primary patency was significantly higher with the drug eluting stent compared to patients treated with bare metal stent (OR 4.51, 95% CI 2.90–7.02, and NNT 3.5) [19]. The drug eluting stent increased the odds of vessel patency 4.5-fold higher than the metal stent and 3.5 patients had to receive a drug eluting stent to prevent loss of patency in one patient in the control arm.

Another technique in meta-analysis utilizes a weighted average of the results. Larger trials are given more weight since the results of smaller trials are more likely to be affected by chance [4, 8]. Either the fixed-effects or random-effects model can be used for determining the overall effect. The fixed-effects model assumes that all studies are estimating the same common treatment effect; therefore, if each study were infinitely large an identical treatment effect could be calculated. The random-effects model assumes that each study is estimating a different treatment effect and hence yields wider confidence intervals (CI). A meta-analysis should be analyzed using both models. If there is not a difference between the models, then the studies are unlikely to have significant statistical heterogeneity. If there is a considerable difference between the two models, then the most conservative estimate should be reported, which is usually the random effects model [2].

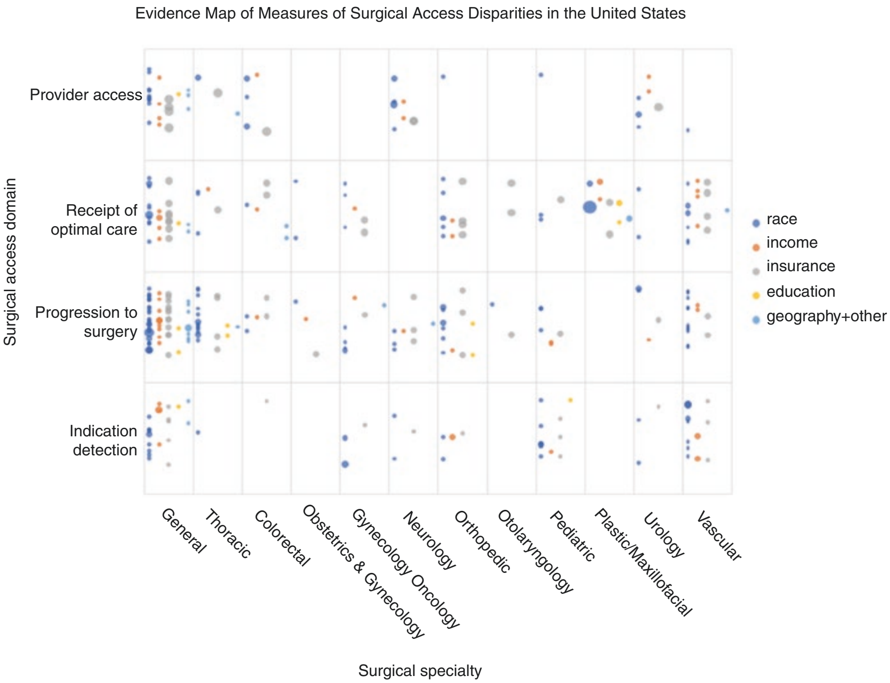


Fig. 17.2 Illustration of evidence map: Improvement in length of stay (relative to baseline), by intervention type and resource use. Evidence map of measures of surgical access disparities in the United States. Bubble size indicates number of studies supporting each measure. Plotting of the bubbles in each cell is systematic to increase readability of the figure. Horizontally, the bubbles are in five rows based on the disparity domain (color). A random placement generator was used to distribute bubbles vertically inside each cell. de Jager E, Levine AA, Udyavar NR, Burstin HR, Bhulani N, Hoyt DB, Ko CY, Weissman JS, Britt LD, Haider AH, Maggard-Gibbons MA. Disparities in Surgical Access: A Systematic Literature Review, Conceptual Model, and Evidence Map. J Am Coll Surg. 2019 Mar; 228(3):276–298. doi: https://doi.org/10.1016/j.jamcollsurg.2018.12.028. Review. PubMed PMID: 30803548; PubMed Central PMCID: PMC6391739

For example, an evidence map can help give perspective to a highly variable evidence base and illustrate a weighted average of results as demonstrated in Fig. 17.2 [19]. This type of mapping can also reveal gaps in the literature and can help guide research planning so that future studies will be more useful for decision-makers.

There are additional types of meta-analysis. One approach is to run the analysis based on individual patient data. While this method requires a greater amount of resources, it lessens the degree of publication and selection bias, thus potentially resulting in more accurate results. Another example is the cumulative meta-analysis, which involves repeating the meta-analysis as new study findings become available and allows for the accrual of data over time [5]. It can also retrospectively pinpoint the time when a treatment effect achieved statistical significance.

If an outdated systematic review is discovered during a literature search, it can be used as a stepping-stone for an updated systematic review. The researcher should modify and refine the search terms from the original systematic review to generate an explicit and relevant pool of data for the updated systematic review. Next, a new search is completed, confirming that the studies from the original systematic review are accounted for, and new literature is identified. Lastly, the entire data set (e.g., total set of studies) is analyzed. In general, one should not depend on the original systematic review's data analysis or use the older systematic review's conclusions. The new review is a distinct entity from the older review in order to minimize bias, maintain the integrity of the data and generates the most up-to-date conclusions.

While in general a meta-analysis produces an overall conclusion with more power than looking at the individual studies, results must be interpreted with consideration of the study question, selection criteria, method of data collection, and statistical analysis [4].

A note about the Cochrane Library, the Cochrane group utilizes a systematic method of producing relevant and informed systematic reviews for the promotion of evidence-based decision-making. This group has published handbooks detailing their methodology for the production of high-quality evidence, which includes the maintenance of their independence from commercial funding, intolerance of biases, and commitment to accessibility. The Cochrane group's review articles are available to anyone with access to the internet so that the global community has a reliable source of information for clinical decision-making. Additionally, the Cochrane group's methodical and reliable strategy for generating reviews are updated more frequently than paper-based journals [20].

17.3.2 Potential Limitations of Meta-Analysis

The main limitation of a meta-analysis lies in the potential for multiple types of bias. Pooling data from multiple studies unavoidably includes biases of the individual studies [21–24]. Moreover, despite the establishment of study selection criteria, authors may tend to incorporate studies that support their view, leading to selection bias. There is also potential for bias when identifying studies because they can be selected by investigators within the field and unintentionally used their individual opinions to determine inclusion criteria. Language bias may exist when literature searches fail to include foreign studies, because those with significant results are more likely to be published in English. Studies with significant findings tend to be cited and published more frequently, and those with negative or non-significant findings are less likely to be published, resulting in citation bias and publication bias. Since studies with significant results are more likely to be indexed in the literature database, database bias is another concern. Studies which have not been published in traditional journals, like a dissertation or a chapter, are referred to as "fugitive" literature and are less likely to be identified through the traditional database search. Finally, multiple publication bias can occur if several publications are generated from a multi-center trial or a large trial reporting on a variety of outcomes. If the same set of patients is included twice in the meta-analysis, the treatment effect can be overestimated. These potent bias factors can affect the conclusions and must be considered during interpretation of the results.

To combat these sources of bias, several tools are available. A sensitivity analysis can help assess for bias by exploring the robustness of the findings under different assumptions. Exclusion of studies based on specified criteria (e.g. low quality, small sample size, or studies stopped early due to an interim analysis) should not significantly change the overall effect if the results of the meta-analysis are not significantly influenced by these studies. The degree of study heterogeneity is another major limitation and the random-effects model should be used when appropriate [2]. Another approach to measure potential bias is the funnel plot, which is a scatter plot illustrating each study's effect with reference to sample size. The underlying principle is that as the sample size of individual studies increases, the precision of the overall estimate or effect difference improves. This is shown graphically as smaller studies would distribute widely while the spread of large studies should be narrow. The plot should show a symmetrical inverted funnel if there is minimal or no bias, as demonstrated in Fig. 17.3. By the same logic, the plot would be asymmetrical and skewed when bias exists.

One standardized method of assessing and reporting the potential for bias is the Cochrane Risk of Bias [25]. RCTs included in a meta-analysis are evaluated for seven potential biases: random sequence generation, allocation assignment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting and other bias. Each item is scored on a scale of high, indeterminate, or low bias. Reporting this potential for bias helps the reader assess the overall level of bias for the selected studies of interest (Fig. 17.4).

Other areas of criticism involve the interpretation of meta-analysis results. One potential problem can occur when a meta-analyst neglects to consider important covariates, which could lead to misinterpretation of the results. For example, in a

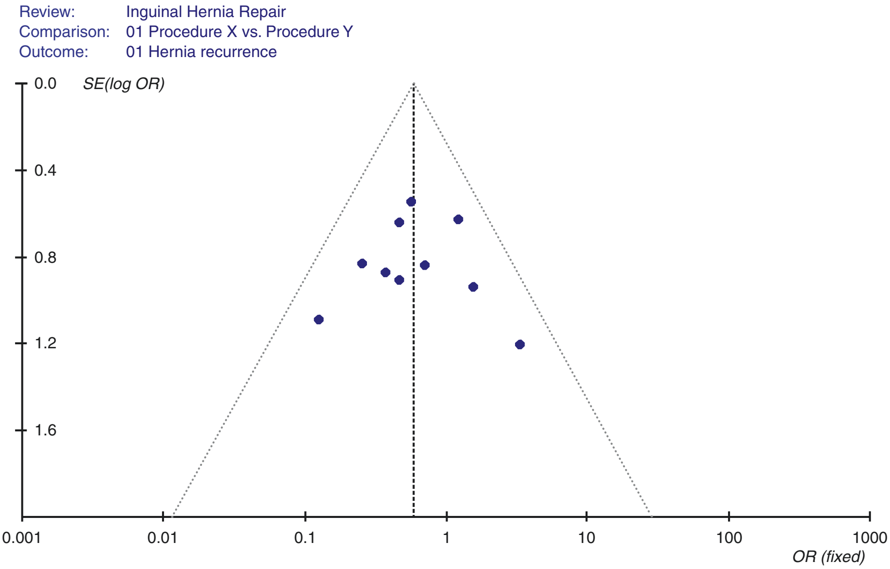
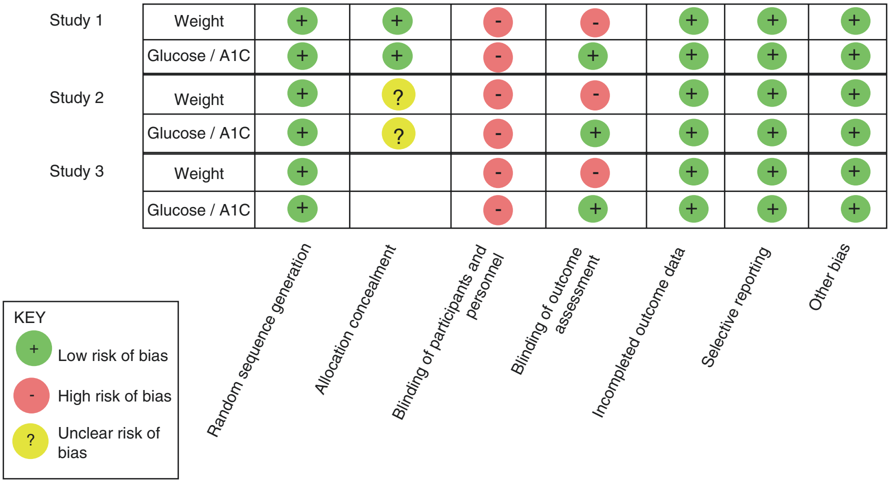


Fig. 17.3 Illustration of a funnel plot for a hypothetical meta-analysis comparing hernia recurrence incidence following Procedure X versus Procedure Y. The y-axis reflects individual study weights as the log of the effect estimate, *SE (log OR)*. The x-axis represents the odds ratio (*OR*) for each study. The symmetry of the plot distribution suggests absence of publication bias



		Random sequence generation	Allocation concealment	Blinding of participants and personnel	Blinding of outcome assessment	Incompleted outcome data	Selective reporting	Other bias
Study 1	Weight	+	+	-	-	+	+	+
	Glucose / A1C	+	+	-	+	+	+	+
Study 2	Weight	+	?	-	-	+	+	+
	Glucose / A1C	+	?	-	+	+	+	+
Study 3	Weight	+		-	-	+	+	+
	Glucose / A1C	+		-	+	+	+	+

KEY
+ Low risk of bias
- High risk of bias
? Unclear risk of bias

Fig. 17.4 Cochrance Risk of bias for hypothetical baritric surgery randomized controlled trials. Two trials are depicted with ratings of high, indeterminate, or low for each of the seven bias categories

study involving cerebrospinal fluid drainage (CSFD) in thoracic and thoracoabdominal aortic surgical repair, the expertise of the surgical team varies among the included studies and could play a critical factor in the outcomes of interest—prevention of paraplegia [26] Some argue that the inherent degree of study heterogeneity does not permit the pooling of data to produce a valid conclusion.

Also, the strength and precision of a meta-analysis is in question when the results contradict a large, well-performed RCT. As such, results of any individual study or trial may be overlooked in place of the pooled results. However, it is arguable that findings falling outside the group mean are likely a product of chance and may not reflect the true effect difference, which provides the rationale for formally pooling similar studies. Even if a real difference exists in an individual trial, the results of the group will likely be the best overall estimate (also known as Stein's Paradox) [27].

Lastly, caution should be exercised when employing subgroup analysis to make decisions on individual patients. Meta-analysis approximates the overall effect of a treatment across a wide range of subjects and thus subgroup analysis is susceptible to bias as well. Figure 17.1 [18] shows two subgroup analyses of single versus double ring wound protector, suggesting better effect of the double ring protector at reducing surgical site infections. However, a cumulative meta-analysis seen in Fig. 17.5 [18] suggests that other changes overtime may also play a role. Clinicians should consider the risks and co-morbidities of the studied population in comparison with their own patients to help decide whether the findings are clinically applicable.

The recent review article of meta-analyses within general surgery by Dixon found many inadequacies in the quality of these studies [28]. Overall, the majority of the meta-analyses had major methodological flaws—median score of 3.3 on a scale from 1 to 7. Areas of weakness included errors in validity assessment,

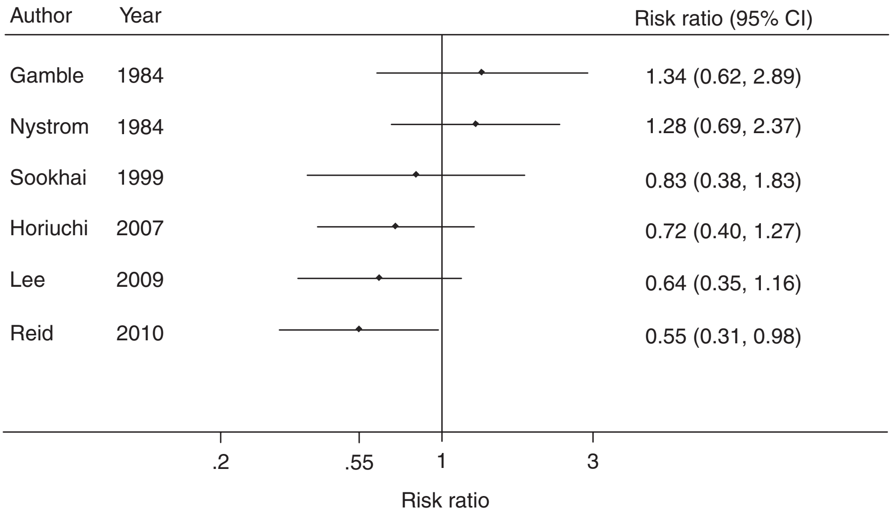


Fig. 17.5 Cumulative meta-analysis detailing the evolution of the pooled effect estimate with addition of subsequent available trial data. Edwards JP, Ho AL, Tee MC, Dixon E, Ball CG. Wound protectors reduce surgical site infection: a meta-analysis of randomized controlled trials. Ann Surg. 2012 Jul;256(1):53–9

Fig. 17.6 Pearls for performing meta-analyses

Pearls for performing meta-analyses

- First the ask a good clinical question
- Statistics are mechanical but must be based in clinical knowledge
- Ideally constructed of only RCTs with little heterogeneity
- Evidence tables will allow the reader to judge the appropriateness of combining the studies
- Combining outcomes with different length of follow-up must be justified
- Risk of bias for RCTs should be assessed and reported

selection bias of patient populations, poor reporting of search strategies, and improper pooling of data. They found that meta-analyses of poorer quality tended to report a greater effect difference than the higher quality ones. As previously discussed, these results emphasize the importance of performing meta-analysis using rigorous and high-quality methodology. Specific general suggestions to follow when conducting meta-analyses are outlined in Fig. 17.6.

17.4 Conclusions

Like primary research, systematic reviews and meta-analyses involve a step-wise approach to arrive at justifiable and clinically relevant conclusions. Identifying the appropriate clinical question provides the foundation for both methods. Meta-analysis is the superior method and providing an accurate appraisal of the literature in addition to quantitative summation. Meta-analysis overcomes the subjective problem of non-systematic reviews and provides a more transparent and reproducible appraisal of the data. It is imperative to understand the strengths and weaknesses of these approaches and in order to be able to critically judge the conclusions, since those results have the potential to influence clinical practice.

References

1. Glass GV. Primary, secondary, and meta-analysis of research. Educ Res. 1976;5:3–8.
2. The Cochrane Collaboration Web site. http://www.cochrane.org. Accessed 11 Nov 2013.
3. Kelly GA. Meta-analysis: an introduction. http://www.pitt.edu/~super1/lecture/lec3221/index.htm. Accessed 11 Nov 2013.
4. Wolf FM. Introduction to systematic reviews and meta-analysis. http://depts.washington.edu/k30/Meta-analysis/Meta-analysis%20clinical%20research%200603_files/frame.htm. Accessed 11 Nov 2013.
5. Egger M, Smith GD. Potentials and promise. BMJ. 1997;315:1371–4.
6. Sauerland S, Lefering R, Neugebauer E. Laparoscopic versus open surgery for suspected appendicitis. Cochrane Database Syst Rev. 2004;4:CD001546.
7. Early Breast Cancer Trialists' Collaborative Group. Effects of radiotherapy and surgery in early breast cancer – an overview of the randomized trials. N Engl J Med. 1995;333:1444–55. [Erratum, N Engl J Med 1996;334:1003.]
8. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. BMJ. 1997;315:1533–7.
9. Meade MO. Selecting and appraising studies for a systematic review. Ann Intern Med. 1997;127:531–7.
10. Ahrq.gov. *Evidence-based Practice Centers (EPC) Program Overview | Agency for Healthcare Research & Quality*. 2019.
11. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med. 1997;127:380–7.
12. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;i4919:355.
13. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609–13.
14. Guyatt GH, Oxman AD, Kunz R, et al. What is "quality of evidence" and why is it important to clinicians? BMJ. 2008;336(7651):995–8.
15. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
16. Buchwald H, Avidor Y, Braunwald E, et al. Bariatric surgery: a systematic review and meta-analysis. JAMA. 2004;292(14):1724–37.
17. http://tech.cochrane.org/Revman. Accessed 21 Apr 2014.
18. Edwards JP, Ho AL, Tee MC, Dixon E, Ball CG. Wound protectors reduce surgical site infection: a meta-analysis of randomized controlled trials. Ann Surg. 2012;256(1):53–9.
19. de Jager E, Levine AA, Udyavar NR, Burstin HR, Bhulani N, Hoyt DB, Ko CY, Weissman JS, Britt LD, Haider AH, Maggard-Gibbons MA. Disparities in surgical access: a systematic literature review, conceptual model, and evidence map. J Am Coll Surg. 2019;228(3):276–98. https://doi.org/10.1016/j.jamcollsurg.2018.12.028. Review
20. Jadad AR, Cook DJ, Jones A, et al. Methodology and reports of systematic reviews and meta-analyses. A comparison of Cochrane reviews with articles published in paper-based journals. JAMA. 1998;280(3):278–80. https://doi.org/10.1001/jama.280.3.278.
21. Egger M, Smith GD. Bias in location and selection of studies. BMJ. 1998;316:61–6.
22. LeLorier J, Grégoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. N Engl J Med. 1997;337:536–42.

23. Bailar JC III. The practice of meta-analysis. J Clin Epidemiol. 1995;48:149–57.
24. Bailar JC. The promise and problems of meta-analysis. N Engl J Med. 1997;337:559–61.
25. Higgins JPT, Altman DG, Sterne JAC. Cochrane handbook for systematic reviews of interventions version 5.1.0. In: Higgins JPT, Green S, editors. Oxford: The Cochrane Collaboration; 2011.
26. Cina CS, Abouzahr L, Arena GO, Lagana A, Devereaux PJ, Farrokhyar F. Cerebrospinal fluid drainage to prevent paraplegia during thoracic and thoracoabdominal aortic aneurysm surgery: a systematic review and meta-analysis. J Vasc Surg. 2004;40(1):36–44.
27. Effron B, Morris G. Stein's paradox in statistics. Sci Am. 1977;236:119–227.
28. Dixon E, Hameed M, Sutherland F, Cook DJ, Doig C. Evaluating meta-analyses in the general surgical literature: a critical appraisal. Ann Surg. 2005;241(3):450–9.

18 Medical Decision-Making Research in Surgery

Susan C. Pitt and Clara N. Lee

Abstract

As surgeons, we make complex decisions every day about whether or not to operate on a patient, and which operation to perform. Some of these decisions are challenging because they involve significant tradeoffs between the risks and benefits of the surgery, such as the risk of complications versus curing a cancer. Other decisions are challenging because they involve deeply personal issues for the patient, such as continence or breast appearance. The ability of surgeons and patients to make these decisions has important implications for health outcomes, including quality of life and health care resource utilization. The science of evaluating, facilitating, and intervening on medical decisions is a relatively young field, which has evolved from other disciplines, including psychology, economics, health behavior, and engineering.

Thus, medical decision-making encompasses a broad range of research methods—from investigating the process of patient decision-making, to the development of patient decision-aids and eliciting patient preferences. This chapter describes this growing field and its applications in surgical research. It first delineates measurement of the decision-making process and decision outcomes, and explains how to develop and test decision aids. It further reviews techniques of preference elicitation.

S. C. Pitt
Department of Surgery, University of Wisconsin School of Medicine and Public Health, Madison, WI, USA
e-mail: pitt@surgery.wisc.edu

C. N. Lee (✉)
Division of Health Services Management and Policy, Department of Plastic and Reconstructive Surgery, College of Medicine, College of Public Health, The Ohio State University, Columbus, OH, USA
e-mail: clara.lee@osumc.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_18

18.1 Clinical Versus Preference Sensitive Decisions

The approach to studying medical decisions depends on the nature of the decision—clinical versus preference sensitive—and the perspective of the decision-maker. Many decisions have clear clinical indications based on a reasonable amount of evidence. For example, the indications for sigmoidectomy in a patient with perforated diverticulitis are almost exclusively clinical and not based on patient preferences, because very little variability exists among patients' preferences about dying from a perforated viscus. In other decisions, however, whether or not to operate, or which operation to perform, depends heavily on patient preference, because patients differ in how they feel about the potential outcomes. For example, some women with breast cancer may want to remove their unaffected breast with a contralateral prophylactic mastectomy to avoid future mammograms and biopsies, while other patients may feel that the changes to their body image or the risks of complications are not worth losing their breast. Another example in which patient preference and approach to risk are essential factors, is the difficult choice between active surveillance and surgery for very low-risk thyroid cancer. These latter examples are "preference sensitive" decisions.

18.2 Evaluating Preference Sensitive Decisions

Preference sensitive decisions can be evaluated with process measures or outcome measures. Considerable debate has taken place over which type of measure is more appropriate. The best approach for your study will depend on your specific aims and research question(s). When using these measures (also called instruments), it is important to keep in mind the context and timing (before or after the decision) for which the instrument was designed. Another important aspect of study design to keep in mind is that each instrument is validated as an entire measure, so an individual question may not be valid by itself.

18.2.1 Process Measures

Measures of the decision-making process typically assess how patients and providers communicate with each other during a consultation. Some of these measures are designed to be used while observing the consultation itself. Others are based on patient or provider report of what was said. The OPTION Scale measures the extent to which a provider involves a patient in decision making [1]. The evaluator completes the scale while observing an actual visit, or while viewing or listening to a video or audio recording of a visit. The main advantage of using some type of recording is that multiple evaluators can independently apply the scale to the same encounter, which improves the reliability of the results.

Other process measures rely on patient or provider report of what took place, rather than direct observation. Patient or provider-reported measures are easier to implement

Table 18.1 Examples of decision process and outcome measures

Measure	Use	Construct measured
Decision process		
OPTION [1]	Observation	Patient involvement
Informed decision-making scale [3]	Observation	Informed participation
Rochester participatory decision-making scale [4]	Observation	Participation
Measure of patient-centered communication [5]	Observation	Patient-centered communication
Medical interaction process system (MIPS) [6]	Observation	Patient-centered communication
Control preference scale [2]	Patient or physician reported	Preference for decision control
CollaboRATE [7]	Patient reported	Shared decision-making
Decisional engagement scale (DES-10) [8]	Patient reported	Engagement in decision-making in cancer care
Decision outcome		
Decisional conflict scale [9]	Patient reported	Decisional conflict
Decision regret scale [10]	Patient reported	Decision regret
Satisfaction with decision scale [11]	Patient reported	Decision satisfaction
SURE test [12]	Patient reported	Decisional conflict
Decision-making confidence scale [13]	Patient reported	Confidence in decision

than direct observation, but the results are less reliable due to recall bias, social desirability bias, or the respondent forgetting. The Control Preference Scale (CPS) asks the patient to report how much involvement s/he prefers to have in the decision-making process [2]. A provider version has been developed, which asks the provider his/her opinion of the patient's preference for involvement in decision making.

Study Design Considerations. When designing a study using process measures, one must consider when to measure the process and which aspect of decision making to focus on. For example, the CPS can be given before or after a patient-provider encounter. It can assess a patient's *general* preference for participation in decision making or a patient's *specific* preference related to a particular clinical decision. Table 18.1 lists some of the most commonly used measures that evaluate different aspects of the decisional process, such as the quality of information exchange, participation, and the degree to which communication is patient-centered.

18.2.2 Outcome Measures

Patient-reported outcome measures. Most, but not all, measures of decisional outcomes are patient-reported outcome measures that evaluate a patient state during or after a decision. For example, the Decisional Conflict Scale is a commonly used outcome measure that consists of 16 questions about perceived uncertainty, factors affecting uncertainty, and perceived effectiveness in decision making. The Decision

Regret Scale is a frequently-used five-item scale that measures regret after a decision, while the Satisfaction with Decisions Scale is a six-item scale that measures patient satisfaction with the decision itself, as opposed to satisfaction with care or with outcomes of care. Table 18.1 lists some additional decision outcome measures.

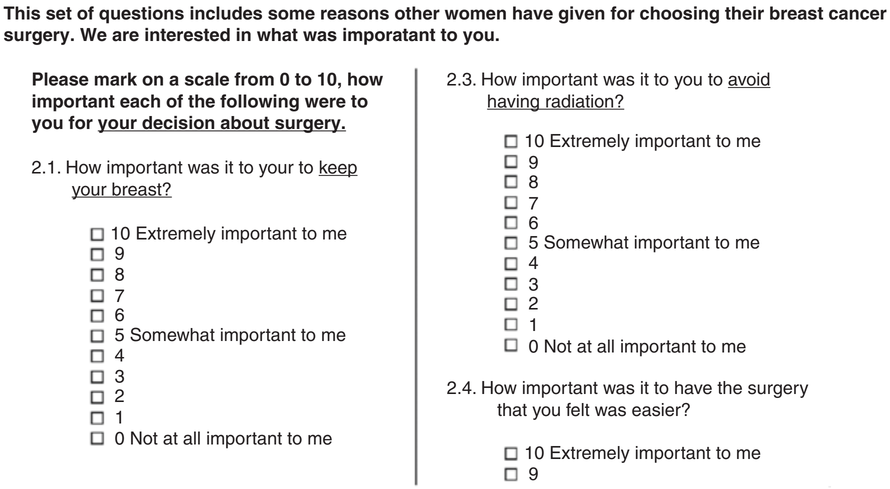
Treatment Choice. Another approach to measuring the outcome of a decision is to consider the choice itself—i.e., which surgical procedure was chosen. Did the patient choose active surveillance, brachytherapy, or prostatectomy for treatment of his low-risk prostate cancer? This approach is particularly appropriate for studies that seek to understand how decision-making affects utilization and practice variations. The choice itself may be the primary outcome of a study that aims to identify factors associated with treatment.

Decision Quality. A more recent approach to measuring decisional outcomes is the development of decision quality measures. An international group of experts in decision science and health care quality defines decision quality as the degree to which the decision is informed and consistent with patient preferences [14]. Based on this definition, decision quality instruments for various surgical decisions have been developed (Fig. 18.1). Each measure has two key components: a knowledge scale and a scale to measure patient preferences or values. Each scale is specific to the clinical decision in question, such as breast cancer surgery or hip replacement.

18.3 Interventions to Improve Decisions

18.3.1 Why Intervene?

Geographic variations in surgical practice have raised questions about the quality of decisions for those procedures. For example, the rates of breast reconstruction, total



This set of questions includes some reasons other women have given for choosing their breast cancer surgery. We are interested in what was imporatant to you.

Please mark on a scale from 0 to 10, how important each of the following were to you for your decision about surgery.

2.1. How important was it to your to keep your breast?

□ 10 Extremely important to me
□ 9
□ 8
□ 7
□ 6
□ 5 Somewhat important to me
□ 4
□ 3
□ 2
□ 1
□ 0 Not at all important to me

2.3. How important was it to you to avoid having radiation?

□ 10 Extremely important to me
□ 9
□ 8
□ 7
□ 6
□ 5 Somewhat important to me
□ 4
□ 3
□ 2
□ 1
□ 0 Not at all important to me

2.4. How important was it to have the surgery that you felt was easier?

□ 10 Extremely important to me
□ 9

Fig. 18.1 Example of questions about personal preferences from the Decision Quality Instrument © for breast cancer surgery

thyroidectomy, and many other procedures vary considerably across the country. For some conditions where a high degree of variability exists, such as back surgery, patients have reported high levels of regret about surgery. They have also reported wishing they had known more about their options. In an effort to reduce unwarranted practice variation, improve patient knowledge, and help patients make choices consistent with their personal preferences, clinicians and researchers have developed patient decision aids.

18.3.2 Decision Aids

Patient decision aids have three primary components—provision of information, clarification of patient values, and preparation of the patient for interaction with the provider. They come in different formats, including video, paper, booklet, poster board, website (Fig. 18.2), and mobile app. Decision aids are generally intended to be used by the patient before the clinical consultation (at home or in the health care setting). Some decisions aids are designed to be used by the patient and provider together. Decision aids are an adjunct to and not a replacement for patient-provider communication.

Over 105 randomized controlled trials of decision aids have been conducted, including many for surgical decisions [15]. These trials demonstrate that decision

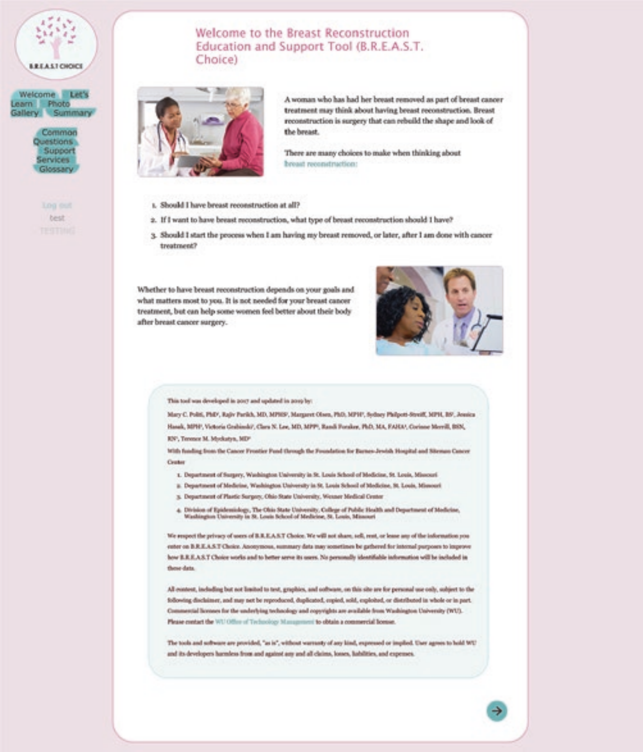


Fig. 18.2 Examples of web and paper-based decision aids

Welcome to the Breast Reconstruction Education and Support Tool (B.R.E.A.S.T. Choice)

A woman who has had her breast removed as part of breast cancer treatment may think about having breast reconstruction. Breast reconstruction is surgery that can rebuild the shape and look of the breast.

There are many choices to make when thinking about breast reconstruction:

1. Should I have breast reconstruction at all?
2. If I want to have breast reconstruction, what type of breast reconstruction should I have?
3. Should I start the process when I am having my breast removed, or later, after I am done with cancer treatment?

Whether to have breast reconstruction depends on your goals and what matters most to you. It is not needed for your breast cancer treatment, but can help some women feel better about their body after breast cancer surgery.

This tool was developed in 2017 and updated in 2019 by:

Mary C. Politi, PhD[1], Rajiv Parikh, MD, MPHS[2], Margaret Olsen, PhD, MPH[2], Sydney Philpott-Streiff, MPH, BS[1], Jessica Hasak, MPH[2], Victoria Grabinski[1], Clara N. Lee, MD, MPP[3], Randi Foraker, PhD, MA, FAHA[4], Corinne Merrill, BSN, RN[1], Terence M. Myckatyn, MD[1]

With funding from the Cancer Frontier Fund through the Foundation for Barnes-Jewish Hospital and Siteman Cancer Center

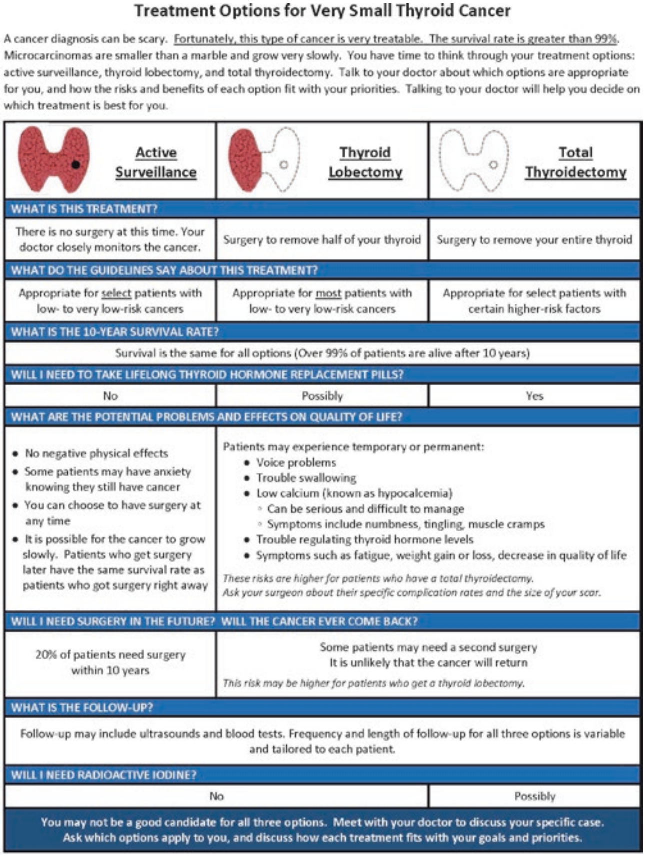
1. Department of Surgery, Washington University in St. Louis School of Medicine, St. Louis, Missouri
2. Department of Medicine, Washington University in St. Louis School of Medicine, St. Louis, Missouri
3. Department of Plastic Surgery, Ohio State University, Wexner Medical Center
4. Division of Epidemiology, The Ohio State University, College of Public Health and Department of Medicine, Washington University in St. Louis School of Medicine, St. Louis, Missouri

We respect the privacy of users of B.R.E.A.S.T Choice. We will not share, sell, rent, or lease any of the information you enter on B.R.E.A.S.T Choice. Anonymous, summary data may sometimes be gathered for internal purposes to improve how B.R.E.A.S.T Choice works and to better serve its users. No personally identifiable information will be included in these data.

All content, including but not limited to text, graphics, and software, on this site are for personal use only, subject to the following disclaimer, and may not be reproduced, duplicated, copied, sold, exploited, or distributed in whole or in part. Commercial licenses for the underlying technology and copyrights are available from Washington University (WU). Please contact the WU Office of Technology Management to obtain a commercial license.

The tools and software are provided, "as is", without warranty of any kind, expressed or implied. User agrees to hold WU and its developers harmless from and against any and all claims, losses, liabilities, and expenses.

Fig. 18.2 (continued)



Treatment Options for Very Small Thyroid Cancer

A cancer diagnosis can be scary. Fortunately, this type of cancer is very treatable. The survival rate is greater than 99%. Microcarcinomas are smaller than a marble and grow very slowly. You have time to think through your treatment options: active surveillance, thyroid lobectomy, and total thyroidectomy. Talk to your doctor about which options are appropriate for you, and how the risks and benefits of each option fit with your priorities. Talking to your doctor will help you decide on which treatment is best for you.

Active Surveillance	Thyroid Lobectomy	Total Thyroidectomy
WHAT IS THIS TREATMENT?		
There is no surgery at this time. Your doctor closely monitors the cancer.	Surgery to remove half of your thyroid	Surgery to remove your entire thyroid
WHAT DO THE GUIDELINES SAY ABOUT THIS TREATMENT?		
Appropriate for select patients with low- to very low-risk cancers	Appropriate for most patients with low- to very low-risk cancers	Appropriate for select patients with certain higher-risk factors
WHAT IS THE 10-YEAR SURVIVAL RATE?		
Survival is the same for all options (Over 99% of patients are alive after 10 years)		
WILL I NEED TO TAKE LIFELONG THYROID HORMONE REPLACEMENT PILLS?		
No	Possibly	Yes
WHAT ARE THE POTENTIAL PROBLEMS AND EFFECTS ON QUALITY OF LIFE?		
• No negative physical effects • Some patients may have anxiety knowing they still have cancer • You can choose to have surgery at any time • It is possible for the cancer to grow slowly. Patients who get surgery later have the same survival rate as patients who got surgery right away	Patients may experience temporary or permanent: • Voice problems • Trouble swallowing • Low calcium (known as hypocalcemia) ◦ Can be serious and difficult to manage ◦ Symptoms include numbness, tingling, muscle cramps • Trouble regulating thyroid hormone levels • Symptoms such as fatigue, weight gain or loss, decrease in quality of life *These risks are higher for patients who have a total thyroidectomy.* *Ask your surgeon about their specific complication rates and the size of your scar.*	
WILL I NEED SURGERY IN THE FUTURE?	**WILL THE CANCER EVER COME BACK?**	
20% of patients need surgery within 10 years	Some patients may need a second surgery It is unlikely that the cancer will return *This risk may be higher for patients who get a thyroid lobectomy.*	
WHAT IS THE FOLLOW-UP?		
Follow-up may include ultrasounds and blood tests. Frequency and length of follow-up for all three options is variable and tailored to each patient.		
WILL I NEED RADIOACTIVE IODINE?		
No		Possibly

You may not be a good candidate for all three options. Meet with your doctor to discuss your specific case. Ask which options apply to you, and discuss how each treatment fits with your goals and priorities.

aids improve patient knowledge and accuracy of risk perception. They can increase the likelihood that patients choose an option that aligns with their values. Decision aids can also reduce decisional conflict, increase patient activation, and increase the probability that patients reach a decision. Overall, decision aids have a positive effect on communication and increase satisfaction with decisions and the decision-making process, often without adding significant time burden and sometimes even reducing visit time.

In trials that measured treatment choice as a primary outcome, decision aids have been associated with fewer invasive treatments. For example, a population-based study of decision aid implementation in a large HMO found lower rates of hip and knee replacement surgery in those receiving the decision aid [16]. Similar findings have been shown for treatment of low back pain.

18.3.2.1 Decision Aid Development

Patient decision aids must be carefully and systematically developed in a manner that ensures the quality, validity and reliability of the aid. The development of patient decision aids is an iterative process that commonly includes five stages prior to production

Table 18.2 Steps for patient decision aid development

Steps	Description
1	*Scoping: Define the scope* (a) Describe the health condition or problem (b) State the decision that is being considered (c) Specify the target audience (d) Explicitly identify the guiding theoretical framework
2	*Steering: Convene a steering committee* (a) Include all relevant stakeholders: Patients, families, clinicians, policy makers, decision-making experts, patient educators, etc. (b) Identify any COI
3	*Design (iterative process)* (a) Elicit patients' views on patients' information and decision support needs (b) Elicit clinicians' views on patients' information and decision support needs (c) Identify format (media) (d) Identify setting (e) Determine distribution plan (consider timing of introduction into patient pathway—How and when decision aid will be distributed to patients and/or clinicians) (f) Appraised and summarized quality of clinical evidence relevant to the decision and options, described methods for evidence review[a] (g) Develop prototype
4	*Alpha testing* (a) Review by patients/family members (for comprehensibility and usability) (b) Review by clinicians (for acceptability and usability) (c) Review by other experts (d) Review results with steering committee and revise as necessary
5	*Beta testing* (a) Field test with patients to assess feasibility (b) Field test with clinicians to assess feasibility (c) Review field test results with steering committee and finalize/distribute

COI conflicts of interest

[a] Opportunity to perform and publish a systematic review and meta-analysis

of a final product: defining the scope, convening a steering committee, designing a prototype, alpha testing with patients and physicians, and beta testing in 'real life' conditions [17]. Table 18.2 describes the details of each of these steps. Different aspects of the development process can be turned into individual research projects. For example, you could perform a study to assess the informational and decisional needs of the patient population and their requirements for decisional support. You could also perform a systematic review to summarize the available evidence.

18.3.2.2 Reporting and Evaluating Decision Aids

Once a patient decision aid has been rigorously developed, it is critical to transparently report the development process and to evaluate the decision aid. The International Patient Decision Aid Standards (IPDAS) Collaboration developed and published standards for the evaluation and reporting of decision aids [18]. These include criteria for: the development process, how probabilities are presented, the use of patient testimonials, how decision aids are disseminated on the internet, and addressing health literacy. The IPDAS criteria have generally met with wide acceptance in the medical decision-making research community and should be considered in any study to evaluate the quality of decision aids. The IPDAS Collaboration also published "Standards for Universal reporting of patient Decision Aid Evaluation studies" referred to as the SUNDAE checklist [19]. The list includes 26-items and is intended for authors and researchers designing and reporting on studies evaluating patient decision aids. Decision aid development is generally followed by testing of its effectiveness and then dissemination and implementation. The design of studies for these additional steps is covered elsewhere in this book.

18.3.3 Other Interventions to Improve Decisions

While decision aids are the most commonly studied intervention to improve patient decision-making, other types of decision support interventions (DSIs) exist. For example, question prompt lists are structured lists of questions believed to improve patient engagement by increasing patient activation. Decision boards, coaching, and audio-recording of consultations with feedback are additional types of DSIs.

18.4 Preference Assessment

Scientifically rigorous research is key to understanding the priorities and preferences of patients, their caregivers, and other stakeholders, and facilitating patient-centered care. Eliciting patient preferences involves the quantitative assessment of the relative desirability or acceptability of two or more alternative health states, services, or interventions. It is central to health economics and outcomes research and encompasses many types of methods. Preference elicitation can be used as a research technique or incorporated into a decision aid to generate a recommendation to the user. Preference exploration utilizes qualitative methods, such as semi-structured interviews, to examine the intensiveness (depth) and extensiveness (breadth) of patient experiences.

In general, four approaches to preference elicitation exist: stated preferences, health-state utility, structure weighting, and revealed preferences. Stated preferences directly measure the desirability or acceptability of potential health care interventions to a patient. With health state utilities, preferences for a particular health state frequently represent an entire patient population and are normalized are on a scale from 0 to 1. For structured weighting, methods often involve comparing and ranking or rating specific attributes related to a decision. Revealed preferences quantify the treatment or intervention the patient actually chose.

These approaches for eliciting patient preferences commonly use discrete choice based techniques where patients have to choose between two or more alternative that each have two or more attributes as well as threshold-related techniques that aim to determine the point at which a patient would change their decision and choose an alternate treatment by varying the level of one or more attributes (Fig. 18.3). Ranking attributes or alternatives in an order from best to worst and rating attributes on an ordered scale are also methods commonly used to elicit patient

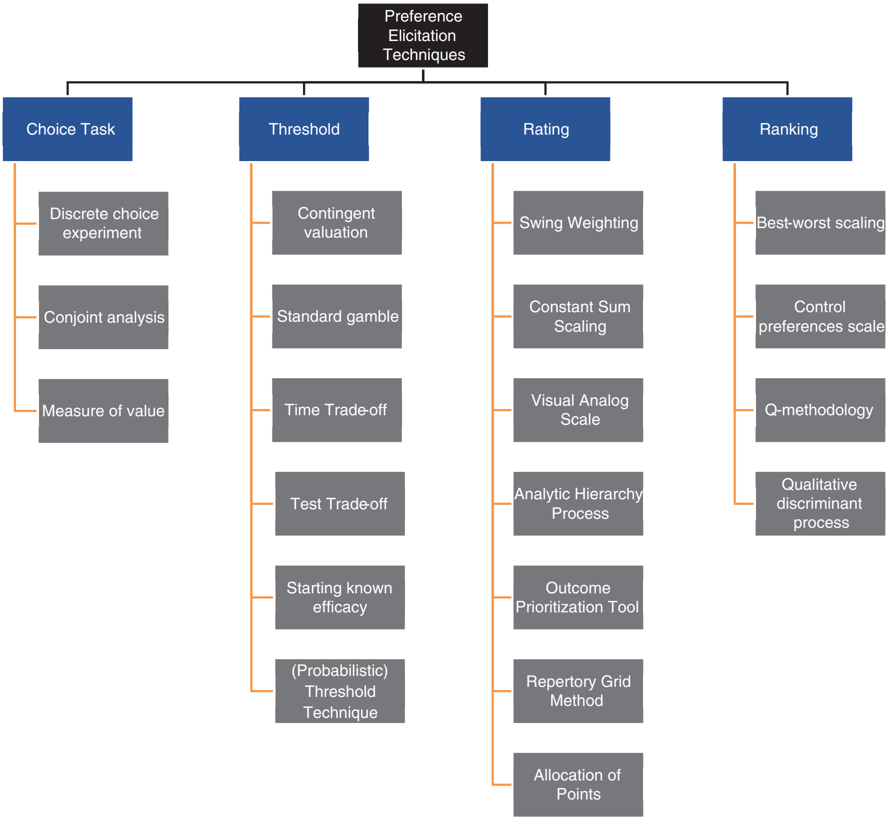
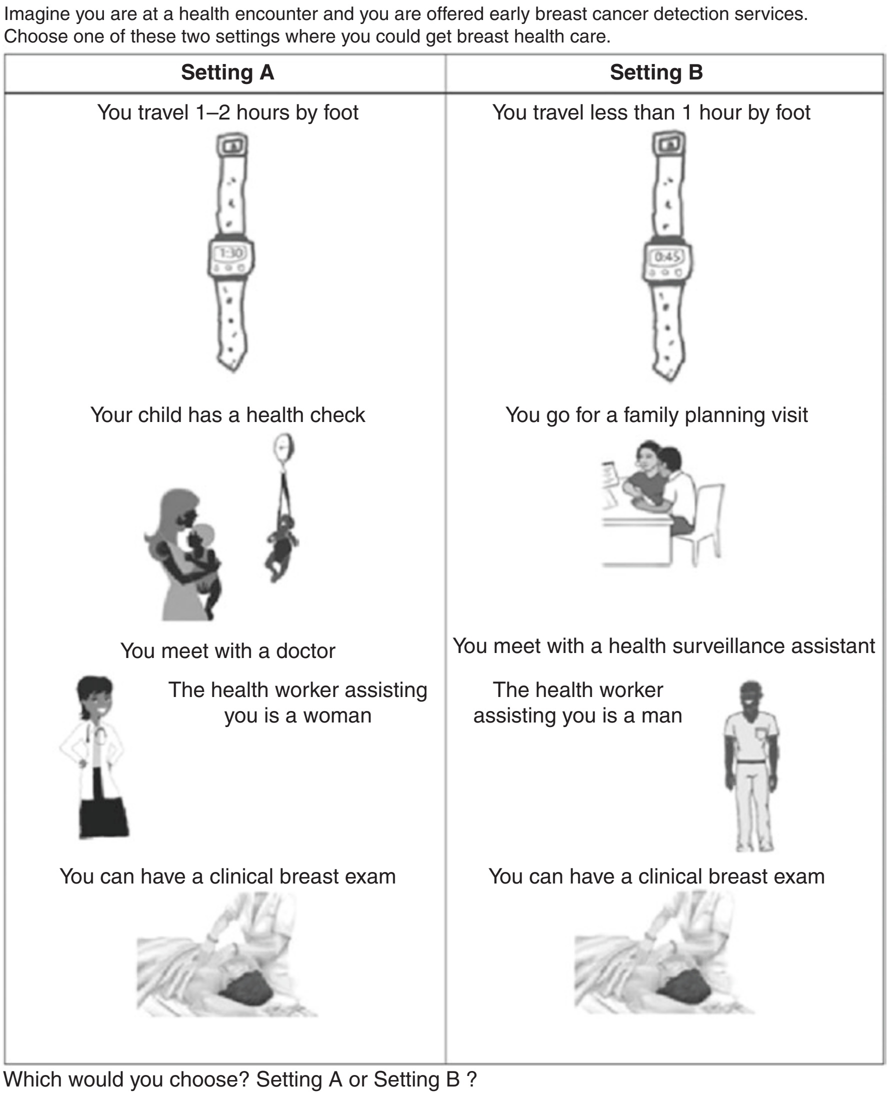


Fig. 18.3 Techniques for measuring patient preferences

preferences. The goal of preference elicitation can be identifying the attribute that matters most to the person, eliciting the relative importance of each attribute, or forcing tradeoffs between the attributes.

18.4.1 Stated Preference Evaluation

Stated preference methods, also known as contingent valuation, ask participants to *express* how desirable or acceptable they find the different attributes of alternative health treatments or interventions. Stated preferences differ from patient-reported outcomes and revealed preferences in that they measure a future state of what the patient *would* like, as opposed to a realized outcome of what the patient *did* like. The major types of stated preference methods used in health care research include direct elicitation, threshold technique, discrete choice experiments/conjoint analysis, and best-worst scaling exercises. These methods are complex and can be time consuming, but can provide valuable information. We recommend working with someone who specializes in these methods, such as a health economist.



Imagine you are at a health encounter and you are offered early breast cancer detection services. Choose one of these two settings where you could get breast health care.

Setting A	Setting B
You travel 1–2 hours by foot	You travel less than 1 hour by foot
Your child has a health check	You go for a family planning visit
You meet with a doctor	You meet with a health surveillance assistant
The health worker assisting you is a woman	The health worker assisting you is a man
You can have a clinical breast exam	You can have a clinical breast exam

Which would you choose? Setting A or Setting B ?

Fig. 18.4 Example of a choice scenario in a discrete choice experiment (DCE) [20]

Direct Elicitation. In direct elicitation, the respondent is presented with two real-world alternative treatments, each with a list of specific features or attributes. The respondent is asked a single, direct question about which treatment he or she prefers. Figure 18.4 shows an example task. This type of study evaluates the proportion of respondents who chose each treatment option or who would choose a particular tradeoff between attributes. One can use regression and other statistical techniques to evaluate respondent characteristics that are associated with choosing a treatment.

Threshold Techniques. The hallmark of threshold techniques is that a respondent is presented with two alternatives, typically a reference treatment (standard of care) and a hypothetical alternative. The respondent then has to choose between the two alternatives. The level of one attribute in the target treatment is systematically varied until the respondent prefers the other treatment. Two specific types of threshold techniques for measuring stated preferences are discussed below in the section on health utility measurement, standard gamble and time trade-off.

Discrete-Choice Experiments (DCE). Studies involving DCEs allow the researcher to elicit preferences without the participant having to state his/her preferred option or treatment. In a DCE, participants are typically presented with a series of alternative hypothetical scenarios that do not necessarily represent realistic choices. Each scenario contains a set of variables or attributes (usually ≤5), each of which has multiple levels. For example, in a study examining preferences for early detection of breast cancer, an attribute might be the distance to the screening center and the levels might be 30 min, 1 h, and 2 h away [20]. Other attributes might be who performs the screening (with levels of doctor, nurse practitioner, or health surveillance assistant) or the gender of the person performing the screening (male or female). Participants are systematically presented with two or three competing scenarios (two are shown in the figure—setting A and B). Participants are then asked to state their preferred choice between scenarios with varying combinations of attributes and levels.

Best-Worst Scaling (BWS). This is a type of choice experiment that is becoming increasingly popular. Design of a BWS exercise starts with identifying the specific features (or “attributes”) about the intervention/question that are most important to respondents, using qualitative data or other high quality evidence. For example, attributes for a BWS study of a new medication may identify side effects (mild, moderate, severe) and frequency of intake (daily, twice daily, weekly) among others. The number of attributes a study can investigate can range from 6 to >50 [21]. Respondents are then presented with a survey that includes a number of choice tasks (typically around 10 to 16). The participant has to rank a set of attributes from best to worst. The task is repeated multiple times, systematically varying specific attributes according to a pre-determined statistical design. Analysis results in ‘relative importance weights’ for the sample and for subgroups. The advantages of BWS are its ability to examine multiple attributes and reduced cognitive burden compared to DCEs.

18.4.2 Health State Utility

Health state utilities are based on expected utility theory and assess the strength of preference for a given health state. Techniques for calculating health state utilities tend to be threshold related and include: standard gamble, time-trade off, person trade-off, contingent valuation, and visual analogue scales. Health state utilities are often used in decision analysis to calculate quality-adjusted life years (QALYs). Some of the most common methods are:

Standard Gamble. This technique offers a respondent a hypothetical choice between two outcomes: (A) a sure outcome, living their remaining life expectancy in a specific health states and (B) the gamble—a risky treatment with two possible outcomes. The two outcomes have a probability, *p*, of life with optimal health or probability *1-p* of death. The objective is to identify the probability at which the respondent is ‘indifferent’ to A versus B. This then demonstrates the ‘disutility’ of a health state by observing the willingness to accept a certain risk of death in order to avoid the state.

Time Trade-Off. This technique is similar to standard gamble. It asks the respondent to choose between two states: (A) Living X years in their current health state or (B) living for a shorter period in full health. The respondent is essentially being asked whether he or she would be willing to trade years of remaining life expectancy to avoid some outcome, such as vocal cord paralysis. The duration of X can then be varied to identify the point at which the respondent is indifferent to the choice between the two options.

Visual Analog Scale. This scale is similar to a pain scale where the respondent is asked to rate a state by placing a make on a 10 cm line (horizontal or vertical) that is anchored by the best and worst possible health states.

18.4.3 Structured Weighting

Structured weighting methods are a heterogeneous group of techniques to evaluate preferences or priorities. Examples of structured weighting methods include constant sum scaling, visual analogue scale, simple direct weighting, ranking exercises, swing weighting, analytical hierarchy process, outranking methods, outcome prioritization tool, and repertory grid technique.

Constant Sum Scaling. This comparative scaling method provides a rank and weight for factors associated with a decision. Respondents are asked to allocate a fixed amount (or constant sum) of points among a set of factors. In medical decision-making research, the factors could represent a set of attributes associated with treatment outcome, such as time off work, hospital length of stay, and the need for a drain for a woman choosing between types of breast reconstruction—autologous vs. implant.

Swing Weighting. This choice task, also called multiple criteria decision analysis, can be used to elicit preferences related to benefit-risk trade-offs. Thus, it is also a stated preference method. A participant is presented with a matrix of three or more treatment alternatives (for example: no treatment, medication, and surgery) and three or more attributes (such as treatment cost, living as long as possible, and being able to work). The relative importance of each attribute is assessed by asking which treatment would result in the most important ‘swing’ from worst possible achievement of that attribute to best possible achievement. Swing weighting is similar to DCEs, but uses different method (see Fig. 18.5).

Analytical Hierarchy Process (AHP). This mathematical tool evaluates priorities in complex decisions that have subjective and objective components that are difficult to quantify. The process is based on the assumption that humans, when faced with a complex decision, naturally cluster decision elements according to their common characteristics. AHP involves creating a ranking (“hierarchy”) of decision elements and then making pairwise comparisons in each cluster (as a matrix). The result provides a weight for each element or

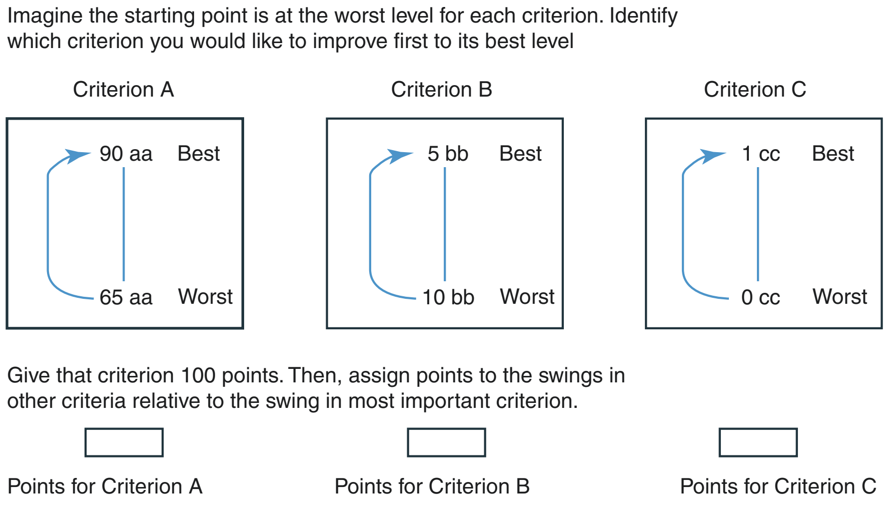

Imagine the starting point is at the worst level for each criterion. Identify which criterion you would like to improve first to its best level
Criterion A
90 aa
Best
65 aa
Worst
Criterion B
5 bb
Best
10 bb
Worst
Criterion C
1 cc
Best
0 cc
Worst
Give that criterion 100 points. Then, assign points to the swings in other criteria relative to the swing in most important criterion.
Points for Criterion A
Points for Criterion B
Points for Criterion C

Fig. 18.5 Example of a swing weighting task. (Reproduced from Thokala P, Devlin N, Marsh K, Baltussen R, Boysen M, Kalo Z, et al. Multiple Criterion Decision Analysis for Health Care Decision Making—An Introduction: Report 1 of the IPSOR MCDA Emerging Good Practices Task Force. Value in Health 2016;19:1–13.)

decision-making factor within a cluster (or hierarchy level) and also a consistency ratio (for checking data).

18.4.4 Revealed Preferences

Unlike stated or expressed preferences, revealed preferences do not ask the participant for their opinions or views. Revealed preferences are "revealed" by directly *observing* actual responses to choices about health states or services. Revealed preference methods include patient preference trials and direct questions in clinical trials. An example of a revealed preference would be the choice of health insurance on the open market.

18.5 Opportunities for Surgeon Scientists

Surgeons have unique opportunities to contribute to medical decision-making research. Unlike many decisions in medicine, most surgical decisions take place at discrete times that are readily identifiable. For example, if you want to study satisfaction with decisions in patients undergoing surgery for bladder cancer, it would be feasible to identify the specific dates of the surgical consultation, surgery, and postoperative follow-up which could be opportunities for objective measurement or intervention. Similarly, the treatment options and potential outcomes for surgical decisions tend to be discrete, making them amenable to being studied.

Few surgical decisions have been well-studied, leaving major opportunities for junior investigators to make a contribution and develop their research expertise. The ideal decision for a young surgeon to investigate is one for which some evidence exists, but for which clinicians disagree about the ideal management, or patients differ in their personal preferences. For example, the decision about sphincter preservation in rectal cancer surgery draws on a growing body of evidence about efficacy. Surgeons differ in their judgment, however, about who is a candidate for sphincter preservation, and patients differ in how they feel about living with an ostomy. Many medical decision-making researchers who have methodological expertise are actively seeking opportunities to collaborate with surgeons who have front-line experience, clinical insight, and access to patients. Surgeon investigators who develop their own skills in medical decision-making research methods could build their academic career by creating a unique niche in surgical decision-making research.

18.6 Conclusion

The decision sciences encompass a broad range of disciplines and content areas. For surgical decisions that are preference-sensitive, many opportunities exist to evaluate the quality of decisions and to develop and test interventions, such as decision aids, to improve decisions. Preference elicitation methods are powerful tools that can be incorporated into decision aids or used on their own for studying complex surgical decisions. Medical decision-making research offers multiple fresh opportunities for surgeons who are beginning their research careers to develop an area of expertise and make a lasting impact.

References

1. Elwyn G, Hutchings H, Edwards A, et al. The OPTION scale: measuring the extent that clinicians involve patients in decision-making tasks. Health Expect. 2005;8:34–42.
2. Degner LF, Sloan JA, Venkatesh P. The control preferences scale. Can J Nurs Res. 1997;29:21–43.
3. Braddock CH, Edwards KA, Hasenberg NM, Laidley TL, Levinson W. Informed decision making in outpatient practice: time to get back to basics. JAMA. 1999;282:2313–20.
4. Shields CG, Franks P, Fiscella K, Meldrum S, Epstein RM. Rochester participatory decision-making scale (RPAD): reliability and validity. Ann Fam Med. 2005;3:436–42.
5. Brown JB, Stewart MA, Ryan BL. Assessing communication between patients and physicians: Measure of Patient-Centered Communication (MPCC). Working Paper Series, Paper #95-2, 2nd ed. London, Ontario, Canada: Thames Valle family Practice Research Unit and Centre for Studies in Family Medicine; 2001.
6. Ford S, Hall A, Ratcliffe D, Fallowfield L. The medical interaction process system (MIPS): an instrument for analyzing interviews of oncologists and patients with cancer. Soc Sci Med. 2000;50:553–66.
7. Elwyn G, Barr PJ, Grand SW, Thompson R, Walsh T, Ozanne EM. Developing CollaboRATE: a fast and frugal patient-reported measure of shared decision making in clinical encounters. Patient Educ Couns. 2013;93:102–7.
8. Hoerger M, Chapman BP, Mohile SG, Duberstein PR. Development and psychometric evaluation of the decisional engagement scale (DES-10): a patient-reported psychosocial survey for quality of cancer care. Psychol Assess. 2016;28:1087–100.

9. O'Connor AM. Validation of a decisional conflict scale. Med Decis Making. 1995;15:25–30.
10. Brehaut J, O'Connor A, Wood T, et al. Validation of a decision regret scale. Med Decis Making. 2003;23:281–92.
11. Holmes-Rovner M, Kroll J, Schmitt N, et al. Patient satisfaction with health care decisions: the satisfaction with decision scale. Med Decis Making. 1996;16:58–64.
12. Legare F, Kearing S, Clay K, Gagnon S, D'Amours D, Rousseau M, O'Connor A. Are you SURE? Assessing patient decision conflict with a 4-item screening test. Can Fam Physician. 2010;56:e308–14.
13. Song MK, Donovan HS, Piraino BM, Choi J, Bernardini J, Verosky D, Ward SE. Effects of an intervention to improve communication about end-of-life care among African Americans with chronic kidney disease. Appl Nurs Res. 2010;23:65–72.
14. Stacey D, Légaré F, Lewis K, Barry MJ, Bennett CL, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev. 2017;(4):CD001431. https://doi.org/10.1002/14651858.CD001431.pub5.
15. Sepucha KR, Fowler FJ Jr, Mulley AG Jr. Policy support for patient-centered care: the need for measurable improvements in decision quality. Health Aff (Project Hope). 2004;Suppl Web Exclusive:VAR54–62.
16. Arterburn D, Wellman R, Westbrook E, et al. Introducing decision aids at group health was linked to sharply lower hip and knee surgery rates and costs. Health Aff (Project Hope). 2012;31:2094–104.
17. Coulter A, Stilwell D, Kryworuchko J, Mullen PD, Ng CJ, van der Weijden T. A systematic development process for patient decision aids. BMC Med Inform Decis Mak. 2013;13(Suppl 2):S2.
18. Elwyn G, O'Connor A, Stacey D, Volk R, Edwards A, Coulter A, et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ. 2006;333:417.
19. Sepucha KR, Abhyankar P, Hoffman AS, Bekker HL, LeBlanc A, Levin CA, Ropka M, Shaffer VA, Sheridan SL, Stacey D, Stalmeier P, Vo H, Wills CE, Thomson R. Standards for UNiversal reporting of patient decision aid evaluation studies: the development of SUNDAE checklist. BMJ Qual Saf. 2018;27:380–8.
20. Kohler RE, Lee CN, Gopal S, Reeve BB, Weiner BJ, Wheeler SB. Developing a discrete choice experiment in Malawi: eliciting preferences for breast cancer early detection services. Patient Prefer Adherence. 2015;9:1459–72.
21. Cheung KL, Wijnen BFM, Hollin IL, Janssen EM, Bridges JF, Evers SMAA, Hiligmsmann M. Using best-worst scaling to investigate preferences in healthcare. Pharmacoeconomics. 2016;34:1195–209.

Landmark Papers

Braddock CH, Edwards KA, Hasenberg NM, Laidley T, Levinson W. Informed decision making in outpatient practice: time to get back to basics. JAMA. 1999;282:2313–20.

Charles C, Gafni A, Whelan T. Decision-making in the patient-physician encounter: revisiting the shared treatment decision-making model. Soc Sci Med. 1999;49:651–61.

Elwyn G, O'Connor A, Stacey D, et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ. 2006;333:417.

Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weinstein M. Decision making in health and medicine. New York: Cambridge University Press; 2009.

Sepucha KR, Fowler FJ Jr, Mulley AG Jr. Policy support for patient-centered care: the need for measurable improvements in decision quality. Health Aff (Project Hope). 2004;Suppl Web Exclusive:VAR54–62.

Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the panel on cost-effectiveness in health and medicine. JAMA. 1996;276:1253–8.

19 Decision Analysis and Cost-Effectiveness Research

Benjamin K. Poulose

Abstract

Physicians, surgeons, and hospital systems face challenging decisions on a daily basis. In certain complex decisions, the formal methods of decision analysis can be useful to incorporate the many known factors involved as well as the uncertainties associated in the decision making process. Often, benefits to a group of individuals is weighed against limited resources. Decision analytic techniques can range from simple decision trees to complex cost-effectiveness models involving simulation techniques. The overall advantage of these approaches is the ability to incorporate uncertainty associated with the decision into the modeling process and to perform sensitivity analysis across important decision factors that may impact the outcome of the decision. The ultimate impact and trustworthiness of decision analytic models rely upon the integrity, truthfulness, and rigor imparted by investigators and clinicians.

19.1 Decision Trees and Probabilities

The building block of decision analysis and analytic techniques involves creation of a decision tree to precisely evaluate a decision and its important downstream consequences. Figure 19.1 illustrates a classic decision tree comparing Surgical Therapy to Medical Therapy to treat a particular disease process. It is important to note that the most useful decision trees, or models, are ones where reasonably robust evidence exists in terms of the data going into the model. If a particular decision option has essentially no data or very poor quality data associated with it, no amount of

B. K. Poulose (✉)
Division of General & Gastrointestinal Surgery, Center for Abdominal Core Health, The Ohio State University Wexner Medical Center, Columbus, OH, USA
e-mail: benjamin.poulose@osumc.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_19

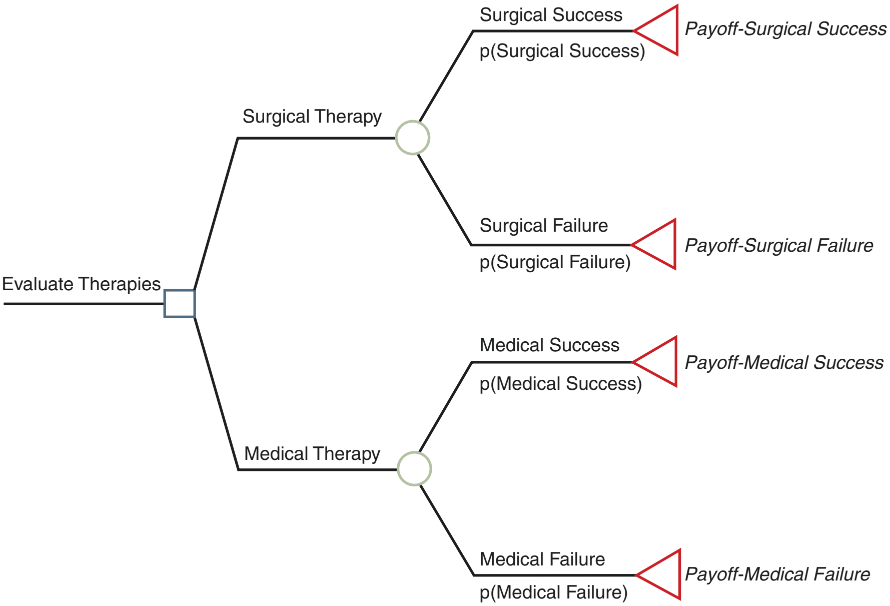


Fig. 19.1 Basic decision tree comparing a surgical therapy to medical therapy

modeling or complex analysis can compensate for this lack of information. It is the academic responsibility of the investigator to determine this up front before studies are published and referenced with convincing titles based on meaningless or non-existent data. This is especially true in the surgical literature where many peer reviewers may not be familiar with techniques used in decision analysis.

In the sample decision tree, each Therapy choice is has two mutually exclusive outcomes of Success or Failure. As modeled, there can only be these two ‘fates’ of Surgical Therapy or Medical Therapy. Obviously, this is a simplistic representation of reality, but serves as a useful starting point in creating a model to reflect clinical decision making. Novice investigators have a tendency to create overtly complex models when, in fact, more simplistic models will suffice to address the decision at hand. As models become increasingly complex, additional data will be required to populate that model-data that may be poor quality or nonexistent. The chance of experiencing Success or Failure with either Medical or Surgical Therapy is governed by the probability of each outcome p(Surgical Success), p(Surgical Failure), p(Medical Success), p(Medical Failure). As these are mutually exclusive-and the only possible-outcomes, adding the two probabilities for each Therapy choice must equal 1. For example, if the known probability of success for Surgical Therapy, p(Surgical Success), is 0.6, then p(Surgical Failure) must equal 0.4. Additionally, if the known probability of success for Medical Therapy, p(Medical Success) is 0.35, then p(Medical Failure) must equal 0.65. Each decision choice and associated Success or Failure are then associated with defined outcomes known as ‘payoffs.’ Payoffs, in the modeling sense of the term, can be expressed in single units such as costs or combined unit ratios of cost/effectiveness. As with any outcome, payoffs can be beneficial (decreased cost or increased quality/

quantity of life) or harmful (increased cost or decreased quality/quantity of life). Payoffs and associated possible values are discussed later in this chapter.

It is also important to keep in mind that the creation of a decision tree represents a static model or 'snapshot in time' given known inputs into the model at the time of creation. There is no concept of time passage inherent to these models (i.e. moving from left to right does not imply passage of time). This becomes important when interpreting results and equally important when trying incorporate time passage into models.

Many commercially available and open source software packages exist to facilitate model creation. It is highly recommended that investigators utilize these resource to help create, organize, and troubleshoot models. In addition, formal training is highly recommended to understanding the advantages and limitations of these methods.

19.2 Types of Cost-Effectiveness Analysis

The connotation of 'cost-effectiveness' analysis (CEA) implies some assessment of the economic impact of a decision along with its practical consequence (beneficial or harmful). Decision trees and other modeling and analytic methods are used to perform CEA. In traditional cost-effectiveness analysis, the economic impact is measured in monetary terms (i.e. dollars) while effectiveness is measured in a combined outcome measure known as a quality adjusted life year (QALY). This combined measure incorporates quantity of life (in years) with a quality of life 'adjustment.' The two most common subtypes of cost-effectiveness analysis are cost-utility analysis and cost-benefit analysis. Cost-utility analysis represents a specific type of CEA that focuses on life expectancy in a healthcare context [1]. In cost-benefit analysis, the benefits and harms of a treatment are expressed solely in monetary terms. Gains in life years for an individual are expressed as the productive value (i.e. in dollars) of that individual to society during those years.

19.3 The Basics

19.3.1 Base Case and Reference Case

Central to performing CEA is the concept of the Base Case and Reference Case. The Base Case refers to evaluation of a CEA model using a preferred set of assumptions and model inputs [2]. In surgical economic analyses, this commonly involves identifying a 'typical' patient profile from epidemiologic studies and using that profile as the archetypical patient for model analysis. This approach is useful from a clinical standpoint as it allows easy grasp of results in the context of a clinical scenario. To make comparisons across CEA increasingly comparable, the Reference Case concept has been accepted as the approach of choice in model development and results presentation. The Reference Case provides a common set of methods, processes, and reporting for CEA, especially for healthcare purposes [2].

19.3.2 Perspective

With any type of CEA, the perspective of the study should be predefined prior to modeling and analysis. This has crucial bearing in terms of the economic component of analysis, but also is important when considering what is included in the model. For example, hospitals and providers may be responsible for certain costs and insensitive to others, while payors may be responsible for a separate set of costs. Patients, on the other hand, might be sensitive to neither hospital/physician or payor costs and incur costs of their own. Traditional methods of CEA have specified the precise perspective of the analysis in the past (provider, hospital, payor, patient, etc.) which has led to much confusion and difficult in comparing studies. To help mitigate this, newer recommendations simply define two perspectives important to healthcare evaluations. The first perspective is that of the 'healthcare sector' which includes medical costs borne by payors and patients [3]. The second is the societal perspective which incorporates all costs and health effects regardless of which group incurs the costs and which group obtains benefits [3].

19.3.3 Time Horizon

The time span involved for a particular decision has an impact on the methods used and also the outcomes themselves. For example, very short term results (<30 days after an intervention) have a different set of considerations as opposed to evaluations that span the entire life of a patient. For this reason, the time horizon of the analysis should be specified at the study design stage and explicitly stated when reporting results. Simple decision trees (and basic analyses) can suffice for short term decisions and decision with a fixed time frame (i.e. evaluate surgical outcomes 5 years after intervention). More complex methods (e.g. Markov models, below) can be used to model a patient's health state over the course of their entire life. The three important aspects of incorporating time into models is outlined in Table 19.1.

The concept of discounting bears special mention. When evaluating costs and effects into the future, one must account for the differential value of these concepts between the present and the future. As a specific example, a dollar is worth more today than it would be in the future. This also applies to effects (i.e. health benefits). Because the modeling exercise is occurring in the present, the future costs and effects need to be 'discounted' as they are assessed into the future. The standard rate of discounting is 3% each year.

Table 19.1 Important aspects of incorporating time into decision analysis models [4]

1. How far into the future to extend the analysis
2. How to simulate time (discrete-time and continuous-time models, cycle lengths, converting rates and probabilities)
3. Discounting future costs and effects to their present values

19.4 The Payoffs: Measuring Costs and Effectiveness

As CEA models are developed, key events (e.g. Surgical Success, Medical Failure, see Fig. 19.1) are assigned costs and measures of effectiveness. During this process, it is important to keep track of the specific sources of these model inputs to assess them for validity and reporting. Only after the perspective has been defined, can individual cost and effectiveness measures be appropriately incorporated into models.

19.4.1 Costs

Healthcare costs incorporated into CEA models should take into account the perspective of the analysis and be accurately reflective of the value for the goods and/or services evaluated. An exhaustive discussion of healthcare costs is beyond the scope of this text. However, a few key concepts should be kept in mind when determining costs for use in CEA. Substituting charges (instead of costs) for CEA should be avoided. In certain situations where no additional information is available other than charge information, cost-to-charge-ratios can be utilized to help estimate costs. Costs used in models should be normalized to a particular year of analysis. Costing information is usually derived from a variety of sources obtained over different years. Utilization of the Consumer Price Indices can convert monetary amounts from different years into a common year for analysis. The cost components for the two recommended CEA perspectives (Health Care Sector and Societal) are summarized in Table 19.2.

Table 19.2 Cost components of the health care sector and societal perspectives [3]

	Reference case perspective	
Cost component	Health care	Societal
Formal health care sector		
Costs paid by third-party payors	Yes	Yes
Costs paid out-of-pockets by patients	Yes	Yes
Informal health care sector		
Patient-time costs	No	Yes
Unpaid caregiver-time costs	No	Yes
Transportation costs	No	Yes
Non-health care sector		
Productivity	No	Yes
Consumption	No	Yes
Social services	No	Yes
Legal or criminal justice	No	Yes
Education	No	Yes
Housing	No	Yes
Environment	No	Yes

19.4.2 Effectiveness

Effectiveness in CEA is typically measured using QALYs. A QALY is a generic measure of disease calculated by the number of years spent in a health state adjusted for the disease state itself. To operationalize QALYs in a CEA model, the concept of health utility must be understood. Based in health utility theory, a value of zero utility is assigned to the state of death, while a value of one is assigned to perfect health. Disease states fall between 0 and 1; many methods exist to determine what value is associated with a particular disease state. Several groups have compiled available health utility measurements into searchable databases, greatly facilitating model development [5, 6]. Decision tree software packages facilitate calculation of QALYs while incorporating health utility values for particular disease states. Great care must be taken to correctly account for harmful outcomes (i.e. complications) to ensure accurate modeling.

19.5 Practical Tips for Developing a CEA Model

Although at first glance developing a robust CEA model may seem like a daunting and difficult task, several practical tips can be employed. Either acquiring the skills and knowledge to work with decision trees or collaborating with someone with this skillset is essential. A carefully designed investigative question evaluating two options for a decision is best. A reasonable amount of information in the established literature should exist to populate the model inputs. In general, information should be gathered from high level evidence sources, and moving toward lower level evidence when needed. Keeping diligent track of probabilities, variable names, costs, utilities, and associated sources of information is essential during model development. This can easily be done using a spreadsheet which can often be linked directly to the model itself. Several versions of the model itself should be stored in separate files, making it possible to 'resurrect' portions of older models that are found to be useful.

19.6 Analyzing Models

Once a model has been refined, checked for errors, inconsistencies, and redundancies, it is ready for analysis. Analytic methods range from simple 'roll back' of the model to complex, high-resource techniques such as Monte Carlo simulation. Using more complex techniques when a simpler method will suffice should be avoided.

19.6.1 Rolling Back and Expected Values

At its most basic level a model 'roll back' utilizes all inputs (probabilities, costs, measures of effectiveness) and determines the expected values for each therapeutic option or strategy. The therapeutic option with the lowest cost and highest accrual of QALYs becomes the preferred strategy and answers the question at hand. For a typical model, results are given for each therapeutic option in terms of cost, incremental cost, effectiveness, incremental effectiveness, incremental cost-effectiveness ratios, and cost-effectiveness ratios. Strategies that are least costly and most effective are designated as 'dominant' to other strategies. A useful way of testing a model is to place extreme inputs for probabilities, costs, or measures of effectiveness and determine if the roll back results behave as expected.

19.6.2 Sensitivity Analysis

Sensitivity analysis is a useful way to determine the impact of a particular variable or set of variables on the choice of treatment. Specifically, sensitivity analysis can be used to evaluate the uncertainty associated with a particular variable. In one way sensitivity analysis, a single variable's value is plotted over its plausible range and the effect on an outcome is evaluated. If the variable and range have a significant impact in the model, the outcome measure will vary widely as the variable values also change. If the variable has minimal impact on the model, there will be no change in the outcome variable across the range evaluated. Two way sensitivity analysis performs a similar function, except two variables are simultaneous changed across their plausible ranges. Optimal strategies are indicated by different areas on a graph corresponding to the values of the two variables. In three way sensitivity analysis, multiple two way analyses are performed across the plausible range of a third variable.

19.6.3 Performing Complex Analyses

Many variations on analysis of decision trees exist. These range from multi-way probabilistic sensitivity analysis that incorporate many aspects of uncertainty to achieve the optimal strategy, to Markov modeling and Monte Carlo simulation. Markov modeling is often used to represent events that recur over time using the concept of state-transition models. In these models, an 'individual' experience is defined by existence in multiple states. A simple state-transition model allows transition between states of being well, existing in disease, or death. Markov models are very useful to evaluate disease process and treatments over time. Monte Carlo simulation is a resource-intensive computational method that helps evaluate high level of uncertainty within models using random values and probability distributions.

19.7 Standards for Methodology and Reporting of Cost-Effectiveness Analysis

Sanders et al. have summarized the most recent standards for methodology and reporting of CEA [3]. As clinicians and research teams increasingly use these methods to answer clinical questions, great care should be taken to present all model assumptions up front. In addition, care should be taken to avoid extreme use of jargon often associated with CEA, possibly alienating the intended group of individuals the study was meant to impact. The results of even the most complex CEAs can be presented in a fairly straightforward manner, making it easy to understand the results and potential impact. Finally, it cannot be overemphasized that the quality and integrity of results from any CEA is directly related to the quality of information input into the model, the quality of the model itself, and the integrity of the investigators in seeking truthful answers with minimal bias.

References

1. Introduction to Drug Utilization Research: Chapter 4: Economic aspects of drug use (pharmacoeconomy): 4.4 Cost-utility analysis. http://apps.who.int/medicinedocs/en/d/Js4876e/5.4.html#Js4876e.5.4. Accessed 31 Mar 2019.
2. Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the panel on cost-effectiveness in health and medicine. JAMA. 1996;276:1253–8.
3. Sanders GD, Neumann PJ, Basu A, et al. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: second panel on cost-effectiveness in health and medicine. JAMA. 2016;316:1093–103.
4. O'Mahony JF, Newall AT, van Rosmalen J. Dealing with time in health economic evaluation: methodological issues and recommendations for practice. Pharmacoeconomics. 2015;33:1255–68.
5. CEA Registry - Center for the Evaluation of Value and Risk in Health. https://cevr.tuftsmedicalcenter.org/databases/cea-registry. Accessed 31 Mar 2019.
6. A Web-based Registry of Cost-Utility Analysis | Harvard Center for Risk Analysis | Harvard T.H. Chan School of Public Health. https://www.hsph.harvard.edu/hcra/a-web-based-registry-of-cost-utility-analysis/. Accessed 31 Mar 2019.

20 Survey Research

Karen J. Brasel

Abstract

Surveys, and survey research, have become ubiquitous; as such, the value of survey research in the eyes of many has diminished. However, there are certain things that are best studied via survey, including beliefs and attitudes. Importantly, quality of life must be ascertained by survey. This chapter highlights the elements of high-quality survey research, focusing on instrument development, mode of administration, response burden, response rate, nonresponse bias, and reporting survey results. Specific strategies for developing good questions and increasing response rates are outlined. Specific analytic techniques related to nonrandom sampling frames, statistical packages, and Likert-type questions are reviewed.

20.1 Introduction

It seems to happen on an almost daily basis anymore—an email invitation to respond to a survey. Survey evaluations of Grand Rounds, departmental events, CME activities. Surveys from associations, surveys soliciting expert opinion. Surveys targeting me as a member of a particular association, or from an investigator concerned about a particular research question. Due to the plethora of survey requests, and the ubiquitous nature of those that are poorly performed, survey research has become somewhat trivialized. Some of this is due to the use of surveys for market research, continuing medical education activities, and solicitation of opinions; for many of these surveys, scientific rigor is neither desired nor required. However, there are certain research questions that are best answered by survey, and it is possible to

K. J. Brasel (✉)
Department of Surgery, Oregon Health & Science University, Portland, OR, USA
e-mail: brasel@ohsu.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_20

perform high-quality survey research. However, as with any other type of scientific inquiry, poorly done survey research will lead to bad science. Which translates into potentially misleading, if not dangerous, information and implications [1].

20.2 When to Use a Survey

Surveys can easily answer questions about knowledge, and best answer questions about attitudes and beliefs. They may also be useful to gather information on behaviors and practices, although this is dependent on the type of behavior or practice. The burgeoning space of medical education research uses surveys liberally, as approximately 50% of original medical education research is survey-based [2]. For sensitive subjects, an anonymous survey may be more likely to reveal truth. For other subjects, data sources such as medical records may provide more reliable information.

20.3 Developing Questions

It may seem obvious, but the first and most important step in survey research is deciding what your research question is. This will allow you to determine whether you can use a previously validated survey or whether you must develop one on your own. Whenever possible, it is preferable to use a previously validated instrument without modification. This ensures that the questionnaire is reliable, valid, and responsive to change. Modifying a validated instrument reduces both its power and validity [3].

If you are unable to use a validated survey to answer your question, spend time developing your survey instrument. This most important step is the aspect of survey research that is most often neglected. The quality of survey data will only be as good as the questions asked on the survey. This is true regardless of the ultimate mode of survey administration, and the techniques described below work for all modes. The development of a survey follows a standard process of item generation, item reduction, questionnaire formatting, composition, and pilot testing [4].

Often, both qualitative and quantitative techniques must be employed to create the optimal research tool. Clearly, the investigator and the research team have an idea of what questions they would like to include on a particular survey. Employing qualitative techniques using focus groups during the item generation stage provides additional information about the topic area being investigated, and helps examine the assumptions brought by the research team during initial drafting of questions. Focus groups also help with how specific terms and vocabulary are understood by the population to be surveyed [5]. As an example, in developing a post-injury quality of life survey, you might ask a group of trauma patients “We are interested in the factors that have affected your quality of life since your injury. Can you tell us some of the things that have affected your quality of life both positively and negatively?” You would also want to ensure that each member of the group has an understanding of the term or concept quality of life.

Each question that you include must be critical, which is the purpose of the item reduction stage. Avoid including questions "just because you're interested" or "because it would be nice to know". All questions should be clear and without bias. Avoid questions with two possible answers and leading questions that have a socially desirable answer. Normalizing statements, such as "It can be difficult to…" prior to asking about a sensitive item increase the likelihood of an honest answer. Questions that require a closed response are the easiest to analyze, so if at all possible try to frame the question to require a closed response [5–7]. Focus groups can also help in the item reduction stage, as can ranking or rating questions by pilot groups or content experts.

Acquiescence is the tendency to endorse any assertion made in a question, regardless of content. Using declarative statements lessens this effect, although when level of agreement is what needs to be measured it is impossible to avoid this effect completely. The magnitude of the acquiescence effect is approximately 10%; approximately 52% of respondents agreed with an assertion, while 42% of respondents disagreed with its opposite [7]. There are two approaches to minimizing the effect of acquiescence—the first is to ask all questions in one direction (usually positive), and the second is to use two questions, one asked positively and one negatively, to check on the magnitude of the effect in a specific survey.

Missing data tends to be more prevalent at the end of surveys, likely related to response burden. Therefore, ask the most important questions at the beginning, leaving the demographic questions to the end. Questions should be grouped thematically, as respondents fatigue if they must keep returning to a particular topic. In addition, questions should progress either from the general to the specific (funneling) or from the specific to the general (reverse funneling). If appropriate to the population, have your survey tested for language level and target the lowest likely educational level in your sample.

Pilot testing your survey will provide valuable information that can help you improve the quality of data in addition to your response rate. There are two options for initial pilot testing, with and without observation. Observing a small number of people answering your survey will give you information about which questions take respondents longest to answer, suggesting that they may be awkwardly or ambiguously worded. You are also able to debrief this small group about areas of concern and get their ideas for improvement. Field testing a pilot group without observation relies on their willingness to answer each question as well as provide written feedback for each question. Both methods should ask this pilot group to give feedback on the aesthetics of the survey, the ease with which it was completed, and their interest in completing the entire questionnaire. The completed questionnaire should then be tested with a final pilot group, which allows calculation of the psychometric properties of the questionnaire. These data may be helpful in further item reduction by eliminating items that cluster together and do not add additional information or eliminating items that have significant floor or ceiling effects [4]. The pilot groups should be similar to your intended population, but should not be included in your final sample.

20.4 Population

For very small populations, it is desirable to obtain information from the entire population. However, for most surveys this is both impractical and expensive and therefore a sample of the population must be surveyed. The sampling frame is a list of the entire population. When qualitative surveys, such as focus groups, are performed, non-random sampling may be appropriate. However, for all other surveys, random sampling provides the most generalizable information. This can be accomplished using a random number generator or using a more systematic approach, selecting those to be surveyed from the sampling frame starting from a random point on the list at equal intervals (every fifth person on the list, for instance) [8]. In order for the results to be valid, the sampling frame must represent the population of interest.

Obtaining the sampling frame can be quite problematic for surveying health professionals, as many organizations have specific policies about handling of membership lists that specifically prohibit contact for survey research. Some are specific for email contact, allowing mail-based survey research to proceed. Many lists contain out of date information, and much effort must be spent cleaning a "dirty" or inaccurate list.

Although data from a simple random sample from the sampling frame is optimal, it is often both extremely difficult and expensive to collect. It also may not capture specific groups of interest, and may not be efficient. From a practical standpoint, probability samples are often used to obtain data more cost-efficiently using complex sample designs. Probability samples divide the sampling frame into strata, and often to clusters within strata from which the sample is subsequently drawn. Complex sample designs are likely to result in unequal probability of selection for individual units of analysis, lack of independence of individual units within randomly sample clusters, and variable effect on estimates of precision. As a result, this approach which simplifies survey administration and data collection on the front end requires more complex statistical analysis [9, 10].

The size of the sample necessary for a particular survey depends on the statistical analysis that will be performed, highlighting the importance of a well-thought analytic plan. The sample size can be calculated using a number of different computer packages or by consulting a statistician. Estimated non-response rates must be included in the sample size calculations.

20.5 Mode of Administration

Most often, administration mode is considered in light of cost, practicality, and response rate. However, it is important to know that mode of administration may introduce important specific biases related to participant demographics. It is tempting in today's internet age to consider only electronic modes of administration. This would be a big mistake, as very recent data suggest that mailed surveys and phone calls achieve the highest response rate and electronic methods miss important populations [11].

20.5.1 Face to Face

Face to face surveys are important for qualitative research, and are the primary method used for conducting focus groups. For quantitative surveys, complex questions can be asked and visual aids can be used. Response rates are generally higher than other methods, as many people find it harder to refuse a face-to-face request. Disadvantages include inefficiency, the need for training interviewers, and prohibitive costs for a sample of any reasonable size [12, 13].

20.5.2 Phone

Phone surveys allow a two-way interaction between interviewer and respondent similar to face-to-face surveys. This allows the use of complex questions, and also allows the interviewer to probe for explanatory answers. The refusal rate is higher than with face-to-face administration, but the costs are much less. In general, refusal rate is lower than with mail or internet surveys. Generating a sampling frame may be problematic, as telephone numbers may be more difficult to obtain than mail or email addresses. Random digit dialing surveys may or may not include cell phone numbers, limiting the representativeness of the population sampled [6, 12].

20.5.3 Mail

Mail surveys are similar to internet surveys in that they are self-administered, without interaction between interviewer and respondent. A basic mail survey includes the survey tool and a return envelope—this is likely to achieve a response rate around 20%. The Dillman approach, or tailored design method described below outlines the optimal process for mailed surveys in order to achieve the highest response rate, which should be greater than 60%. A strategy that includes at least a partial mail or paper component is the mode of choice when surveying physicians in order to balance response rates and efficiency. They also allow the inclusion of an incentive, one of the factors with the greatest impact on response rate [6, 12, 13]. All of these contribute the increased costs associated with mail surveys compared to all other modes of administration. Importantly, participants who are older, unmarried, retired, non-White, with less education, lower income and without private insurance may have a higher response rate when mailed surveys are used [11, 14]. Another disadvantage of mailed surveys is their tendency to have a higher rate of missing data.

20.5.4 Internet

There are several advantages to email or internet survey administration. It is the least costly method of survey administration, and in general produces the highest quality data for the least amount of effort because the data can be logic-checked during survey administration [6]. Costs of an internet-based survey are approximately 20% of a similar survey administered by mail. In addition, multimedia can be incorporated to enhance interest and engagement [8]. Incentives can be used, although they require additional effort to be expended to track and deliver.

However, there are also several disadvantages to this approach. In order to get an appropriate sampling frame, correct email addresses must be obtained for all potential respondents. Particularly for physicians, response rates to internet surveys are significantly lower than response rates to mail-only or mixed-methods surveys [8, 13]. There is an increasing concern about the lack of anonymity with internet-based surveys, particularly when compared to mail surveys. In addition, the elderly, non-English speakers and those of lower socioeconomic status have a much lower response rate to internet-based surveys [11, 14].

20.5.5 Mixed-Mode

Mixed mode surveys, which primarily include both mail and internet options for response, have some promise in terms of response rates. This is particularly true when electronic options are followed by standard mail and the respondents are given the option of completing the survey either online or by returning the paper questionnaire. The representativeness of the respondents in a mixed-mode approach does not differ when the population of interest is younger and more educated; in other populations it is less clear although equivalent results have been shown in a sample of men with prostate cancer [15–17]. This is a promising approach for surveys of physicians [13].

20.6 Response Burden

Response burden is related to response rate, with increased perceived burden related to diminished response rate. Burden is most directly related to time required to complete the survey, with several factors contributing to time. Most easily measured is length, with shorter surveys having a decreased response burden. Other factors contributing to response burden include number of pages or internet screens, poorly worded questions, difficult questions, internet screens that are difficult to navigate and other technical difficulties accessing or responding to the survey [8].

20.7 Response Rates

The response rate is simply the number of completed surveys received divided by the number of surveys sent to nonrespondents and respondents. Simple on the surface, that is. There are actually six possible ways to define response rates [18]. The first decision point in defining survey response rate is determining the eligibility of the nonrespondents. The easiest way is to assume that all nonrespondents are eligible; however, this results in a correct response rate only if that assumption is known to be factually true. The second is to estimate the probability of nonrespondent eligibility. For mailed surveys, ineligible respondents include those with a wrong mail or email address and those surveys returned to sender. In order to estimate the true number of ineligible respondents, addresses should be checked on a sample of both respondents and nonrespondents. Expressed mathematically, ARR = (R)/([R] + e[T–R–NE]), where ARR = adjusted response rate, R = eligible respondents, e = the proportion of nonrespondents estimated to be ineligible, T = total number of surveys, and NE = ineligible respondents (including return to sender). For other surveys, if it is known that a certain percentage of a survey population would be eligible (e.g. are assistant professors), that percentage of nonrespondents is used to calculate response rates. The second item to determine when calculating response rates is to determine which surveys to include as respondents. The American Association for Public Opinion Research suggests that a complete survey contain responses to a minimum of 80% of the questions [18]. An alternative definition requires 80% of the questions of interest, allowing for nonresponse to demographic questions. Best practice is to identify which of the AAPOR definitions has been used when calculating response rate [19].

Response rate is a critically important aspect of survey research, as low response rates can introduce significant bias. For instance, if 50% of respondents respond in a particular way to a specific survey item, the true percentage is between 45%–55% if the survey has a response rate of 90%, but ranges between 5%–95% if the response rate is only 10%. Higher response rates also provide greater statistical power, reducing the chances of a type II error. Finally, higher response rates allow for greater generalizability to the population the respondents represent [20].

Given the critical importance of response rates, there are a variety of strategies available to maximize return. The Dillman approach, or tailored design method (TDM), is considered standard for mail questionnaires regardless of the population being studied. The first element of the Dillman approach is to make the questionnaire respondent-friendly. This includes a survey that is easy to read, makes use of bolded sentences, indentation, clear and unambiguous questions, and is relatively short. Other elements involve the cover letter and number of contacts, all by first-class mail. The first contact may be a postcard or letter letting the potential respondent know of the upcoming survey, or it may be the initial survey with a personalized cover letter and return envelope with a real first-class stamp. The next contact is a reminder postcard 2 weeks later, and the last contact is a resending of the survey with a return envelope [21, 22].

Response rates for physicians are generally below response rates of the general public. Specific strategies to increase response rates for physician questionnaires include use of a phone contact or registered mail contact after the three first-class mail contacts and the use of an unconditional monetary incentive sent with the initial survey. These strategies can increase response rates up to 20% beyond that achieved when using the TDM approach alone [8, 23]. The amount of monetary incentive resulting in optimal response rates is unclear, with some studies showing amount directly related to response rate and others unable to establish a direct relationship between incentive amount and response rate. Often even a small monetary incentive (5$) achieves significant improvements in response rates [24]. If a monetary incentive is not possible, a nonmonetary incentive such as a pen, lottery ticket, or laser pointer can be used, although the effect may not be quite as great [25]. CME credit does not appear to be a worthwhile incentive [26].

20.8 Nonresponse Bias

A response rate of 60% or greater has been required by many journals in order for survey research to be considered for publication [27]. This figure was originally targeted in order to minimize nonresponse bias, and it is certainly true up to a point that greater response rates minimize nonresponse bias. However, the strength of the relationship between nonresponse rate and nonresponse bias may not be as strong as previously thought [22, 28]. This is because the entire population may be viewed as having both a propensity to respond or not respond based on specific characteristics of the survey, the mode of administration, and personal characteristics that may vary over time. Key to determining how close the link is between nonresponse rates and nonresponse bias is how strongly correlated the survey variable of interest is with the likelihood of responding. Some analyses suggest that efforts to improve response rates above a particular target may actually worsen the quality of the data, as nonresponders who are converted or coerced into responding may provide inaccurate information. Effects of tools to improve response rates on nonresponse bias remain unclear although examples similar estimates using the same survey with a less aggressive protocol and a lower response rate are plentiful [28, 29].

Nonetheless, calculation of nonresponse bias is essential and can be performed in a number of ways. The simplest way to do this is to compare respondents and nonrespondents (as defined by AAPOR standards) by demographic information, and provide estimates of response and nonresponse for key subgroups within the target population. Similar demographic profiles and response rates support lack of nonresponse bias, although use the simplistic assumption that the subgroup variables are the only possible causes of propensity to respond [22].

Wave response compares early and late wave respondents (based on whether they respond to the initial, second, or third request) in terms of demographics, response rates, and responses. This analysis assumes a continuum of response, suggesting that late wave respondents are more similar to nonrespondents than early wave respondents. Its weakness is that there is no direct information provided about the nonrespondents [18].

If an external source exists with which to compare information, respondent survey data can be compared to this external source. Although this does not provide direct information about the nonrespondents, it allows an estimation of whether nonresponse bias has influenced the results. The availability of external benchmarks for health-related survey research is fairly rare.

Other methods to assess nonresponse bias include collecting auxiliary variables on respondents and nonrespondents to guide attempts to balance response rates, and the use of postsurvey adjustments to test the sensitivity of the responses obtained. Whenever possible, multiple approaches to assess nonresponse bias should be used [22].

20.9 Likert Scales

A Likert scale is most commonly used to measure agreement with a particular statement. Most commonly a 5-point scale is used. One common scale has two levels of agreement, a neutral option, and two levels of disagreement (strongly agree, agree, neutral, disagree, strongly disagree). If you want to force either agreement or disagreement, the neutral option can be eliminated leaving a 4-point scale. Data reliability is not affected by offering a neutral or no opinion option, and there is some evidence that this option is most often chosen to avoid the cognitive work to generate an optimal answer. Often the levels of agreement and disagreement are collapsed in the analysis phase, leaving a functional 2- or 3-point scale. Another common scale begins at one extreme, increasing at approximately equal intervals to the opposite extreme (poor, fair, good, very good, excellent).

Most questions that use a Likert scale are actually Likert-type questions, which are single items that use the Likert response alternatives. A Likert scale is a series of four or more Likert-type items combined into a single score during data analysis. The importance of this difference is that single Likert-type items are ordinal scale observations. These observations are ranked in that each successive response is greater than the one before, but how much greater is not specified and there is no assumption of equal distance between one observation and another. Ordinal refers to the position in a list. Likert scale items, in contrast, are interval scale items which have a similar relative distance between points on a scale which does not have an absolute zero. Ordinal and interval scales are analyzed differently from each other as well as differently from nominal data (named categories without any position ranking or relative distance) [9, 30, 31].

20.10 Analysis

Survey data are often presented using descriptive statistics, measures of central tendency, estimates of parameters, and procedures to estimate relationships between multiple variables. Before using any parametric statistical test, look again at the population surveyed and the sampling frame. If the survey population was a complex sample design rather than a random sample, nonparametric statistics and other statistical approaches must be used. Complex sample designs require the use of sampling weights, in order to reduce the potential sources of bias introduced by the use of probability rather than true random samples. Sampling weight is usually included as a weight variable in addition to the stratum and cluster variables used in generating the probability sample from the sampling frame [9].

Independent of sampling weight, ordinal Likert-type data should be presented using median or mode for central tendency, frequencies to describe variability, and

Kendall's tau to analyze associations. Chi-square analysis may also be appropriate. Means can be used for interval Likert scale data, with standard deviation used for variability, and Pearson's r to describe associations. Other analyses that may be appropriate include ANOVA, t-test, and regression [31].

There are several software packages that are available to analyze survey data. These include SAS, SPSS using the separately purchased complex samples add-on module, Stata, and SUDAAN. Stata and SUDAAN offer the greatest flexibility and variety of options for analysis [9].

20.11 Reporting Results

The key points in reporting survey research begin with explaining the purpose of the research and explicitly identifying the research question. Unfortunately, reporting of survey data in the medical field is extremely inconsistent, compromising both the transparency and reproducibility of the results [32]. As with most research, the methods section is extremely important and will be the basis on which your readers will determine whether they are able to generalize your results. The research tool, or questionnaire must be described. If an existing tool is used without modification this section can be brief; if a new tool was used a detailed section on how the tool was developed and tested is important. Description of the sample includes how the potential subjects were identified, how were they contacted and how many times they were contacted, how many agreed to participate, how did the nonresponders differ, what type of definition was used for response rate and what was that rate. The analytic plan is followed by the results.

20.12 Conclusions

Obtaining high-quality data, particularly information on beliefs and attitudes, is possible via survey methodology. The survey must use a well-designed and thoroughly tested instrument, a representative sample from an appropriate sampling frame, an appropriate definition for respondents and nonrespondents, and minimize nonresponse bias. Choice of survey method balances costs, effort, intended population and response rates. Analytic methods must account for nonrandom sampling and nonresponse.

References

1. Scholle SH, Pincus HA. Survey research: think…think again. Acad Psychiatry. 2003;27:114–6.
2. Phillips AW, Friedman BT, Utrankar A, Ta A, Reddy ST, Durning SJ. Surveys of health professions trainees: prevalence, response rates, and predictive factors to guide researchers. Acad Med. 2017;92:222–8.
3. Alderman AK, Salem B. Survey research. Plast Reconstr Surg. 2010;126:1381–9.
4. Burns KE, Duffet M, Kho M, et al. A guide for the design and conduct of self-administered surveys of clinicians. CMAJ. 2008;179:245–52.

5. Fowler FJ. Improving survey questions: design and evaluation. Thousand Oaks, CA: Sage Publications; 1995.
6. Jones TL, Baxter MAJ, Khanduja V. A quick guide to survey research. Ann R Coll Surg Engl. 2013;95:5–7.
7. Krosnik JA. Survey research. Annu Rev Psychol. 1999;50:537–67.
8. Klabunde CN, Willis GB, McLeod CC, et al. Improving the quality of surveys of physicians and medical groups: a research agenda. Eval Health Prof. 2012;35:477–506.
9. West BT. Statistical and methodological issues in the analysis of complex sample survey data: practical guidance for trauma researchers. J Trauma Stress. 2008;21:440–7.
10. Thoma A, Cornacchi SD, Farrokhyar F, Bhandari M, Goldsmith CH, for the Evidence-Based Surgery Working Group Users' Guide to the Surgical Literature. How to assess a survey in surgery. Can J Surg. 2011;54(6):394–402.
11. Feigelson HS, McMullen CK, Madrid S, et al. Optimizing patient-reported outcome and risk factor reporting from cancer survivors: a randomized trial of four different survey methods among colorectal cancer survivors. J Cancer Surviv. 2017;11(3):393–400.
12. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. International J Qual Health Care. 2003;15:261–6.
13. Scott A, Jeon SH, Joyce CM, et al. A randomised trial and economic evaluation of the effect of response mode on response rate, response bias, and item non-response in a survey of doctors. BMC Med Res Methodol. 2011;11:126.
14. Hagan TL, Belcher SM, Donovan HS. Mind the mode: differences in paper vs. web-based survey modes among women with cancer. J Pain Symptom Manage. 2017;54(3):368–75.
15. Kroth PJ, McPherson L, Leverence R, et al. Combining web-based and mail surveys improves response rates: a PBRN study from PRIME Net. Ann Fam Med. 2009;7:245–8.
16. Zuidgeest M, Hendriks M, Koopman L, Spreeuwenberg P, Rademakers J. A comparison of a postal survey and mixed-mode survey using a questionnaire on patients' experiences with breast care. J Med Internet Res. 2011;13(3):e68.
17. Broering JM, Paciorek A, Carroll PR, et al. Measurement equivalence using a mixed-mode approach to administer health-related quality of life instruments. Qual Life Res. 2014;23(2):495–508.
18. The American Association for Public Opinion Research. Standard definitions: final dispositions of case codes and outcome rates for surveys. 7th ed. Chicago, IL: AAPOR; 2011.
19. Phillips AW, Friedman BT, Durning SJ. How to calculate a survey response rate: best practices. Acad Med. 2017;92(2):269.
20. Draugalis JR, Plaza CM. Best practices for survey research reports revisited: implications of target population, probability sampling, and response rate. Am Pharm Educ. 2009;73:1–3.
21. Gore-Felton C, Koopman C, Bridges E, et al. An example of maximizing survey return rates: methodological issues for health professionals. Eval Health Prof. 2002;25:152–68.
22. Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q. 2006;70:646–75.
23. Thorpe C, Ryan B, McLean SL, et al. How to obtain excellent response rates when surveying physicians. Fam Pract. 2009;26:65–8.
24. Ulrich CM, Danis M, Koziol D, et al. Does it pay to pay? Nursing Res. 2005;54:178–83.
25. Olsen F, Abelsen B, Olsen JA. Improving response rate and quality of survey data with a scratch lottery ticket incentive. BMC Med Res Methodol. 2012;12:52–62.
26. Viera AJ, Edwards T. Does an offer for a free on-line continuing medical education (CME) activity increase physician survey response rate? A randomized trial. BMC Res Notes. 2012;5:129.
27. Livingston EH, Wislar JS. Minimum response rates for survey research. Arch Surg. 2012;147:110.
28. Davern M. Nonresponse rates are a problematic Indicator of nonresponse Bias in survey research. Health Serv Res. 2013 Jun;48(3):905–12.
29. Olson K. Survey participation, nonresponse bias, measurement error bias, and total bias. Public Opin Q. 2006;70:737–58.

30. Gob R, McCollin C, Ramalhoto MF. Ordinal methodology in the analysis of Likert scales. Qual Quant. 2007;41:601–26.
31. Boone HN, Boone DA. Analyzing Likert data. J Extension. 2012;50:2T0T2.
32. Story DA, Gin V, na Ranong V, et al. Inconsistent survey reporting in anesthesia journals. Anesth Analg. 2011;113:591–5.

Qualitative Research Methods

21

Dawn M. Elfenbein and Margaret L. Schwarze

Abstract

This is a short introduction to the field of qualitative investigation. There are many methodologies and methods that support rigorous qualitative research; however, there are as many controversies about the "right way" to do qualitative research and how qualitative research should be defined and judged. In this space, it is impossible to do more than provide a general overview of study design, instruments for qualitative data collection and introductory guidance for analytic processes. Many of the references cited provide excellent examples of rigorous qualitative work in the medical literature and will expose the reader to multiple options for future study design and execution.

Imagine you have 10 blind men who know nothing about elephants. You place them in a circle around an elephant and ask them to briefly examine it with their hands. If you use quantitative analysis to synthesize their sampling you might conclude that, in general, the elephant has rough, hard skin and short, spiky hair. This conclusion about the elephant would be accurate but also, incomplete. If instead, you asked two blind men to examine the elephant's trunk for an extended period of time, after a few hours they might be able to tell you that the elephant has a long and unusual appendage. This appendage can pick things up off the ground, blow water out the

Not everything that can be counted counts. Not everything that counts can be counted.
—William Bruce Cameron, "Informal Sociology: A Casual Introduction to Sociological Thinking"

D. M. Elfenbein
University of California, Irvine, Irvine, CA, USA

University of Wisconsin, Madison, WI, USA

M. L. Schwarze (✉)
University of Wisconsin, Madison, WI, USA
e-mail: Schwarze@surgery.wisc.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_21

end and explore the world around it. Qualitative analysis of these new data would also lead to an accurate description of the elephant, one that would resonate with an outside observer who has actually seen an elephant. However, the observation would not be generalizable, i.e. you could not conclude that each man sampling the elephant would observe a trunk, nor would it be accurate to say that the elephant is covered with trunks. Nonetheless, qualitative analysis tells us something that is quintessentially important for description of an elephant.

In this chapter we introduce the technique of qualitative analysis. For health services researchers, qualitative methods provide an essential adjunct for many quantitative endeavors and have robust power as a stand-alone methodology provided the study design and execution are performed with rigor.

21.1 When to Use Qualitative Analysis

While quantitative analysis may provide a mile-high or bird's eye view of the population being studied, qualitative analysis starts from the ground and moves upward. Qualitative analysis is ideally suited for examining processes or interactions between people within a specific context, for example doctors, nurses, and technicians in an operating room [1] and can be particularly helpful in identifying subtle and critical distinctions that are not appreciable using quantitative analysis. Qualitative analysis is useful for describing social constructs for example, the (now outdated) taboo against disclosure of medical errors to patients [2] and can be instrumental for outcomes research given its power to identify latent or non-obvious processes for example problems at high or low performing institutions [3]. Qualitative analysis can also be used to understand language. This is called discourse analysis, where the language used to describe certain phenomena is the unit of study, instead of the phenomena itself [4]. While quantitative researchers tend to cast their findings as being objective, the reality is that even cold, hard numbers are described using language that has been chosen by the researchers. Language reflects the society in which the research was performed and shapes the way the data is interpreted and that language has consequences beyond just the numbers. Thus, qualitative analysis is critical to health services research as it helps us to know what "things" we are talking about when we are counting things. Finally, qualitative analysis is a critical tool for policy creation and evaluation as it enables investigators to examine perspectives and interactions among different stakeholders [5] and it is vitally important to understand implementation of new policies and procedures [6].

Qualitative research can be hypothesis generating. It is a good method to start with when your question is "What is going on here?" as it allows the investigator to be open to theories or constructs that arise from the data as opposed to using the data to test a preexisting theory or hypothesis. The flexibility of the methodology helps the researcher avoid the problem of seeing only what they are looking for [7]. Finally, many investigators will use qualitative methods as a first step for survey design, both to identify important questions to ask respondents and to insure the internal validity of survey questions through the use of cognitive interviews [8, 9].

21.2 Formulating a Research Question

Developing a concise, important and feasible research question is a challenge for all investigations and is particularly important for a qualitative study. To start, the investigator must acknowledge their theoretical assumptions and use these assumptions to focus the boundaries around the case to be sampled. In order to study a population or phenomenon in depth, the sample size for a qualitative study is, by necessity, typically small. As such the investigator is confined to study only *some* actors, in *some* contexts dealing with *some* decisions [10]. To define these boundaries, the investigator posits their theoretical assumptions outright in order to determine the case to be studied, aspects in which variability is desired and dimensions where homogeneity is important.

In a quantitative investigation, it is critical that the sample is representative of the population studied in order to generalize the results. As such, the sampling mechanism for a diverse population should ensure inclusion of a range of ages, socioeconomic status, racial and cultural backgrounds (if this diversity is present in the population). For a qualitative investigation, the goal is not to achieve generalizability but rather to capture the phenomenon as it exists at a certain point in time for a particular group (further investigation may or may not demonstrate variability between groups). As such, the investigator needs to explicitly state their theoretical assumptions up front. For example they might state either that they do not believe age, race, or socioeconomic status will have an appreciable effect on the results (and give supporting evidence) or state that these effects are unknown but, due to study constraints demographic differences are not the focus of the present study but will be critical to evaluate in the future. In essence, the investigator is not ignoring or excluding a specific segment of the population, but is explicitly acknowledging the choices made as well as the assumptions behind these choices in order to develop an appropriate study design to answer a specific and discrete research question.

Although boundaries are described from the start, because much of the research strategy and sampling methods (often called theoretical sampling) are grounded in the data, these boundaries need to be flexible or emerging [10]. This has led to the pejorative characterization of qualitative research as “make it up as you go along research” because this strategy is distinctly different than those used for quantitative investigations. However, the iterative process involved allows the investigator to shift the sampling frame in order to follow and expand important findings as they emerge from the data. Some examples of questions that are ideally suited for qualitative investigation include, “What processes are used to decrease mortality from gastric bypass in high performing centers?”, “What are the drivers of robotic surgery?” and “How do policies to improve access to general surgery in underserved populations impact different stakeholder groups?”

21.3 Sampling Strategy

Once the research question is determined, the next step is to select a sampling strategy that reflects the theoretical assumptions and enables the desired analysis. Typically this is called purposeful sampling where the selected respondents or observation units (hospital wards, operating rooms, texts [4], critical documents [11]) are specifically chosen to reflect the case that you desire to study. The sample is usually small so that it can be studied in depth and typically, the investigator does not set a predetermined sample size. Instead, respondents are sampled until the investigation reaches theoretical saturation, a point in the analysis where the themes and trends encountered occur with a degree of regularity. This can pose problems with the IRB, grant solicitation, and for study staff. To avoid this, investigators often generate an estimate of their sample size but should offer a large enough range to allow sampling beyond initial estimates if needed. In addition, researchers more familiar with quantitative methods may bristle when, in a qualitative study, it is necessary for the analysis to proceed before all of the data has been collected. This step is required to determine whether additional respondents are needed, to allow for interrogation of unanticipated results emerging in the data and to determine whether theoretical saturation has been achieved.

Because the sampling strategy is theory driven, it is important to state clearly the reasoning behind the design of the sampling method used as the rigor of the study will be judged on whether there is alignment of the data collection process with the study purpose. There is a large number of sampling strategies, well described in Chap. 2 of "Qualitative Analysis" by Miles and Huberman [10]. These strategies include maximum variation sampling where respondents are selected to include high variability in order to identify common patterns, contrasting case sampling where respondents or units are analyzed against each other in order to demonstrate differences [3] and snowball sampling where respondents with a unique or distinct trait are used to identify subsequent respondents for in depth investigation of an atypical point of view or phenomenon [12].

21.4 Structured Vs. Unstructured Data Collection

In addition to selecting a sampling strategy, the investigator will need to select an approach for data collection. The approach can range from a highly structured instrument using open-ended interviewing to a completely unstructured method, for example participant observation.

21.4.1 Focus Groups

One example of a highly structured approach is the use of focus groups. Focus groups, like market research, are ideally suited to obtain feedback on actual practices or proposed interventions [13]. Respondents are chosen to meet specific characteristics, frequently homogeneous on some levels and heterogeneous on others, and are studied in a group to capture important interactions between respondents. The groups are typically small enough for all participants to become engaged (range 4–12) and investigators will typically use more than one focus group per investigation. The focus group is formally moderated and carefully scripted with predetermined questions. A helpful reference for focus group design is "Focus Groups: A Practical Guide for Applied Research" by Richard Kruger and Mary Anne Casey (2009) [14].

21.4.2 Open-Ended Interviews

Open-ended interviewing is a less structured approach that still retains a large degree of structure. Although the interviewing process should be iterative, the process typically starts with a pre-determined interview guide. The investigator designs open-ended questions with care to avoid questions for which a yes or no answer would be possible. Rather than asking the respondent, "Can you tell me why you are having surgery?" (A question that can easily be answered with a "no"), a better example of an open-ended question is, "Tell me the story about how you came to have surgery."

Instead of providing an "interview script" the pre-designed questions serve as a starting point and suggested probes are supplied in order to direct the respondent to the salient issues. Analysis should be ongoing with data collection so that during analysis, researchers can provide feedback to the interviewer to ensure that they will flush out important themes or concepts in subsequent interviews. This iterative process allows for investigation of unexpected results and enables the investigator to explore concepts or themes in great depth.

It is critical that the interviewer is well trained in this type of questioning and is intimately familiar with the research question and relevant background information. It is frequently suggested that the principal investigator perform interviews because their background and understanding will have a significant impact on the direction of the interview. This may not always be possible given time constraints. At times, this also may not be desirable if respondents are familiar with the investigator or their background and are likely to provide socially desirable answers. For example, it is often not desirable for a physician to interview patients, as patients (and physicians) come to a patient-doctor relationship (even a non-clinical one) with their own set of beliefs and norms about how to talk to physicians and what physicians might want to hear. Not surprisingly, evolving interview questions are problematic for the institutional review board. Rather than providing the IRB with a script that will be read verbatim, it can be more effective to present the IRB with a list of question domains and sample questions with follow up probes for each domain. This will help to avoid returning to the IRB after each iteration for approval of a new line of questioning.

A helpful tool to consider in interviews or focus groups is the use of a vignette. Although it is challenging to design a clinical vignette that captures all of the complexities of clinical decision making, the presentation of a narrative account of a specific case can prompt a more instinctive and less abstract answer from the respondent. An additional helpful tool is the use of deliberative debriefing. Oftentimes, respondents will have socially created notions of reality that are not based in fact or clinical data, or the construct under consideration is complex and requires precision in description so that all respondents start from the same definition of terms. Examples of deliberative debriefing include a focus group study where female respondents were debriefed about the impact of breast cancer screening [15] and an interview study where family members of people who were unable to donate organs after circulatory death (DCD) were debriefed on the difference between donation after brain death and DCD [16].

21.4.3 Directed Observation

For directed observation, the researcher is embedded in the study environment but directed to study specific elements or constructs within the environment. Explicit acknowledgement of the investigator's theoretical assumptions upfront is critical for this type of investigation. Typically the researcher will perform prolonged observation of processes or events but enter the field with a list of predetermined elements to focus the observation. This has the advantage of facilitating data accrual rapidly with the cost of missing an important issue or construct because it was not recognized *a priori*. This is a useful approach when the unit of observation is documents, published work, videos or other media as is typical for discourse analysis [17]. A helpful reference for this type of qualitative research is, "Discourse as Data: A guide for Analysis" by Margaret Wetherell, Stephanie Taylor and Simeon Yates (2001) [18].

21.4.4 Ethnography

The least structured method for qualitative analysis is prolonged observation of participants in the field. This enables the investigator to share the daily environment of the study subjects including social interactions, language and habitual activities which allows for a rich description of processes and constructs. This is often described as ethnography (note: ethnography also refers to a qualitative methodology, not a method, so this can be confusing for those with more qualitative experience) and is a method commonly used by anthropologists and sociologists. The method is extremely time-intensive as the investigator is literally inserted into the daily routine of the population they are studying for a prolonged period of time, often on the order of months to years. The researcher is able to observe actions and counter-actions rather than simply eliciting the respondent's perspective about what they might do in a specific situation. A well-known ethnography about surgeons is "Forgive and Remember" by Charles Bosk [19]. For this research, Bosk, a sociologist, embedded himself on the surgical service at the University of Chicago Hospital for 18 months. The resulting text is a powerful description of the customs and practices that govern surgical care and are determined by surgical training. The rituals, culture and normative behavior described will resonate with anyone who has experienced surgical residency. For a shorter reference, another well-done example of ethnography from the medical literature is Joan Cassel's "Surgeons, intensivists, and the covenant of care: Administrative models and values affecting care at the end of life" [20].

21.4.5 Cognitive Interviewing

Survey studies require special mention at this point. While quantitative data from a survey are often published in the medical literature, readers should be aware that how the survey was designed and the language used within survey questions is of critical importance. Qualitative methods behind survey design are often not well described and are frequently omitted, and this has massive influence on the quality of the response and our ability to interpret the findings. While the numbers from a survey can appear valid, reliable, and generalizable, if the survey design did not use high-quality cognitive interviews to design the questions, the results may be problematic. The words and language matter and poorly designed surveys that do not ensure shared understanding of language within the line of questioning can lead to claims that may not be true. For example, from 2008 to 2014 researchers performed survey studies to assess resident confidence upon completion of surgical training after new duty hour rules, yet these studies failed to use a shared definition of the theoretical construct "confidence" they were measuring. Although many authors found similar results, specifically surgical trainees who reported "low confidence," the lack of rigor in survey development allowed this group of researchers to create "evidence" for a constructed claim [21]. Strong qualitative work to both develop the line of survey questions coupled with cognitive interviewing of a small sample of respondents, where respondents are asked to "think out loud" to describe what each survey question means to them, can avoid these deep pitfalls with serious consequences.

21.5 Analysis

After defining the study population and study design, the next step is to analyze the data as it is collected. This process is particularly noxious to those familiar with quantitative methods but it is critical for a robust qualitative study as the analysis is used to feedback into data collection to ensure that important themes and trends are examined in depth. For most qualitative investigations this next step requires coding the data. To this end, the investigator (or a team of investigators) will examine transcribed notes, transcripts of audio recordings, audio recordings themselves, texts, video tapes or other media and code snippets of the data as events, processes, ideas, or concepts appear. The coding can either proceed deductively or inductively. For deductive coding, a specific theory is used to analyze the data. For example, the investigator might use the "theory of clinical inertia" [22] to analyze why surgeons fail to refer patients to high volume centers for pancreaticoduodenectomy or the "Input-transformation-output" model of healthcare professional performance [23] to analyze the structures and processes that enhance safety in the operating room.

Alternatively, the coding might proceed inductively, which helps to anchor the empirical structure of the study and is particularly useful when non-obvious or latent issues are suspected or not previously well described. To do this, the investigator will use a technique called constant comparison where each new code is iteratively tested against previous uses of the code to ensure that the use of the code is consistent across the data set. Although inductive purity is frequently difficult to achieve, inductive analysis has the distinct advantage of allowing the investigator to discover new theories or constructs. Ultimately, whether the coding proceeds inductively or deductively the process is used to develop a coding taxonomy that is used to code subsequent data as it is collected. This taxonomy is flexible and allows for inclusion of new codes as they emerge from the data, refinement of existing codes based on the in-process data analysis, and termination of a code if it cannot represent the phenomenon as it exists in the data. The ability to reformulate or refine the concept or data unit increases the accuracy of the coding scheme based on empirical data which has an advantage over forcing the data into predefined and rigid categories.

Coding can be performed by a single investigator but may be enhanced through the inclusion of multiple investigators, particularly if members of the coding group have different backgrounds. The use of multiple coders has two distinct advantages. First, it provides an immediate availability of different perspectives to derive meaning from the data. Second, the process of sorting through varied interpretations of the data can reveal assumptions based on each coder's background and allow the group to attend to biases throughout analysis. Investigators will use different strategies to manage the variability that may result from multiple coders. Some investigators will retain the diversity of coding as it may provide a critical signal for analysis of the data. Others will work to come to group consensus about the specific code, as the consensus building can be a gateway to higher level analysis. Finally, some authors will have coders to code all the data independently and report inter-rater reliability. Although some strategies might be more productive than others, it is important for the investigator to both chose the strategy that will enable them to answer the specific research question and clearly describe and justify the reasoning for this choice.

Next a second level of data analysis needs to occur, sometimes referred to as higher level analysis or axial coding (that complements the first step of open coding). This is the process of making sense of large volumes of data, drawing connections between concepts and processes and refining or developing theories and hypotheses. In contrast to quantitative analysis where the goal is to condense and reduce the data for presentation, the goal of qualitative analysis is to expand the data and develop ideas, theories and thematic constructs [24]. This process can be particularly challenging as the data is frequently voluminous and unwieldy and the techniques for analysis are not standardized and have been described in myriad ways by leaders in the field. Simple diagrams mapping interactions and relationships can be useful, provided the investigator can escape from linear thinking and diagram relationships and connections as they exist in the data rather than with a reductionist view. Miles and Huberman demonstrate multiple methods for higher level analysis. For example, a construct chart or matrix can be helpful tools to ensure maximal fit and faithful data representation [10].

Given the data produced in a qualitative investigation is unstructured and typically massive it can be daunting to manage. There are several commercially available programs designed to catalogue qualitative data that can assist the investigator with organization. Programs such as NVIVO (QSR International–Melbourne, Australia) have the capacity to maintain many different types of primary data from simple word documents to video and enable the user to record and arrange the coded data. Other programs include Atlas.ti (ATLAS.ti Scientific Software Development GmbH–Berlin, Germany), Dedoose (SocioCultural Research Consultants–Manhattan Beach, California), and QDA Miner (Provalis Research–Montreal, Canada). Microsoft Excel (Microsoft Corporation–Redmond, Washington) can also be used to catalogue data but has less capacity to compare coded data elements than these other programs. The computer programs do not actually perform the analysis, but they can certainly retain the data and associated codes in a manner that allows for future retrieval and higher level analysis.

21.6 Ensuring Rigor

The standards for qualitative research are less familiar to most investigators and readers of the surgical literature. Although some may bristle at a direct comparison of the standards for quantitative methods because the standards are not perfectly analogous, this structure may prove a useful introduction for judging qualitative research. Rigorous quantitative research has internal validity, external validity, reliability and generalizability. In turn, though not perfectly in parallel, qualitative research should be judged by its credibility, dependability, confirmability and transferability [25]. Although qualitative researchers have developed a checklist for standardized reporting of qualitative research, similar to CONSORT for clinical trials, called "Consolidated Criteria for Reporting Qualitative Research" (COREQ) [26], use of this checklist is controversial [27] in respect to its overly dogmatic prescription of study processes and procedures, which when followed do not by themselves confer rigor [28].

21.6.1 Credibility

Credibility refers to the internal consistency of the research wherein the prolonged exposure of the investigator to study participants or subjects allows for a thick and rich description that attends to culture, context and setting. Credibility can be enhanced by reflexivity which is the process of stating one's positions and biases upfront. Another technique common in qualitative research is called member checking where, after analysis, the investigator returns to the study participants with the results to see if the analysis rings familiar with them [29].

21.6.2 Dependability

Dependability is enhanced by a clear and in-depth description of the processes and design choices used throughout the study. Many qualitative researchers will refer to an “audit trail” that enables the reader to fully understand the investigator’s steps and assess the validity of the conclusions based on the choices made and the procedures presented. This includes a clear statement about the investigator’s theoretical assumptions, a robust description of the iterative processes that influenced data collection and a detailed explanation of the coding process and higher level analysis.

21.6.3 Confirmability

Confirmability speaks to bias and perspective with respect to the investigator. Although such biases exist in quantitative analysis (for example, the goal is typically to confirm the hypothesis tested through rejection of the null) these biases are not often explicitly presented to the audience. In contrast, qualitative researchers explicitly state their biases upfront and the steps taken to manage these biases. One commonly used technique is triangulation. This refers to incorporation of multiple perspectives in order to describe the studied phenomenon or resultant theory as objectively as possible. To triangulate, the investigator may use multiple different frames of reference to gather data on the study population. For example, Bradley and colleagues interviewed physicians, quality managers and administrators from hospitals with both high and low use of beta-blockers post-myocardial infarction [3]. This design enabled the investigators to provide a rich description of the hospital culture that determined practice, a result that might have been missed or inaccurate had they interviewed only physicians. Another form of triangulation is to construct a study team for analysis that represents multiple perspectives. Although no one person can be expected to represent an entire group, a mix of professional and personal identities can enable the investigator to use multiple perspectives to interpret the data.

21.6.4 Transferability

Where the goal of quantitative research is to make statements that are generalizable, generalizability is not typically within the power of qualitative research. Instead, qualitative researchers aim for resonance. Although an assessment of resonance is left to the reader, the goal is for the investigator to present enough information about the context, processes and participants for the reader to judge how the results may transfer to other settings or domains [30]. While qualitative analysis does not aim to generate universal statements about populations, the characterization of specific behaviors, rituals, and actions and the conditions under which these occur, can illuminate and enlighten many health care practices.

One frustrating aspect of performing high quality qualitative research may be the challenges related to publishing these studies in high-impact journals. This is a problem that has worsened over time as some journals have explicitly stated they are not particularly interested in qualitative inquiry [31, 32] and others have been more implicit in their policy. Journal editors and reviewers frequently rely on heuristics related to large samples and small p-values without understanding the reliability of the variables included in the study or the lack of strength in the author's conclusions. In turn, this epistemological crisis of overconfidence in "big data" trickles down to assessments of qualitative research whereby reviewers note that the sample is too small, or not representative, and the findings are not generalizable. Although qualitative and quantitative research can both be done poorly, and journals should guard against publication of lousy research, investigators whose qualitative research is judged as "non-representative" or "not generalizable" should consider an appeal to the journal's editor, asking for reviewers with qualitative expertise.

21.7 Summary

Qualitative investigation is a powerful tool for health services researchers as it gets at processes, concepts, trends and constructs that are difficult to identify with quantitative methods. Although completion of a qualitative study is quite time consuming the results can have real impact in the field leading to new discoveries and improvements in surgical care.

References

1. Hu YY, Arriaga AF, et al. Protecting patients from an unsafe system: the etiology and recovery of intraoperative deviations in care. Ann Surg. 2012;256(2):203–10.
2. Gallagher TH, Waterman AD, et al. Patients' and physicians' attitudes regarding the disclosure of medical errors. JAMA. 2003;289(8):1001–7.
3. Bradley EH, Holmboe ES, et al. A qualitative study of increasing beta-blocker use after myocardial infarction: why do some hospitals succeed? JAMA. 2001;285(20):2604–11.
4. Neuman MD, Bosk CL. What we talk about when we talk about risk: refining surgery's hazards in medical thought. Milbank Q. 2012;90(1):135–59.
5. Robinson JC, Casalino LP. Vertical integration and organizational networks in health care. Health Aff (Millwood). 1996;15(1):7–22.
6. Giori NJ, Amanatullah DF, et al. Risk reduction compared with access to care: quantifying the trade-off of enforcing a body mass index eligibility criterion for joint replacement. J Bone Joint Surg Am. 2018;100(7):539–45.
7. Johnson RB, Onwuegbuzie AJ. Mixed methods research: a research paradigm whose time has come. Educ Res. 2004;33(7):14–26.
8. Lee CN, Hultman CS, et al. What are patients' goals and concerns about breast reconstruction after mastectomy? Ann Plast Surg. 2010;64(5):567–9.
9. Schwarze ML, Bradley CT, et al. Surgical "buy-in": the contractual relationship between surgeons and patients that influences decisions regarding life-supporting therapy. Crit Care Med. 2010;38(3):843–8.
10. Miles MB, H. A. Early steps in analysis. Qualitative data analysis. Thousand Oaks, CA: SAGE Publications; 1994. p. 50–89.
11. Steinman MA, Bero LA, et al. Narrative review: the promotion of gabapentin: an analysis of internal industry documents. Ann Intern Med. 2006;145(4):284–93.
12. Curlin FA, Dinner SN, et al. Of more than one mind: obstetrician-gynecologists' approaches to morally controversial decisions in sexual and reproductive healthcare. J Clin Ethics. 2008;19(1):11–21. discussion 22-13

13. Frosch DL, May SG, et al. Authoritarian physicians and patients' fear of being labeled 'difficult' among key obstacles to shared decision making. Health Aff (Millwood). 2012;31(5):1030–8.
14. Kruger, Richard, Casey, Mary Anne, Focus Groups: A Practical Guide for Applied Research, SAGE, Los Angles, ed. 2015.
15. Hersch J, Jansen J, et al. Women's views on overdiagnosis in breast cancer screening: a qualitative study. BMJ. 2013;346:f158.
16. Taylor LJ, Buffington A, et al. Harms of unsuccessful donation after circulatory death: an exploratory study. Am J Transplant. 2018;18(2):402–9.
17. Cabana MD, Rand CS, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282(15):1458–65.
18. Wetherell, Margaret, Taylor, Stephanie, Yates, Simeon, Discourse as Data, SAGE, Los Angeles, ed. 2001.
19. Bosk CL. Introduction. Forgive and remember. Chicago, IL: The University of Chicago Press; 1979. p. 2–34.
20. Cassell J, Buchman TG, et al. Surgeons, intensivists, and the covenant of care: administrative models and values affecting care at the end of life--updated. Crit Care Med. 2003;31(5):1551–7.
21. Elfenbein DM. Confidence crisis among general surgery residents: a systematic review and qualitative discourse analysis. JAMA Surg. 2016;151(12):1166–75.
22. O'Connor PJ, Sperl-Hillen JAM, Johnson PE, Rush WA, Biltz G. Clinical inertia and outpatient medical errors. In: Battles JB, Henriken K, Marks ES, Lewin DI, editors. Advances in patient safety: from research to implementation, Concepts and Methodology, vol. 2. Rockville, MD: Agency for Healthcare Research and Quality; 2005. p. 293–308.
23. Karsh BT, Holden RJ, et al. A human factors engineering paradigm for patient safety: designing to support the performance of the healthcare professional. Qual Saf Health Care. 2006;15(Suppl 1):i59–65.
24. Murphy E, Dingwall R, et al. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess. 1998;2(16):iii–x. 1–274
25. Malterud K. Qualitative research: standards, challenges, and guidelines. Lancet. 2001;358(9280):483–8.
26. Tong A, Craig J, et al. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349–57.
27. Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? BMJ. 2001;322(7294):1115–7.
28. Eakin JM, Mykhalovskiy E. Reframing the evaluation of qualitative health research: reflections on a review of appraisal guidelines in the health sciences. J Eval Clin Pract. 2003;9(2):187–94.
29. Marrow SL. Quality and trustworthiness in qualitative research in counseling psychology. J Couns Psychol. 2005;52(2):250–60.
30. Kuper A, Lingard L, et al. Critically appraising qualitative research. BMJ. 2008;337:a1035.
31. Greenhalgh T, Annandale E, et al. An open letter to The BMJ editors on qualitative research. BMJ. 2016;352:i563.
32. Loder E, Groves T, et al. Qualitative research and The BMJ. BMJ. 2016;352:i641.

Part IV

Career Development

Writing Scientific Manuscripts

22

Andrew M. Ibrahim and Amir A. Ghaferi

Abstract

One of the key pieces of academic currency is a manuscript. It is how academic surgeons communicate their craft to others in the field. There are numerous forms of manuscripts including systematic reviews, meta-analyses, commentaries, editorials, and primary research papers. Writing not only permits you to disseminate your findings through peer-reviewed journals, but it also provides you the opportunity to critically review the literature, evaluate your results, and place your findings into the greater context of your scientific area. While oral presentations are no doubt important, the written word remains a document that can be referenced and reviewed for many years to come. This chapter will address how to approach scientific writing through a systematic and strategic approach to deliver a clear and concise product.

22.1 Why Is Writing Critical in Academic Surgery?

There are many reasons why writing is critical to not only success in academic surgery, but also a vital form of making an impact on your field of research. First, Publishing in high impact journals, such as *New England Journal of Medicine* or the *Journal of the American Medical Association*, can lead to significant changes in

A. M. Ibrahim · A. A. Ghaferi (✉)
Institute for Healthcare Policy and Innovation, University of Michigan,
Ann Arbor, MI, USA
e-mail: aghaferi@med.umich.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_22

clinical practice or policy. The key reason these journals maintain high impact factors and are broadly read is they value not only excellent science, but also the ability to convey a clear narrative that allows a diverse audience to understand and potentially apply the new knowledge. However, writing clear, concise, and compelling manuscripts is not reserved for high impact journals. Maintaining a structured and disciplined approach to writing effectively will increase the reach of your work regardless of where it is ultimately published. Second, publication deficiency can be a barrier to promotion and tenure. The number and quality of publications are a significant measure of academic productivity. Finally, writing regularly allows you to crystallize your thoughts and new ideas. While this chapter is about writing manuscripts, these methods described below will apply to grants, policy briefs, blogs, and any other written communication about your work.

22.2 Where Do I start?

Every prolific manuscript started with a blank page. Don't let that intimidate you. Finding the time to write can be extremely difficult given the numerous demands on your time and the innumerable distractions (i.e., email, social media, clinical demands, etc.) In order to be successful at writing, it must be scheduled into your day to day just like your clinic, operating days, and teaching or administrative obligations. Some say that you should write for at least 1 h a day. While this is a laudable goal, don't beat yourself up if you don't hit the mark. However, keep tabs on when you are writing and reward yourself when you've been successful. You don't need entire days of writing blocked off on your calendar. It can be very difficult to maintain concentration for such a prolonged period of time. The majority of your manuscript can be written in 20–30 min chunks of time. Setting yourself up for success with these short bursts requires pre-planning with an effective outline of your manuscript. The remaining sections of this chapter will outline the contents of a manuscript and the key components of each section. We would encourage you use this outline as a template to plot out the topic sentences or ideas of each paragraph. This will allow you to then write in small, efficient, and effective blocks of time. Also, don't overly self-edit during the early writing process. Allow your thoughts to flow to the page and reserve the editing for later. The easiest way to overcome a blank page is to just start with something. Believe it or not, the rest will quickly follow. You will find yourself with too much written and reducing the contents will be your new "problem."

22.3 The Components of a Manuscript

Scientific manuscripts typically have the following components—Title, Abstract, Introduction, Methods, Results, Discussion, Tables/Figures, References. Some of these sections may have slightly different names or formatting preferences based on the journal. It is important to review the *Instructions to Authors* that is on every journals website before submitting (and preferably before you start writing). It does not reflect well on the authorship team if the manuscript is formatted incorrectly and will likely delay review.

The general guidance for writing we present draws from a seminal paper by Gil Welch—"Preparing Manuscripts for submission to Medical Journals: The Paper Trail" [1]. The systematic approach laid forth by Dr. Welch has been used by us and our colleagues to develop and refine our writing process.

The order in which you compose your manuscript does not need to prescriptive. You will develop your own style and preference for this. Some authors find it useful to sketch out the *Introduction* very early in the process which helps solidify the motivation and context for the study. The Methods can be a running, live document while you are conducting the study to avoid forgetting about nuances to your analysis. Finally, creating blank tables and figures for what you aim to display in your study helps you and your team focus the analytic effort.

22.3.1 Abstract

The abstract is intended to provide a brief summary of the manuscript. It often follows the same format as the article (i.e., Introduction, Methods, Results, Conclusion). Although your study may have multiple findings, the abstract should only focus on one or two key takeaways. Therefore, you should decide what are the most important. Ensure your introduction, methods, results and conclusion are consistent *within* your abstract. For example, your paper may examine multiple outcomes (e.g. complications, mortality, costs) but you only plan to focus on mortality and costs in the abstract. Your introduction, methods, results and conclusion should all be tailored to those two outcomes. Also, ensure that the components in your abstract are consistent with the manuscript. The abstract may have been written early for a meeting submission and the final manuscript results in changes to the methods or results.

Earlier, we mentioned that there is no correct order to writing the components of your manuscript. However, we like to construct abstract first to help hone the research question. This can be done with placeholders for results. Seeing what you want your final product to look like can help clarify your outcomes of interest and guide your study design.

The abstract is often used by editors and reviewers as an initial assessment of the manuscript. Specifically, journal editors may use the abstract as a screening tool to determine whether peer-review should be initiated. Further, even after publication, readers will use the abstract to determine whether they should read the rest of the manuscript. This is an opportunity to entire readers to read further.

22.3.2 Introduction

The Introduction serves to generate interest in the manuscript by giving context to the research question, identifying a significant knowledge gap, and give an overview of the study design or plan.

The *first paragraph* of the introduction motivates the research question by providing the reader with the necessary background information to understand why the topic is important. The first sentence is especially important to create excitement about the work/words to follow. For example, if the paper is evaluating variation in mortality rates across hospitals, you need to let the reader know why this is an important public health or patient safety problem. Finding the correct "altitude" for this information is key. Avoid giving information that is too broad or too narrow.

The *second paragraph* identifies the specific problem or knowledge gap that your study needs to address. This is where you highlight a couple major themes or examples from the existing literature that highlight the uncertainty in the field that your study question answers. This is where a good review of the literature provides you with the ability to describe these known and unknowns. The most difficult part of this paragraph is creating a narrative that makes your study question interesting and worth answering. You should highlight deficiencies in the literature that fit into the strengths of your study. For example, if the existing literature is entirely using administrative data and you have access to a clinically granular registry, then you should highlight that in this paragraph. Do not raise multiple knowledge gaps that are *not* the focus of your study, or your reader will feel misled!

The third paragraph gives the reader a succinct preview of the methods undertaken in your study. You don't need to go into too much detail because the Methods section will do that for you. However, it gives the reader an overview of the data you used and the outcomes of interest. This should closely parallel the knowledge gaps you established in the second paragraph. Lastly, you should conclude the introduction with a hypothesis statement or study aim.

22.3.3 Methods

The methods section provides a step by step recipe for how the study was executed. There is no one correct way to structure this section. Some journals use standard headers and you should review the *Instructions for Authors* or previously published articles to get a sense of their preferred structure.

It is helpful to the reader to use sub-headers in this section. The most common sub-headers are *Data Source*, *Study Population*, *Outcomes, Statistical Analysis*. Devote at least 1–2 paragraphs for each of these sections. There may be significant similarities to other work published by your group. Don't reinvent the wheel. Look over those manuscripts and place that prose in your own words with a bend toward the specifics of your study. It is reasonable to reference other papers where complex methods have been previously outlined. It is important to give the reader enough detail that they could reasonable replicate your study. At a minimum, be sure to define your primary outcomes if there are not standard definitions established already. Reviewers are looking for a sound study design and ruling out sources of bias—so include inclusion/exclusion criteria, consideration of sample sizes/power calculations, etc. If human subjects are involved, a statement of institutional review board approval or exemption is expected here, as well.

22.3.4 Results

The Results section should provide the output from your analyses but should also be concise and direct. This section is typically 3–4 paragraphs long and should start with a general description of the study population that references your "Table 1". The next 2–3 paragraphs will review the primary and secondary outcomes of your study. This section of the manuscript is not where you provide any editorialization about the results. Everything should be presented objectively and without modifiers like "remarkably" or "interestingly." It is best to present the outcomes in the same order you structured your methods section. You should reference all tables and figures in this section. Do not simply rehash every table and figure in words. Let those items stand on their own as well. Simply introduce them and give a couple highlights about them. Be sure to avoid having a table or figure that is not referenced anywhere in the text. Even the best visual will need some sort of explanation.

22.3.5 Discussion

The discussion section allows you to bring the entire story you have been building all together. This involves linking the information you presented in the introduction, methods, and results and providing the appropriate context and future directions for your work. We have a general framework for how to structure the discussion that allows you to summarize the work and guide the reader about how to place the findings in perspective.

The *first paragraph* should summarize the main findings of the paper. Choose the 1–2 key findings the reader should take away from the results section. When we read a paper, we often read the abstract and the first paragraph of the discussion to quickly get to the authors punchline.

The *second and third paragraphs* should now place the study's main findings into context. This is where you can expand upon some of the concepts and knowledge gaps that you referenced in the introduction. You now have more space to dig deeper and explain why your study has moved the field forward. Be sure to review the literature and include seminal papers germane to your study. Give a fair review of those papers and explain why your study adds to what is already known or unknown. If all you are doing is agreeing with previous work, then you may want to rethink your research question. This is an important area to contrast your work and justify what it worth pursuing.

The fourth paragraph is the limitations paragraph. Every study has some limitations, so do not skip on this. (Many authors do, and reviewers absolutely hate it!) This also give you an opportunity to explain to the reader or reviewer that you've thought this through and acknowledge some limitations. You can also preempt some reviewer comments in this paragraph. Most health services researchers are interested in understanding what about your study threatens the validity of the findings. Plan to present at least three and up to five limitations. The most common categories of limitations are chance, bias, and confounding. Indicate how your study may suffer from each of these *but take the opportunity to provide a counter point as to how you mitigated that limitation* through your study design, data source, or other method.

The final 1–2 paragraphs are your chance to discuss the implications of the study's findings. The implication should not simply be that more research is needed. Of course more research is needed, but what can readers do with your results today. How does it move the field of health services research forward for your area of interest? Give this section some thought and provide your thoughts on the real-world implications of the work. (Note: when journalist want to create a press release or write about your work, this is often where they look for story ideas.) However, be careful not to overreach. Stay within the conclusions that can be reasonably drawn from your results given the limitations that you described in the previous paragraph. Most health services research will speak to one or more of the following stakeholders—patients, providers, payers, and policy makers.

22.3.6 Tables and Figures

Remember that most readers will start by reading the abstract and introduction, then move to the tables and figures to get a taste of the value of your study. Therefore, it is important to make sure these visuals are polished, clear, and organized. Take a look at papers published in high impact journals for good examples. Follow the instructions for the journal you are submitting to regarding the general formatting of the tables as well. Do not embed them in the middle of your text. They should be uploaded separately and will appear at the end of your manuscript. Make sure that you have followed the Instructions for the Authors for formatting guidelines and remember that most people reviewing your manuscript wear reading glasses (use a large font!).

22.3.7 References

The references should be included at the end of your manuscript and follow the recommended format for the journal to which you are submitting. Use the reference management software that you are most comfortable with. Build the reference database as you are conducting your review of the literature. Don't wait to build it as you are writing the manuscript. Most modern programs also allow you to embed the.pdf of the article which can come in handy when you are editing the paper, responding to reviewers, or writing your next manuscript.

22.4 Final Pearls

Scientific writing is not easy and can lead to frustration. You are not the only one who feels this way. To overcome feelings of frustrations or "writer's block" it is important to try to write a little something every day. Remember, there are pieces of the manuscript that can even be written without any data. You can skeletonize the abstract, formulate your introduction, mock up your tables and figures, and write the methods as you conduct your analyses. You should also seek feedback on your manuscript piece by piece. It can be difficult to hone the message in the introduction and discussion. Therefore, go through several iterations of these sections with your mentor. This entails setting up regular meetings and coming prepared. Listen to the feedback you've received in each meeting and put it to work. We find that audio recording (with your mentors' permission) these sessions can allow you to pay more attention to the meeting and have a productive dialogue without worrying about writing down everything. Go back and listen to that recording and internalize the feedback you receive. Then, make some changes. Your mentor or colleague wants your manuscript to get better. Don't be stubborn and believe that your written word is infallible. The better you become at applying the feedback you receive, the more quickly your writing will improve.

Reference

1. Welch HG. Preparing manuscripts for submission to medical journals: the paper trail. Eff Clin Pract. 1999;2(3):131–7.

Disseminating Research and #VisualAbstracts

23

Sarah A. Brownlee and Andrew M. Ibrahim

Abstract

A visual abstract is a visual representation of the key findings of a research manuscript or article. The first visual abstract was used by the journal *Annals of Surgery* in July of 2016 alongside the hashtag #VisualAbstract, and since then over 75 journals, organizations, and institutions have adopted their use as a way to increase dissemination of research findings. This chapter details the tools and steps needed to create a basic visual abstract, discusses ways to leverage the visual abstract for dissemination of research on social media, and explores new ways visual abstracts are being used to improve the reach and impact of research findings across medical and scientific disciplines.

23.1 Introduction to the Visual Abstract

A visual abstract is a visual summary of the information usually found within the abstract portion of an article. Much like the actual text abstract of an article, the visual abstract is meant to convey the key findings of the article in a shorter format. It is not intended to replace the article or serve as a substitute for reading the full text; rather, visual abstracts are meant to help readers quickly and efficiently identify content most relevant to them that they wish to engage with more fully. The success of this practice is evident in the dissemination metrics tracked since the inception of the visual abstract format.

S. A. Brownlee
Department of Surgery, Massachusetts General Hospital, Boston, MA, USA

A. M. Ibrahim (✉)
Department of Surgery, University of Michigan, Ann Arbor, MI, USA
e-mail: iandrew@med.umich.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_23

The first visual abstract was debuted by the Annals of Surgery in July 2016 and was posted to social media with the hashtag #VisualAbstract to increase dissemination and searchability. Though visual summaries of research articles, such as the “Central Illustration” in cardiology or the “Graphical Table of Contents” in basic science literature, have been used since the 1980s, the first visual abstract was unique in its use of a replicable triptych layout, clean single-color icons, and direct reporting of primary outcomes. Today, visual abstracts have evolved to include a variety of visual styles, usually distinct to one of the over 75 journals, organizations, and institutions that have adopted the visual abstract format.

Broader adoption of the visual abstract for dissemination of a range of scientific information has led to a variety of visual styles. However, key components of the visual abstract remain the same despite the visual style being used. These include:

- A summary of the key question being addressed, usually from the title of the article or heading of a key figure
- A summary of key outcomes
- An author and citation, usually including the first author’s name, journal, and year of publication
- A visual display of an outcome, such as a simple single-colored icon
- Data of the outcome, a numeric representation of the outcome with clear labeling of units
- Who created the visual abstract, which is often the journal but may be an individual author

23.2 Creating a Visual Abstract

The creation of an effective visual abstract requires an understanding of essential design principles, basic computer software, access to images or icons, and—most importantly—an appropriate research article. Below is a breakdown of each of these components, along with tips and tricks for maximizing each step of the visual abstract creation process and ways to avoid potential pitfalls.

23.2.1 Principles of Design Thinking

To create a strong visual abstract, it is important to respect design conventions including attention to contrast, repetition, alignment, and proximity. For example, to create hierarchy with information, use contrast techniques such as bold, color, and size. Following design rules will help create a consistent visual abstract that allows audiences to easily understand the key findings of the article. In addition, the design of a successful visual abstract is predicated on principles of effective design. Outlined below are some principles that are useful when designing and creating visual abstracts, regardless of subject matter of the article or abstract.

- *Focus on the user experience:* The first question to be asked before any design decisions are made is, "Who is the audience?". The answer will guide choices such as how much exposition of the subject matter is needed or how much space is devoted to describing research methods.
- *Clarity of purpose:* Regardless of audience, narrowing down the article to the key message is a crucial step. Aim to do this at multiple stages of the design process.
- *Rapid prototyping:* Often the first or second choice for a layout or icon will not be the best option. Remain flexible and create multiple versions of the visual abstract, paying attention to what works best about each version.
- *Iterative improvement:* In addition to rapid prototyping, seeking feedback is an important part of incrementally improving the visual abstract design. Focus less on making the abstract "perfect", and more on improving specific aspects of the abstract in a stepwise fashion.
- *Thoughtful restraint:* Depending on the article chosen for the visual abstract, it can be difficult to balance context with clarity of the key message. More is not always better, as too much detail can distract from the take home message of the study being presented. Choosing articles with a clear narrative (see 'Article Selection' below) can make this job easier.
- *Relevant creativity:* The visual abstract format invites creativity and following the above principles will often lead to new ideas that are worth piloting. However, flashiness or overstated design elements should not be used at the expense of clarity of purpose.

23.2.2 Getting Started with the Right Tools

Software: Online image creators and graphic design software options are plentiful. Expensive or high-powered software such as Adobe Photoshop, InDesign, or Illustrator certainly can be used for visual abstract creation, but most projects actually can be completed using Microsoft PowerPoint or Keynote. In keeping with the thoughtful restraint principle discussed above, one may be better to truly start with less sophisticated software.

Icons: The internet has multiple image banks that are free to use, e.g. Google Image search. If using such a bank for images, it is crucial to ensure that copyright permission is obtained prior to using any copyrighted images—this point cannot be overstated enough. As an alternative, a copyright filter can be used to search for images that are free to use without copyright limitations. For researchers who do a lot of visual presentations or will be making visual abstracts frequently, subscribing to an image or icon bank may be preferable (e.g. Noun Project, Shutterstock, Getty Images). Subscription banks offer access to higher quality images and icons that often do not require copyright attribution. There are a growing list of them (Table 23.1).

Article Selection: Overtime multiple templates have emerged to described all types of studies. However, when starting off and progressing on the visual abstract

Table 23.1 Growing list of online icon banks

Boxicons
EndlessIcons
Evericons
Feather
Flaticon
Iconfinger
Iconfont
Iconmonstr
Icons8
IconStore
Ikonate
Illustrio
Jam
MaterialDesignIcon
Noun Project
Octicons
SimpleIcons
SVGRepo
Webiconio Zondicons

Joy, Felix, "Felix Joy - Designer and Maker: Icons" Available online: https://www.felixjoy.co/designbase/icons Accessed 26 September 2019

learning curve, you'll want to choose more straight forward articles with these characteristics:

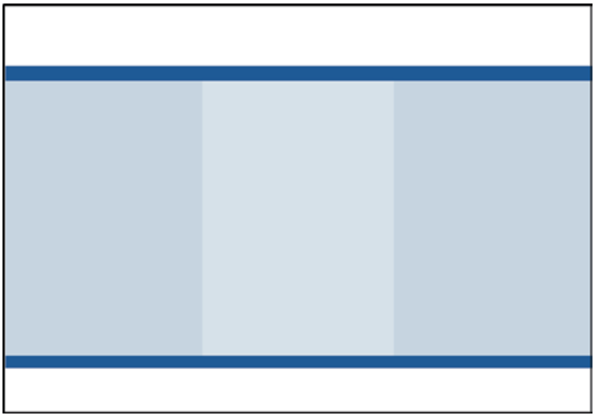
- *Well-known, easily understood study design*. Preferred study designs include (but are not limited to) randomized controlled trials, survey results, retrospective reviews, and qualitative data. It is possible to represent multiple results clearly in a visual abstract format, as long as the viewer can quickly understand the methods used to get those results. A good example of a preferred study design is randomized controlled trial of X (intervention) vs. Y (standard of care). Lesser known designs or complex types of statistical modeling may be compelling, but they are cumbersome to explain visually and take up too much space in the visual format. Many journal article formats that do not describe original investigations can also be adapted for a visual abstract, including commentary/ perspective pieces or committee recommendation reports.
- *Clear binary outcomes.* Choose articles with results reported in simple, binary terms, such as an increase or decrease in frequency, time, or cost with results reported in absolute differences. An example of clear binary outcomes is "mortality decreased from 22% to 17% after X intervention was performed." Absolute relationships are more digestible than something along the lines of "in our three-tier hierarchical model, the OR was 1.7 in favor of mortality". Historically, outcomes measured in odds ratios have been avoided altogether, since they are often misinterpreted even by article reviewers.
- *A story that's easy to interpret.* Pick the most clearly written paper so that the visual abstract is easy to keep in line with the original thrust of the research. A

clear story is most often found in the form of primary and secondary outcomes that go in the same direction. For example, “complications and mortality decreased following X intervention” is more easily understood than a case where one goes up and the other goes down. There isn’t enough room to include every discussion point in the visual abstract, so nuanced results are more likely to be lost or confused in the visual format. If there are doubts or confusion about the overarching message or take-home points from the article when read in its entirety the first time, the article is unlikely to translate well to a streamlined visual format.

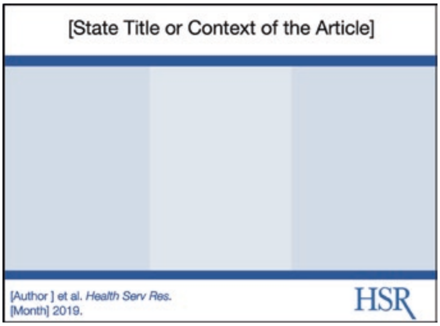
23.2.3 Step by Step Creation of a Visual Abstract

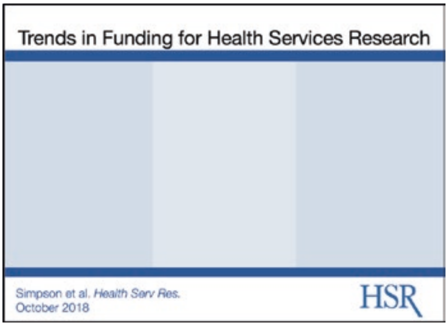
Once the software, icon bank, and article have been selected, the following steps can be followed to create a simple visual abstract:

1. *Create your visual fields*. There are multiple options for dividing the blank canvas into distinct visual sections, but colored boxes (usually one color in multiple shades) can be an easy way to quickly create distinction without adding unnecessary complexity.

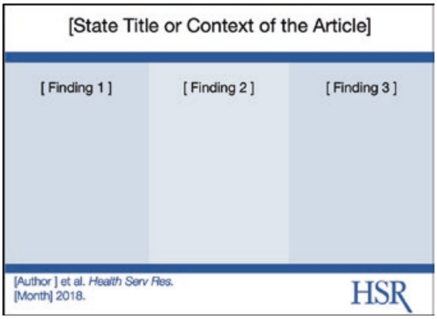


2. *Add author, journal, and title*. This essential information will add an important framework for the rest of the text.


[State Title or Context of the Article]
[Author] et al. *Health Serv Res.*
[Month] 2019.
HSR


Trends in Funding for Health Services Research
Simpson et al. *Health Serv Res.*
October 2018
HSR

3. *Lay out the outcomes*. One easy way to do this is to describe the outcomes across the top of the colored boxes in short comparative phrases (e.g. “increase in infection rates”).

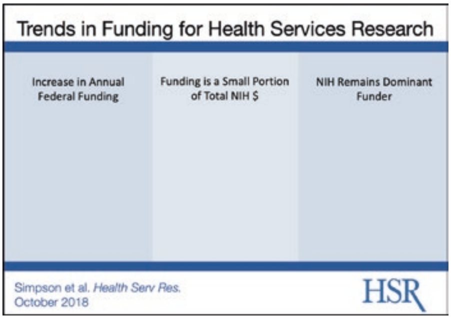


[State Title or Context of the Article]

[Finding 1] [Finding 2] [Finding 3]

[Author] et al. *Health Serv Res.*
[Month] 2018.

HSR



Trends in Funding for Health Services Research

Increase in Annual Federal Funding

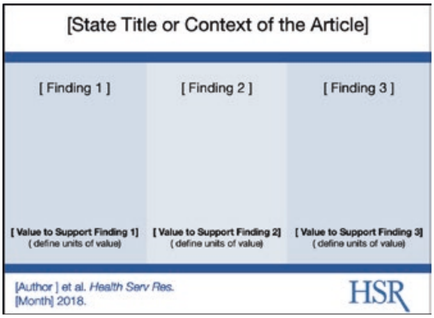
Funding is a Small Portion of Total NIH $

NIH Remains Dominant Funder

Simpson et al. *Health Serv Res.*
October 2018

HSR

4. *Add in data for each outcome*. Add the numeric value for each outcome, including the units.



[State Title or Context of the Article]

[Finding 1] [Finding 2] [Finding 3]

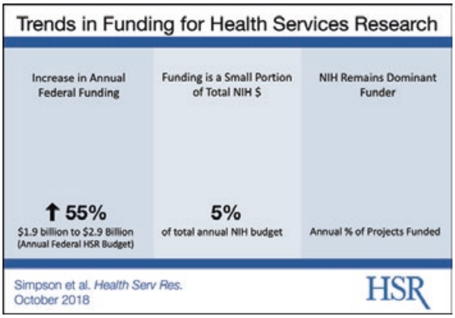
[Value to Support Finding 1]
(define units of value)

[Value to Support Finding 2]
(define units of value)

[Value to Support Finding 3]
(define units of value)

[Author] et al. *Health Serv Res.*
[Month] 2018.

HSR



Trends in Funding for Health Services Research

Increase in Annual Federal Funding

↑ 55%
$1.9 billion to $2.9 Billion
(Annual Federal HSR Budget)

Funding is a Small Portion of Total NIH $

5%
of total annual NIH budget

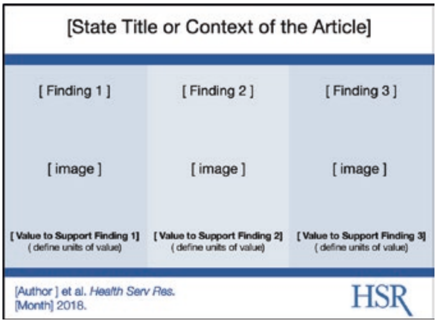
NIH Remains Dominant Funder

Annual % of Projects Funded

Simpson et al. *Health Serv Res.*
October 2018

HSR

5. *Add visuals*. This is often the most difficult step, and it is important to use the principles of Rapid Prototyping to trial multiple icons or images. Again, it is critical to make sure any permissions or rights to icons or images are obtained if needed.



[State Title or Context of the Article]

[Finding 1] [Finding 2] [Finding 3]

[image] [image] [image]

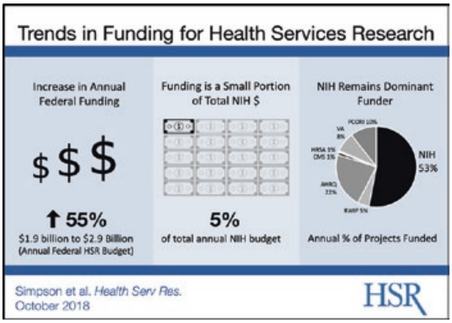
[Value to Support Finding 1]
(define units of value)

[Value to Support Finding 2]
(define units of value)

[Value to Support Finding 3]
(define units of value)

[Author] et al. *Health Serv Res.*
[Month] 2018.

HSR



Trends in Funding for Health Services Research

Increase in Annual Federal Funding

$ $ $

↑ 55%
$1.9 billion to $2.9 Billion
(Annual Federal HSR Budget)

Funding is a Small Portion of Total NIH $

5%
of total annual NIH budget

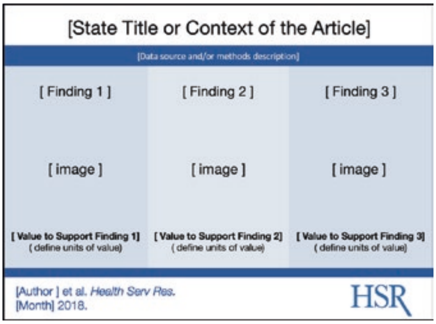
NIH Remains Dominant Funder

Annual % of Projects Funded

Simpson et al. *Health Serv Res.*
October 2018

HSR

6. *Add methods or data sources.* Depending on the methods or data sources used, it can be useful to include a brief description or listing somewhere near the top of the visual abstract.



[State Title or Context of the Article]

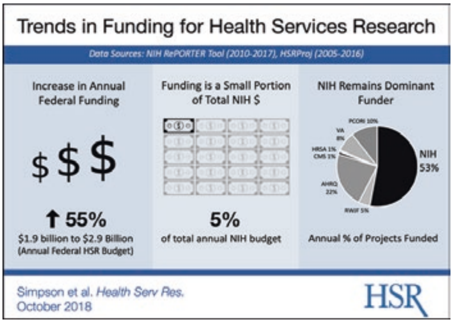
[Data source and/or methods description]

[Finding 1]
[image]
[Value to Support Finding 1]
(define units of value)

[Finding 2]
[image]
[Value to Support Finding 2]
(define units of value)

[Finding 3]
[image]
[Value to Support Finding 3]
(define units of value)

[Author] et al. *Health Serv Res.*
[Month] 2018.



Trends in Funding for Health Services Research

Data Sources: NIH RePORTER Tool (2010-2017), HSRProj (2005-2016)

Increase in Annual Federal Funding
↑ 55%
$1.9 billion to $2.9 Billion
(Annual Federal HSR Budget)

Funding is a Small Portion of Total NIH $
5%
of total annual NIH budget

NIH Remains Dominant Funder
Annual % of Projects Funded

Simpson et al. *Health Serv Res.*
October 2018

7. *Export the file.* Save the file as a JPEG or other image file so that it is easily usable for social media. Proofread and double check that all data used in the visual abstract are consistent with that of the original article.

23.2.4 Avoiding Visual Abstract Mistakes

Outlined below are some strategies to avoid common visual abstract mistakes and ensure that the final visual abstract is an accurate and engaging summary of the research article it represents.

- Whenever possible, use language directly from the article. This will limit any bias in from the interpretation of the article by the visual abstract creator.
- Clearly state the study's conclusion. The visual abstract should encompass the core message of the research, and including a concluding statement is a good way to ensure the authors' message is correctly represented.
- Include methods, data sources, or study design and size. These details can give important context to the viewer and usually can be done succinctly.

23.3 Leveraging the Visual Abstract for Dissemination of Research

Once the visual abstract has been created, the next step is to utilize it for increased dissemination of the research it depicts. This section includes information on the current evidence supporting increased dissemination with visual abstract use, steps for utilizing social media (especially Twitter) to increase the reach of an article through visual abstract and hashtag #VisualAbstract use, and guidelines for onboarding the visual abstract for use by a scientific journal.

23.3.1 Evidence of Increased Dissemination

From July to December of 2016, *Annals of Surgery* conducted a prospective, crossover study to evaluate the impact of visual abstract use on article dissemination [1]. A group of 44 articles published that same year in the journal were chosen for study. In the first phase, half of the articles were tweeted from the journal's Twitter account using the title only, and the other half were tweeted with the title and visual abstract. After a 4-week washout period, the groups were switched, and the articles were retweeted using the opposite protocol. The results demonstrated over a seven-fold increase in article impressions when tweeted alongside a visual abstract, as well over an eight-fold increase in retweets. Most significantly, there was nearly a threefold increase in visits to the full article on the publisher website when article titles were tweeted with a visual abstract.

23.3.2 Use of Social Media to Improve Dissemination

Multiple social media platforms can be used to disseminate research and take advantage of the visual abstract format for improved reach and impact. Facebook, Twitter, and LinkedIn are some of the major platforms that can be used to share visual abstracts, and each generally has its own benefits and drawbacks. The visual abstract itself, alongside the hashtag #VisualAbstract was debuted and has primarily been utilized on Twitter, which will be discussed in more detail below.

Once a visual abstract has been created and the article is live on the publisher's website, there are a few strategies that can be used to maximize the traffic and attention they receive. First, if the journal does not use visual abstracts regularly, it can be helpful to reach out and let them know one has been created. Often a dedicated public relations or social media team will be available to share the visual abstract through the journal's own accounts and social media networks. Second, tagging all of the study's authors and stakeholders along with their institutions will bring the tweet to their attention, and allow them to share it within their own networks. Third, it is worth reaching out to your own institution to determine if there is a public relations or social media team that would be willing to share the visual abstract and article on their own accounts to help amplify the message.

Finally, in order to track dissemination, it is important to understand the most frequently used outcomes for Twitter activity: impressions, retweets, and link clicks. Impressions are the number of times a tweet is seen on Twitter and can be thought of as analogous to the exposure of the visual abstract to a given audience. Retweets are the number of times a tweet is shared. These reflect how worthwhile an audience thinks it is to share the tweet with their network. Link clicks represent the number of times a link included in the tweet is clicked (often this will be a link to the full text of the article). These are the best estimate of engagement with the article in its entirety, and an increase in link clicks is often seen as the primary goal for the use of a visual abstract.

23.3.3 Journal Adoption of Visual Abstracts

Since the first use of a visual abstract by *Annals of Surgery* in 2016, over 75 journals and institutions have adopted the visual abstract for regular use in their article dissemination strategy. While each journal likely has a unique set of goals for their visual abstract use, there are some common strategies that can be used to ensure the success of a visual abstract launch:

- *Obtain author buy-in.* Before creating or publishing a visual abstract for an article that has been accepted for publication in the journal, it is important to ask permission from the authors. While the degree to which authors are involved in the creative process will vary, we suggest giving them the option for final approval to ensure their work is accurately represented in the visual format.
- *Start small*. The process of creating, editing, and approving a visual abstract for final use by a journal will take much longer in the initial stages. Because of this, it is best to start with a modest goal such as producing 1–2 abstracts per month. Plan to review the process and monitor the outcome metrics of these initial abstracts after they are piloted.
- *Article selection is key.* In the early stages, it is best to select articles that naturally lend themselves to visual abstract creation (see "Article Selection" above). Usually these articles have clear binary outcomes and straightforward methods. Establishing the success of the visual abstract for articles with these characteristics will make it easier to tackle more nuanced or complicated articles later on.
- *Identify roles.* A typical visual abstract team may consist of one visual abstract creator, 1–2 independent reviewers (often members of the editorial team familiar with the journal's subject matter but who do not have an individual stake in the articles published), and a member of the social media or communications team. It is important to identify which tasks each team member will be responsible for—for example, the visual abstract creator may produce an initial draft of the visual abstract, which is then reviewed by the independent reviewers for clarity, accuracy, and completeness. The social media or communications team member can provide helpful guidance about branding, logos, or particular images that the journal wants included.
- *Create a timeline.* Work with the editorial board to understand how the visual abstract creation process will fit in to the workflow of producing each issue. For example, establish a calendar that starts as soon as the article(s) are selected for visual abstract creation and has deadlines for each stage of the creation process (initial draft production, reviewer responses, input from social media or communications team, final draft).
- *Stay consistent.* Standardizing elements such as templates, fonts, logos, and color schemes is an important early step which serves the dual purpose of making the visual abstracts faster and easier to produce as well as more recognizable to the audience.

23.3.4 Who Should Make the Visual Abstract?

As mentioned above, anyone *can* make a visual abstract, but there are a number of pros and cons to consider when thinking of who *should* make a visual abstract for a given article (Table 23.2). In general, there are 3 groups of people who might consider creating a visual abstract: the author, the journal, or the audience.

Pros associated with author creation of the visual abstract may seem obvious. The author is generally most familiar with the subject matter, study design, and nuances of the article being published and will understand the overarching message the article is meant to convey. However, it is also possible the author may be too close to the article and may either overstate the impact of the conclusions or ignore methods or results that are either confusing or potentially misleading to the audience. For these reasons, it is a good idea for the visual abstract to go through a similar (though often abbreviated) peer-review process as the article itself.

Journal creation of the visual abstract similarly has benefits and drawbacks. Creation of a visual abstract by a journal often involves multiple people, as described above ("Journal Adoption of Visual Abstracts"), which confers the benefit of multiple perspectives. For example, the authors of the article have given permission for the visual abstract to be created, and are able to give important feedback, but the creation is often performed by a non-author, which helps limit any bias from the author. The drawback to this approach can be a "too many cooks in the kitchen" phenomenon where the overall message of the visual abstract gets lost in various perspectives and opinions about the details or design elements. To avoid this issue, it is important to clearly define the role of each person involved in the process.

Finally, the audience for the article may make a visual abstract. This approach also has advantages and disadvantages. Audience members have proven to be a source of new methods and ideas about how to create and use visual abstracts. Encouraging the creation of visual abstracts by readers who are not directly connected to the article is one of the best ways to continue such innovation. On the other hand, without consultation with the authors or those closer to the article, there is an increased risk of misinterpretation or misuse of the article's results or conclusions. This risk could be minimized by reaching out to the authors and/or journal before self-publishing a visual abstract on social media. The decision to do so should be made on a case-by-case basis.

Table 23.2 Who should make the visual abstract?

Creator	Pros	Cons
Authors	Most familiar with study design, methods, results, and conclusions	May be too close to the topic or have biases about the results
Journal	Process allows for input by multiple people with varied perspectives	Too many opinions can inhibit a clear visual message
Audience	Expand the context and utility of visual abstracts through innovative design	Risk of misinterpretation or misuse of the article's conclusions

23.4 Future of the Visual Abstract

With broad adoption of the visual abstract into scientific communication new applications have emerged into new audiences, new contexts and new purposes.

New Audiences. While visual abstracts were originally designed to spread research amongst researchers, the audience has broadened. The Centers for Disease Control, for example, has begun using the Visual Abstract for weekly public service announcements aimed at the general public. In line with that, it seems possible that visual abstracts could also be created to serve as patient education tools.

New Contexts. The original context of the visual abstract was post-publication dissemination of research articles by journals. The templates have been adopted into to more proximal outputs and now are frequent slide summaries of conference talks. In fact, Dr. Chelsea Harris, has pioneered the #LiveVisualAbstract that summarizes talks given at meetings in almost real time.

New Purpose. For all of us who do research, we know that dissemination of information is only the start of the story to improve our field. Ideally, we would also want to see that information translate into changes in practice. As the field of visual abstract creation matures, it now possible to envision how the visual abstract could service as a point of intervention. For example, a visual abstract with foley best practice guidelines may be placed on units with high urinary tract infection rates to see if practice patterns could be improved.

The future of visual abstracts are now well beyond any single journal or person. The movement has benefited from being an open source endeavor where best practices and lessons learned are openly shared. We hope you'll continue the same spirit.

An Open Source Visual Abstract Primer is updated regularly. To Download the more recent copy, visit: www.SurgeryRedesign.com/resources

Reference

1. Ibrahim AM, Lillemoe KD, Klingensmith ME, Dimick JB. Visual abstracts to disseminate research on social media. Ann Surg. 2017;266(6):e46–8. https://doi.org/10.1097/SLA.0000000000002277.

Engaging Policymakers to Disseminate Research

24

Babak J. Orandi and Jayme E. Locke

Abstract

Researchers have historically been divorced from the policy implications of their work. However, through a process that involves sound science, involvement of key stakeholders, a comprehensive policy strategy, and crafting a statement that resonates with policymakers and the lay public, researchers have the opportunity to significantly increase the impact of their work.

24.1 Introduction

Researchers have historically been divorced from the policy implications of their research. Researchers and policymakers have traditionally worked in silos, with little interaction and large cultural differences dividing the two groups. The reasons for this are manifold, including a legacy effect of a historical emphasis on basic science research, which tends to have fewer direct policy implications. Additionally, despite a plethora of research productivity metrics (e.g., h-index, m-value, number of publications, amount of grant funding, mentions on social media, etc.), none fully recognize the huge efforts that making policy changes require, rendering policy-making counter to the goal of academic promotion and the demands of clinical care [1, 2]. With the vast increase in health services and quality improvement research, many research endeavors have more clear, direct implications for the development of policy. Direct involvement with policy-makers is therefore a natural extension of these efforts; however, medical training does not typically provide sufficient training to disseminate research findings beyond the audience of scientific colleagues.

B. J. Orandi · J. E. Locke (✉)
Division of Transplant Surgery, Department of Surgery, University of Alabama at Birmingham, Birmingham, AL, USA
e-mail: jlocke@uabmc.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_24

Interacting with policy-makers and influencing policy is a complex process quite different from what is often viewed as the orderly scientific method. What follows will serve as a rough guide as to how to engage policymakers to effect change based on scientific research.

24.2 Sound Science

Naturally, the first step in promoting evidence-based policy is producing the evidence. The science must be sound, and of particular relevance, is its external validity. Having a highly homogeneous patient population in a clinical trial, for example, reduces confounding and more elegantly distinguishes cause and effect. However, a study lacking in external validity limits the ability of policymakers to extrapolate the findings to the population level, and may contribute to "slips between the cup of science and the lip of application." [3] In this regard, pragmatic clinical trials are especially useful, as they often have heterogeneous patient populations recruited from a variety of patient care settings, compare clinically relevant interventions, and measure a spectrum of meaningful health outcomes [4]. Ideally, the science is published in reputable, peer-reviewed journals, though it is generally rare that a single publication alone can significantly influence policymakers (or practitioners for that matter). More often, it is a body of work—with studies of varying degrees of internal and external validity—that drives policy changes. Formalized syntheses of the literature, such as those provided by the Cochrane Collaboration, can be especially useful for this purpose.

24.3 Stakeholders

In order to effect policy change, relevant stakeholders need to be involved. A coalition of stakeholders provides a broader reach, shared resources, and more credibility than a single individual or organization. It also allows for concerns to be aired and addressed early and internally. In determining which individuals and organizations should be involved, think broadly and involve any group that stands to gain from the proposed policy change (Table 24.1 for potential stakeholders). Of note,

Table 24.1 Potential stakeholders

Voters	Professional surgical organizations (including local chapters)
Campaign donors	Professional general medical organizations
Patient advocacy organizations	Professional specialist organizations
Health economists	Quality improvement organizations
Insurers	Other academics with policy experience
Hospital organizations	Pharmaceutical/device manufacturers
Granting agencies	Home care providers
Nursing organizations	Accreditation organizations
First responder organizations	Media
Think tanks	

Table 24.2 HIV organ policy equity act supporters [6]

Legislative sponsors in the U.S. Senate: Sen. Barbara Boxer, D-CA Sen. Tom Coburn, R-OK	Legislative sponsors in the U.S. House of Representatives: Rep. Lois Capps, D-CA Rep. Andrew Harris, R-MD Rep. Michael Burgess, R-TX
AIDS United	Gay and Lesbian Medical Association
American Academy of HIV Medicine	HIV Medicine Association
American Civil Liberties Union	Human Rights Campaign
American Medical Association	Infectious Diseases Society of America
American Society of Nephrology	Lambda Legal
American Society of Transplant Surgeons	National Coalition for LGBT Health
American Society of Transplantation	National Minority AIDS Council
American Society for the Study of Liver Disease	Organization for Transplant Professionals
amfAR (American Foundation for AIDS Research)	Treatment Access Group
Association of Organ Procurement Organizations	The AIDS Institute
American Transplant Foundation	United Network for Organ Sharing
Dialysis Patient Citizens	

larger organizations, such as the American Medical Association and the American College of Surgeons, often have legislative advocacy branches and political action committees that can be substantial resources for engaging with policymakers and may be invaluable coalition partners.

For example, in 2013, President Barack Obama signed the HIV Organ Policy Equity Act (the HOPE Act), which permitted the transplantation of HIV-positive organs into HIV-positive recipients, previously been illegal under the National Organ Transplantation Act. The act drew bipartisan support, partially because of the broad coalition of stakeholders supporting it (Table 24.2). The reader is referred to the Prevention Institute's website for a detailed guide to coalition-building that is beyond the scope of this chapter [5].

24.4 Strategy

The details surrounding the actual policy are obviously situation-specific, but there are some general rules to follow. It is critical to gauge the political will for change. Some policy proposals may be effective, worthy of implementation, and backed by good evidence, but will go nowhere if they are not politically tenable. Be practical and realistic. In developing the proposed policy, the problem at hand must be clearly identified, as well as its scope and severity. Just as with clinical care and research, crafting policy is a multi-disciplinary effort. Involve health policy experts and healthcare economists. They may be invaluable in anticipating unintended consequences of the proposed policy. Additionally, they can lend their expertise in defining the financial implications of the policy changes—costs, savings, who will pay, etc. If those estimates do not exist in the healthcare/medical literature, it may be worth defining them in the form of an additional peer-reviewed publication to add to the body of work that underpins the policy proposal. Similarly, throughout the process, new directions may present themselves that may require additional data collection, analyses, and subsequent publications in order to quantify uncertainties that may result from policy change. Finally, metrics will need to be identified to determine if the policy is successful or not.

24.5 Statement

Once the policy has been developed, the statement needs to be crafted in a way that is approachable, succinct, suitable for the target audience, avoids excessive use of jargon, and makes liberal use of figures and images. In a survey of 292 state government policymakers, Sorian and Baugh reported that policymakers are frequently overwhelmed by the amount of information that they receive—only 27% of the information they receive is read in detail and 35% is never read at all [7]. They found that material was more likely to be read if it was timely (related to current debates), avoided jargon, appeared unbiased, and was not too lengthy. Additionally, presentation style matters: "Respondents said that they were more likely to read information using short bulleted paragraphs than information containing large blocks of type. Similarly, respondents found the use of charts or graphs to illustrate key points helpful." Interestingly, legislative staff noted that in addition to the executive summary, they did want more detailed information available as well, suggesting that both should be sent to policy-makers and their staff.

In addition to sending policy briefs, it is critical to meet with policymakers. In a systematic review of studies in which policy-makers were interviewed about factors that promote an evidence-based approach to policy development, a major finding was that face-to-face interaction is vital for the uptake of research and knowledge [8]. One of the first challenges can be getting such a meeting. Constituents can request meetings with their elected officials, though appointments can be very difficult to come by, especially with members of the U.S. Congress. This is an opportunity to leverage any personal contacts and connections, as well as those of others in the coalition. Professional societies, patient advocacy organizations, academic medical centers, and large hospitals often have government liaisons, which may be able to facilitate meetings. Healthcare professionals-turned-legislators may be more receptive to meetings. It may also be helpful to identify policymakers with personal connections to the cause. For example, even though unsuccessful, the first federal bill to fund dialysis and transplantation for patients with end stage renal disease was introduced by Senator Henry Jackson (D, WA) in 1965 because a former classmate of his had become one of the earliest dialysis patients [9]. Finally, as unsavory as it may seem, campaign donations open doors. Potential constituent donors are three times more likely to meet with a member of Congress and five times more likely to meet with his/her chief of staff than non-donor constituents [10].

Once a meeting with a policymaker (or his/her representative) has been established, it is important to provide a concise oral presentation. In a survey of international policy-makers, they uniformly recognized the importance of scientific research to inform their work, but they acknowledged they often do not have the technical expertise to interpret the results and scientists are frequently unequipped to succinctly and clearly explain their work [11]. It may be helpful to practice explaining the pitch in a sentence or two to non-medical friends and family (not in a run-on sentence). It can be surprisingly difficult to do without preparation. Have a mastery of all the details and be prepared to share them if asked, but keep the explanation simple, brief, and easily intelligible. In crafting the message, statistical evidence is more effective than narrative examples, but a combination of both is most effective [12, 13]. Accordingly, having a compelling patient or patient advocate attend the meeting may be helpful in swaying policymakers. For example, in 1971, Shep Glazer, a dialysis patient and advocate testified before Congress to expand Medicare to fund treatment of end stage renal disease. In addition to emotionally powerful testimony, he dialyzed in front of the House Ways & Means Committee (fortunately, the episode of ventricular tachycardia and hypotension that accompanied this dialysis session was so short-lived and reversible with immediate cessation of dialysis that his audience never realized). This display engendered significant public support and helped sway many on the committee, leading to the act's ultimate passage [14].

In addition to developing a statement for policymakers, it is critical to craft a statement for the lay public to get support for policy change. Table 24.3 lists a few public relations strategies to consider. Additionally, it may be helpful to employ the services of a public relations or marketing firm (some do pro bono work). Most medical centers and universities also have media relations teams that can assist with these efforts. In order to effect change, the lay public should be encouraged to contact their local, state, and national elected representatives by phone, mail, and e-mail. They can also try to meet with their representatives in their district offices or their governmental offices, as well as at town hall meetings. They can also sign petitions, including those that can be started at www.petitions.whitehouse.org.

Table 24.3 Public relations strategies

Social media (Facebook, Twitter, Instagram, YouTube, etc.)
Traditional media outlets (television, radio, newspaper, magazines, etc.)
Op-Eds in medical journals and in the lay press
Fundraisers
Educational events
Press releases
Highlight prominent supporters

24.6 Summary

Effecting change can take many years, requiring patience and persistence: "research evidence may hold equal, or even less importance, than other factors that ultimately influence policy, such as policymakers' values and competing sources of information, including anecdotes and personal experience." [15] In other words, politics inevitably impacts policies. However, engaging policymakers to disseminate research findings is a natural extension of the scientific process and a way to maximize its impact. Doing so requires sound science, involvement of key stakeholders, a comprehensive policy strategy, and crafting a statement that resonates with policymakers and the lay public.

References

1. Carpenter CR, Cone DC, Sarli CC. Using publication metrics to highlight academic productivity and research impact. Acad Emerg Med. 2014;21(10):1160–72.
2. Otten JJ, Dodson EA, Fleischhacker S, Siddiqi S, Quinn EL. Getting research to the policy table: a qualitative study with public health researchers on engaging with policy makers. Prev Chronic Dis. 2015;12:E56.
3. Green LW, Glasgow RE, Atkins D, Stange K. Making evidence from research more relevant, useful, and actionable in policy, program planning, and practice slips "twixt cup and lip". Am J Prev Med. 2009;37(6 Suppl 1):S187–91.
4. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA. 2003;290(12):1624–32.
5. Prevention Institute. Developing effective coalitions: an eight step guide. 2018. https://www.preventioninstitute.org/publications/developing-effective-coalitions-an-eight-step-guide
6. American Society of Nephrology. Boxer, Coburn introduce bill to end ban on research into organ donations between HIV-positive patients. 2013. https://www.asn-online.org/policy/webdocs/hopeact.pdf
7. Sorian R, Baugh T. Power of information: closing the gap between research and policy. Health Aff (Millwood). 2002;21(2):264–73.
8. Innvaer S, Vist G, Trommald M, Oxman A. Health policy-makers' perceptions of their use of evidence: a systematic review. J Health Serv Res Policy. 2002;7(4):239–44.
9. Blagg CR. The early history of dialysis for chronic renal failure in the United States: a view from Seattle. Am J Kidney Dis. 2007;49(3):482–96.
10. KJ L, BD E. Campaign contributions facilitate access to congressional officials: a randomized field experiment. Am J Polit Sci. 2016;60(3):545–58.
11. Hyder AA, Corluka A, Winch PJ, El-Shinnawy A, Ghassany H, Malekafzali H, et al. National policy-makers speak out: are researchers giving them what they need? Health Policy Plan. 2011;26(1):73–82.
12. Allen M, Preiss RW. Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. Commun Res Rep. 1997;14(2):125–31.
13. Allen M, Bruflat R, Fucilla R, Kramer M, M S, RD J, et al. Testing the persuasiveness of evidence: combining narrative and statistical forms. Commun Res Rep. 2000;17(4):331–6.
14. Rettig RA. Origins of the medicare kidney disease entitlement: the social security amendments of 1972. In: Hanna KE, editor. Institute of Medicine (US) Committee to Study Decision Making, Biomedical Politics. Washington, DC: National Academies Press (US); 1991.
15. Brownson RC, Chriqui JF, Stamatakis KA. Understanding evidence-based public health policy. Am J Public Health. 2009;99(9):1576–83.

Being a Good Mentee in Outcomes Research

25

Brian Fry and Kyle H. Sheetz

Would you tell me please, which way I ought to go from here?
That depends a good deal on where you want to go
Alice and the Cheshire Cat, Alice in Wonderland

Abstract

Fostering successful mentoring relationships is vital to trainees' development into future leaders in health services research. The right mentors can be hugely beneficial for mentees' skill acquisition, development and refinement of a career vision, creation and growth of a professional network, and the ability to pursue unique career opportunities. In the new era of mentoring relationships, it is exceedingly rare for mentees to find a "one size fits all" mentor. Instead, mentees should look to establish a network of mentors to fulfill their various professional and personal needs. Appropriate mentor selection depends on both a mentee's individual goals and identifying a mentor who meshes well with a mentee's expectations, working style, and personality. Additionally, a successful mentoring relationship largely depends on a mentee's ability to take charge of the relationship and manage up. While there are several best practice behaviors and mentee missteps to avoid, sometimes even the best mentoring relationships can become unproductive, and it's important to acknowledge when it's time to break up with a mentor. Development in academic surgery is a team effort, and finding the right team of mentors is crucial to one's personal and professional growth.

25.1 Introduction

Mentoring has long been an important component of academic medicine, and it should come as no surprise that good mentorship not only helps mentees develop personally and professionally, but also increases research productivity, boosts career

B. Fry · K. H. Sheetz (✉)
Department of Surgery, University of Michigan, Ann Arbor, MI, USA
e-mail: ksheetz@med.umich.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_25

satisfaction, and reduces burnout [1, 2]. Despite the crucial importance of mentorship, it can be difficult for trainees to identify good mentors and to foster successful, productive mentoring relationships. Many mentees also feel uncomfortable taking that first step to initiate a mentor-mentee relationship. Furthermore, mentoring has moved away from the “traditional” one-on-one, apprentice-style model to a “modern” model of a network or team of mentors. Thus, it has become even more critical for a mentee to understand how to effectively manage her mentoring relationships in order to reach her goals.

While there is a large body of literature on mentoring in academic medicine, recent work has largely focused on characterizing the core principles of successful mentoring relationships, both from the mentor and the mentee. This chapter borrows some of the ideas from this literature and adds the authors’ own thoughts as to how to be a successful mentee in the new age of mentoring relationships.

25.2 Building a Team of Mentors

Selecting the right mentor can often seem like a daunting task. It isn’t as simple as looking for the most high-profile person in your area of interest or choosing someone with whom you already have a good working relationship. Often, high-profile mentors will lack the bandwidth to take on additional mentees or give them the consistent, uninterrupted meeting time. More junior-level mentors may be fantastic at teaching you how to write a manuscript or shore up your methodology. In the early stages of learning health services research, this type of mentoring is necessary to help catapult your skills to the next level. The increasing amount of information and complexity in today’s world has required a paradigm shift in mentoring relationships. It’s now very rare for a single mentor to fit all of a mentee’s needs, and the modern model of successful mentoring requires building a team of mentors who play various roles rather than searching for the “one size fits all” perfect mentor.

Each young investigator’s team of mentors will differ depending on individual needs, but it helps to have people with diverse perspectives and at varying stages of their professional careers. Having content mentors is an absolute requirement in order to become a successful health services researcher, as it’s hard to get published and gain respect in the field without first learning how to master a specific content area. Clinical mentors may help guide a mentee’s decision making to support her future clinical aspirations such as fellowship opportunities or faculty jobs after training. Some mentors may not share a mentee’s clinical or research interests, but are vested in his overall career arc and can act as a sounding board or help mitigate conflicts. Other mentors may act as “sponsors” by using their prestige and connections to help mentees find unique, valuable opportunities for growth and advancement. Ideally, the best mentors fill multiple roles and a mentorship team operates as much more than the sum of its parts.

Lastly, having a team of mentors also can serve as a protective net for any mentee. Having several mentors means that a mentee has many people to lean on at any given time, and thus is rarely fully dependent on any one person in his/her mentorship team. For example, if a conflict arises with a specific mentor, others in the mentorship team may be able to help manage the issues at hand or provide leverage to move past them. If a particular mentor becomes hard to reach or lacks the time to commit to a meaningful mentoring relationship, a mentee can divert time and energy to this rest of his mentorship team and pursue a different direction. A mentorship team is like having a diversified stock portfolio—it's helps a mentee manage the volatility of academic medicine and makes it easier to achieve consistent growth over time.

25.3 Mentor Selection

Introspection is the first and perhaps most critical step to finding a good mentor. You cannot search for guidance without first knowing where you want to go, and your potential list of mentors will vary greatly depending on your individual needs and desires. If you are struggling with this, sometimes it's helpful to start with your long-term goals and work backwards to where you are now. It may be helpful to ask yourself the following questions, and be as specific as you can with your answers:

- What are your short- and long-term career goals?
- What are your current knowledge gaps?
- What skills do you need to develop to fill those gaps?
- What are your priorities in the next 6 months, 1 year, and 5 years?

With these answers in mind, start reaching out to people who may be able to help you along your chosen path. If you're stuck, ask your current role models, mentors, and colleagues for contacts they may know given your interests. "Shop around" and set up one on one meetings with as many people as you can. In the end, you should aim to have a diverse mentoring network that can advise you on content, clinical, strategic, and personal matters, with hopeful overlap between categories.

When you meet with prospective mentors, make sure to be up front and honest about your goals and expectations. If you are looking to gain a specific skill (i.e. becoming a better writer), you need to gauge whether your mentor is capable and comfortable helping you develop it. Not only does this ensure transparency from the start, but your discussions will likely help you refine and shape your vision and future desires. Be sure to ask what a potential mentor would expect of you and how he/she usually likes to conduct mentoring relationships (i.e. hands on versus 40,000 foot guidance) to further assess compatibility. Mentoring is a skill that's developed through time and experience, so try to seek out those who have a proven track record of mentoring but who are not too busy or overcommitted with other professional duties. Remember that it's imperative to be selfish when selecting a mentor and as your academic development time is precious—it is okay to decline a mentoring relationship if you do not think it's the correct fit.

25.4 Managing Up & Taking Charge

The best mentees understand how to take charge and "manage up" to ensure that their mentor-mentee relationships stay productive. First, a mentor's time is a limited and valuable resource. It is not a mentor's responsibility to keep a mentee on track; it is imperative to understand that the mentee is ultimately the driving force behind moving a mentor-mentee relationship forward. Work together with your mentor to set clear expectations from the beginning of the relationship. How often will you meet with one another? What is the preferred method of communication on a day-to-day basis? How much lead time does your mentor prefer for looking over drafts of abstracts or manuscripts? The more specific you can get with the general expectations of your mentor-mentee relationship, the better.

Next, the best mentees are self-directed learners who also drive their own professional growth and skill development. While you may require more support and direction at the very start of a mentoring relationship, you should do your best to find answers to questions on your own before referring to your mentor. Simple literature reviews, Google/YouTube searches, and your colleagues are great resources that can answer many initial questions. If you remain stuck and fail to gain traction, you can then go to your mentor with a problem and the steps you've already taken to attempt to solve that problem. This shows both initiative and awareness, and your mentors will appreciate that you did your homework before coming to them for help. Sometimes it's best to put yourself in your mentor's shoes—what would she do you if you were to come to her with the question at hand? You may find that you already have the resources to answer, or at least partially answer, your original question. The more prepared you can be when seeking out your mentor or prepping for your scheduled meetings, the more rich and productive the interaction will be.

A vital piece of any mentoring relationship is consistent and continuous feedback. While feedback will organically be given due to the nature of the mentor-mentee relationship, a good mentee will ensure the he is getting adequate guidance to continue pursuing his ambitions. Mentees should ask for feedback in a goal-oriented manner and be as specific as possible when soliciting evaluation. For example, instead of asking for general edits and revisions on a manuscript draft, ask how the discussion section can be improved to better communicate the impact of your findings. Or instead of asking if there's anything you can be doing better, ask if there are things you should be doing that can help make your mentor-mentee meetings more productive and efficient. It's also important to realize that feedback in a mentor-mentee relationship is a two way street, and that good mentors will solicit feedback from their mentees. The mentoring relationship should be constantly adapting and transforming based on the needs of those involved. The more feedback a mentee gives to her mentor, the better the mentor can understand that mentee's specific needs, and the more effective the mentoring relationship.

25.5 Best Practice Behaviors for Mentees

While taking charge and managing up are critical skills to develop for successful mentoring relationships, there are also specific "best practice" behaviors that will help a mentee maximize the precious time he gets from his mentor (Table 25.1).

Table 25.1 Best practice behaviors for mentees

Behavior	Description
Underpromise and overdeliver	No one enjoys mentoring someone who continually pushes back deadlines or underdelivers results. Strive to be responsible, on time, and produce your highest quality work.
Schedule frequent meetings	Shorter, more frequent meetings are often more productive than the occasional longer meeting. The more touch points a mentee can have with her mentor, the better. Mentors are typically very busy, so schedule meetings out well in advance.
Create an agenda for meetings	Showing up to meetings empty handed or without an agenda is a waste of mentors' precious time. Mentees should always have goals and a plan for every meeting and make sure their mentors are aware of them before the meeting starts. If you are hoping to have a manuscript or grant reviewed it, send a copy early so your mentor has time to edit before your meeting—use your meeting time for discussion.
Audio record conversations (with permission and keep to yourself)	It may seem awkward at first but having an audio transcript of mentoring conversations allows mentees to stay in the moment instead of furiously scribbling down notes while in a meeting. Mentees may often find themselves wishing they remembered that sentence or idea their mentors mentioned a few weeks ago—if mentees audio record conversations, they have easy documentation of important discussions with their mentors.
Minimize talking to listening ratio	Some mentees make the mistake of doing too much talking and too little listening during meetings. Mentees should remember they are seeking advice and guidance from mentors; the more mentees are listening, the more opportunity for natural learning to occur.
Always follow up	It is always the mentee's responsibility to promote closed loop communication as mentors are too busy to keep tabs on all of their mentees progress at all times. Mentees should give mentors updates when appropriate and never "disappear" for long stretches at a time. Mentees should also not be afraid of reminding mentors of important meetings or deadlines.
Maintain a growth mindset	Feedback is not always easy to hear, but good mentees find ways to incorporate even the harshest of comments and use a growth mindset to promote continual learning and development.
Know when to say "no"	It's always tempting to agree to anything a mentor proposes but overcommitting oneself can erode trust in mentoring relationship. Mentees should know their limits and when it's better to pass on a project or opportunity so that they can give their existing work the effort it requires. It is better to do one thing really well than many things poorly.
Always be honest	Transparency and open communication are key in any mentoring relationship, and it starts with the mentee. Mentees should always be honest about what you know/don't know and can/can't do.

Many of the following may seem like common sense, however, they can drastically improve a mentee's effectiveness. This list is not exhaustive and mentees should always consult their mentors for additional best practice behaviors specific to their relationship.

25.6 Breaking Up Is Hard to Do

Another key skill for mentees to develop is the ability to recognize when it is logically time to move on from a mentoring relationship, or in the difficult situation when a mentor is committing "mentorship malpractice" [3]. In the latter example, some behaviors are benign but irresponsible, such as a mentor lacking enough time and bandwidth to truly support a mentee's development (a mentor's version of overpromise and underdeliver). These situations may be successfully navigated by asserting firm boundaries and maintaining an open, honest line of communication. Other behaviors, such as imposing low value tasks on mentees or severe micromanaging that hinders mentees' productivity are more malignant in nature and may require a mentee to break off her relationship with a mentor. Knowing how to combat unproductive mentoring behaviors may help right the ship and are a crucial tool in any mentees' toolbox. However, even if mentees do everything right, it may not be enough to salvage a mentoring relationship gone bad.

When considering whether to end a mentor-mentee relationship, there are a few recommended steps to follow [4]. The first step is fairly simple: evaluate if you are truly getting what you need from the relationship. Try to pinpoint exactly what is lacking or where things need to improve. It may be that you and your mentor don't share a good rapport with one another. Or maybe you are moving in a different direction than your mentor and it doesn't make sense for either of your to continue the current relationship. Whatever the reason(s), think long and hard about why your mentoring relationship has been unsuccessful. After this, decide if these reasons are amenable to being fixed. If you've already invested time and energy into a relationship, it's important not to jump ship at the first sign of conflict. After all, a mentor may not even know something is a issue if the mentee has never mentioned it before. Some issues may be fixed with an open, honest conversation. However, some issues are too big to be overcome and it's worth ending things before the relationship has truly soured.

If you decide to break up with you mentor, it's important to "disengage with gratitude," and leave the relationship as gracefully as possible. Be professional and don't draw things out or slowly stop responding to emails. Be sure to show gratitude and mention to your mentor how important the relationship has been to you and how you've grown along the way. Then frame the discussion in terms of how your needs as a mentee have changed rather than how your mentor is not doing this or that for you. Be direct but not blunt. The goal should not be to make a mentor feel bad, but instead to share your perspective and reasoning for why you need to move on. Lastly, try as hard as you can not to burn bridges—you never know when you may come across this person again, perhaps as a colleague, boss, or subordinate. Your old mentors will likely still be an important people in your field, and you may end up needing them again sometime in the future.

25.7 Conclusion

Finding effective mentorship is crucial in one's development as a trainee. The search for the right mentor starts from within, and it's important for mentees to understand where they want to go before looking for someone to guide them there. Building a team of mentors ensures that mentees have a diverse support network to meet their various needs. The best mentees know how to manage up and take charge of their mentoring relationships. However, they also know when it's time to gracefully break up from mentors when relationships become caustic or unproductive. Successfully navigating the waters of mentorship can have a huge positive impact on one's career trajectory, research productivity, and overall satisfaction.

References

1. Sambunjak D, Straus SE, Marušić A. Mentoring in academic medicine: a systematic review. JAMA. 2006;296(9):1103–15.
2. Elmore LC, Jeffe DB, Jin L, Awad MM, Turnbull IR. National survey of burnout among US general surgery residents. J Am Coll Surg. 2016;223(3):440–51.
3. Chopra V, Edelson DP, Saint S. Mentorship malpractice. JAMA. 2016;315(14):1453–4.
4. O'Hara C. How to break up with your Mentor. Harvard Business Review. May 29, 2014. https://hbr.org/2014/05/how-to-break-up-with-your-mentor

Suggested Readings

Bregman P. How to ask for feedback that will actually help you. Harvard Business Review. December 5, 2014; https://hbr.org/2014/12/how-to-ask-for-feedback-that-will-actually-help-you

Chopra V, Arora VM, Saint S. Will you be my Mentor?—four archetypes to help mentees succeed in academic medicine. JAMA Intern Med. 2018;178(2):175–6.

Cochran A, Elder WB, Neumayer LA. Characteristics of effective mentorship for academic surgeons: a grounded theory model. Ann Surg. 2019;269(2):269–74.

Cho CS, Ramanan RA, Feldman MD. Defining the ideal qualities of mentorship: a qualitative analysis of the characteristics of outstanding mentors. Am J Med. 2011;124(5):453–8.

Straus SE, Johnson MO, Marquez C, Feldman MD. Characteristics of successful and failed mentoring relationships: a qualitative study across two academic health centers. Acad Med. 2013;88(1):82–9.

Vaughn V, Saint S, Chopra V. Mentee missteps: tales from the academic trenches. JAMA. 2017;317(5):475–6.

Finding a Mentor

26

Jessica Ruck and Fabian M. Johnston

Abstract

The benefits of mentorship are increasingly recognized and include accelerated advancement of mentees, improved program retention and diversity, and greater research productivity. However, finding a mentor can be challenging. When searching for a mentor, reflect on what you're hoping to gain. Also, consider how many mentors you might need; having a mentorship *team* might better fulfill your mentorship needs than relying on one busy person. Mentors can be found anywhere, but look for a good fit. Consider what you would like your mentors to have in common with you but also whether diversity among your mentors might offer important, contrasting viewpoints and access to expanded networks. Finally, once you've found a mentor, discuss your expectations, invest in the relationship, assess whether the mentorship is effective, and recognize when a mentorship relationship needs to be terminated. Receiving good mentorship can be invaluable and can prepare you to become a mentor one day, as well.

26.1 Introduction

Mentorship has become a buzzword, something that is often discussed in academic surgical circles but can still seem elusive. You might feel that you're the only one struggling to find a mentor, but the numbers suggest that difficulty finding a mentor

J. Ruck
Department of Surgery, Johns Hopkins Hospital, Baltimore, MD, USA
e-mail: jessicaruck@jhmi.edu

F. M. Johnston (✉)
Department of Surgery, Johns Hopkins Hospital, Baltimore, MD, USA

Division of Surgical Oncology, Department of Surgery, Johns Hopkins Hospital, Baltimore, MD, USA
e-mail: fjohnst4@jhmi.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_26

is a common problem. By some accounts, fewer than 50% of medical students and, in some fields, fewer than 20% of faculty members have a mentor [1]. This, of course, bears the question of who *should* have a mentor, particularly in the field of academic surgery.

In this chapter, we aim to answer that question and provide some tips on how to find a mentor. The short answer is, we believe that everyone should have a mentor—at *least* one. This includes surgeons at all levels—medical students aspiring to become surgeons, residents in the thick of their training, fellows who are interviewing for their first attending position, and attendings of all levels as their careers mature. There is near-universal agreement that mentorship is important to academic success [2], which helps explain why mentorship has become a core value of surgical professional societies including the Association of Academic Surgeons [3]. In the sections that follow, we'll break down why we believe mentorship is so important and why you should prioritize finding a mentor.

26.2 Benefits of Mentorship

Mentorship offers myriad benefits to surgeons, surgery departments, and the field of surgery, including increasing the success and pace of advancement of mentees, retention [2], and diversity [4, 5]. Mentors can help mentees evaluate different fields and decipher the best fit. Once someone identifies their field of choice, mentors can facilitate their entrance into that field as well as their long-term success [4]. This might seem too good to be true (and make a mentor seem more like a fairy godmother), but the link between mentorship and success is based on numerous steps along the way at which mentors help shape and support their mentees' personal development and careers. Within academia, where research is often a necessity for advancement, mentorship has been associated with higher productivity, as measured by publication rate and grant application success [1]. Conversely, lack of mentorship is associated with lower likelihood of achieving research milestones, such as producing first-author peer-reviewed publications and teaching sessions at a national meeting [6]. This should come as no surprise to someone who has struggled to publish a manuscript; good mentorship can help you learn study design, study execution, manuscript drafting and revision, identification of appropriate target journals, and navigation of the peer review process. In these moments, mentorship is invaluable.

26.3 Finding Mentorship

Finding a mentor requires [1] knowing what you're looking for and [2] identifying people who are willing to be mentors. This is a moment to be introspective about who you are and what you're looking for. Before you meet with a potential mentor, you need to understand your own needs and frame discussions with potential mentors around these needs. As your mentor, this person will be accepting the responsibility to answer your questions, create opportunities for you, and provide guidance on how to grow your skills, portfolio, and career. Therefore, choosing a mentor is an important task, and having them accept the role of mentor is equally important. We advocate having multiple mentors, as it is unlikely that one person will be able to provide all of the mentorship and guidance that you need throughout your career and personal development.

26.3.1 Where to Look

There is no wrong place to look for a mentor. You might find them through your division chief or department chair during a meeting to discuss your interests and goals, through a colleague at your institution who hears about your research aims, through collaborators in other departments who mention someone they respect in your field, at conferences during research talks or working groups, at Grand Rounds, through journal clubs, through Twitter discussions about hot topics in your field, or just about anywhere else. If someone inspires you, offers wise words, or has had experiences from which you think you could learn, that is a relationship you can explore and possibly develop into mentorship. Not every person that you approach will be interested in being a mentor, and you might not "click" with certain potential mentors. Developing your mentorship team will be an ongoing process, and you will figure out what type of mentorship works best for you and what qualities you therefore need to seek in potential mentors.

26.3.2 Finding Mentorship as a Woman or an Underrepresented Minority

Some mentors might be similar to you in background, skills, and interests, while others might be quite different. While there are no rules for you your mentor should be, studies have found that people often seek out people who share their characteristics, particularly gender and race.

Women in surgery—as well as in other academic medicine disciplines [7]—report facing gender-based challenges in accessing mentorship. These include less access to collaboration and support [8], greater difficulty finding mentors compared to male colleagues [1], lack of access to same-sex mentors and role models contributing to slower progression through the ranks [9], and differential treatment in general based on sex [10]. The lack of female mentors is repeatedly cited and is consistent with findings that, at least in certain surgical fields, the majority of mentors are older men [11]. These and other challenges for women have led to the creation of organizations such as the Association for Women Surgeons, which provides support, community, and mentorship for women by (mostly) women. Additionally, women in surgery are finding each other within training programs, departments, and online to access support and mentorship around shared experiences. While these are excellent resources for female surgeons and surgeons-in-training, it is just the first step in leveling the mentorship playing field. If you are a woman in surgery, it is important to remember that your mentor does not have to be another woman (although we must keep working to make that an accessible option), that you are not alone in your frustrations with the lack of female mentors, and that the field of surgery is changing and you will be a part of a generation of mentors for younger women one day. In the meantime, we must train men and women to be better mentors for aspiring female surgeons.

This desire to find a mentor who "looks like you" is not restricted to gender, but also found in studies of mentorship and race/ethnicity. Residents from underrepresented racial and ethnic minorities also often actively seek out mentors of the same race or ethnicity, but they may find it difficult to identify mentors that fit these parameters [12]. To the mentees, finding a mentor of a similar background might be important because they feel they will be better able to commiserate about a shared experience or because they view the mentor as a role model. Interestingly, program directors are less likely than their residents to place importance on shared demographic characteristics between trainees and their mentors [13], which might lead program directors to suggest or assign different mentors than the mentees would seek out themselves. The relationship between these shared mentor-mentee characteristics and the effectiveness of the mentorship provided has not been rigorously studied. However, if you desire a mentor with a shared ethnic heritage there are multiple organizations which pride themselves on building community among and mentoring of underrepresented minorities, including the Society of Black Academic Surgeons (SBAS), Society of Asian Surgeons (SAS), and Latino Surgeons Society (LSS), among others.

Therefore, as you search for a mentor, we encourage you to think about what you are looking for in a mentor and why. Which characteristics are most important? Have you prioritized certain mentor characteristics because you believe you will be better understood, because you think that they will open doors for you and help you achieve certain goals, or for another reason? Perhaps a mentor with a similar background will be the best fit, but we also encourage you to reflect on whether a diverse group of mentors will provide additional benefits. Diversity in the backgrounds of your mentors can also help expose you to various perspectives, experiences, and resources.

26.4 Creating a Successful Mentorship Relationship

People seek different things from their mentors depending on their level of training, their strengths and weaknesses, and the mentorship that they are already receiving from others. However, there are qualities that are common to many good mentors: approachability and accessibility [13]; the ability to ensure open communication, maintain confidentiality, and prevent mentor-mentee competition [2]; a strong sense of collaboration [14]; creation of humanized relationships and a holistic view of how to support mentees [14]; commitment to a longitudinal relationship with mentees [15]; and willingness to engage with diverse mentees [10]. Mentors may not possess all of these qualities—accessibility can certainly become an issue at the peak of a mentor's career—but they are certainly qualities that mentors should strive to have and maintain.

Beyond general qualities, there are important steps that mentees should take to create successful mentoring relationships. One of the most important steps is establishing common expectations with the mentor. Multiple studies have found that clear governance arrangements, clearly defined role boundaries, and acknowledgments of mentees' and mentors' limitations increased the likelihood that mentorship relationships will succeed [15, 16]. In contrast, mismatched expectations can be a barrier to successful mentorship [16]. It is important to figure out why you feel a person would be a valuable mentor so that you can explain to them how they can best support you and/or help you advance. Do you need a sense of support or are you looking to them to challenge you? Are you hoping to collaborate on research projects, get advice on whether to take on administrative responsibilities, or gain introductions within your field? Once the "why" of the mentorship relationship is established, you must discuss the "how." How formal of a mentor-mentee relationship do you want, and what will that look like? Will you have regular scheduled check-ins or will you reach out when questions or concerns arise? Mentors and mentees may have different preferences as to the setting and frequency of meetings; one study found that mentees prefer one-on-one meetings and more frequent interaction, while mentors prefer group meetings and less frequent mentor-mentee interaction [11]. This might reflect the time limitations of mentors, an identified barrier to mentorship [16], but nevertheless underscores the importance the practical aspects of mentorship.

Just as important as your expectations and responsibilities are those of your mentors. Be prompt and present when responding to your mentors' emails. Make it easy for them to respond to your emails, edit manuscript drafts, write letters of recommendation, sponsor you for travel grant applications, and generally advocate for you. If you want your mentor to go to bat for you, be clear in your requests and be prepared to provide facts and figures they can use to talk you up. Show respect for your mentors' busy schedule by providing plenty of advance notice before a deadline. By showing that you understand the many demands on your mentors' time—indeed, a "time crunch" from many competing demands is one of the challenges faced by mentors and a frustration to mentees everywhere—you can show that you are thoughtful, respectful of their many obligations, and appreciative of the time they take to mentor you.

Methods of ensuring that mentors and mentees have realistic expectations about their relationship and that both parties commit to their mentorship relationship responsibilities range from discussions via email or in person to the creation of a mentorship contract. This contract forces the mentor and mentee to acknowledge the developing relationship, commit to it, and delineate their expectations. When writing a formal contract, you might find that you are forced to discuss the specifics instead of relying on vague generalizations about what the relationship will be. This is consistent with the SMART acronym for goal-setting: the goals should be specific, measurable, achievable, realistic, and timely. By creating a mentorship contract, you will also have an opportunity to determine which gaps remain in your mentorship network and consider how you can fill them. Indeed, mentors may encourage their mentees to have multiple mentors to address unique domains [17, 18] or provide multiple perspectives. Of note, mentorship relationships can change over time and expectations should be readdressed periodically, particularly if there is a change in professional role for the mentor or mentee.

26.5 Have You Found a Mentor or a Sponsor?

As discussions of mentorship become more common and more nuanced, distinctions between different types of mentor-mentee relationships have grown clearer. One of the most commonly discussed distinctions is mentors versus sponsors. While mentors are present to discuss your next steps with you—indeed, they may be the only people having these important conversations with you—sponsors take on an advocacy role. While mentors serve as a valuable source of advice, ideas, and experience, sponsors openly and actively advocate for you. A sponsor uses their connections and influence to make their mentee more visible in the workplace or in their field [19].

The line between a mentor and a sponsor can be blurry—both are often at a more senior level and may have access to professional societies and networks that you have yet to join. The difference is that if an opportunity arises, a sponsor considers whether their mentee (you) would be appropriate for the award, position, or opportunity and extends an invitation or nominate them. They actively take on the role of opening doors and advocating for you to help you achieve your goals and advance personally and professionally.

26.6 Formalizing the Mentor Role

As the value of mentorship in medicine and surgery has grown increasingly apparent, there has been an increase in formalized mentorship programs within academic centers and national professional organizations aiming to increase and improve mentorship. Essentially, leaders at all levels have realized that finding a mentor can be hard, that mentorship is valuable, and that they can assist others in finding a mentor. Despite this, a recent study found only 54% of departments of surgery around the United States had established mentorship programs, and those that did often had no evaluation forms, no training for mentors or mentees, and no exit strategy for failed mentor-mentee relationships [19]. Additionally, in more than two thirds of departments, faculty mentoring efforts were not recognized formally by the department or institution [19]. These challenges can make it difficult to identify a mentor because there are fewer resources and programs to help you and a lack of support for and prioritization of the mentorship relationship. In short, this is a part of why so many people say that they want mentorship but struggle to find it.

Recognition of the challenges that surgeons-in-training and surgeons face in finding mentorship, there have been numerous initiatives to facilitate the finding-a-mentor process. At the national level, initiatives to improve mentorship have included the Diverse Surgeons Initiative, which seeks to increase the number of under-represented minority faculty members in surgery by creating a grant to fund training of under-represented minority surgical trainees in advanced minimally invasive surgery skills and mentorship in this field [20]. Among participants, 99% completed general surgery residency, 87% completed fellowships, 50% became Fellows of the American College of Surgeons, 76% had contributed to peer-reviewed literature, 41% were on faculty, and 18% held local, regional, or national leadership positions [21]. With so many of the program participants holding leadership initiatives, we hope that the Diverse Surgeons Initiative will help to strengthen mentorship and promote diversity within the field.

At the center level, initiatives have targeted mentorship relationships at the faculty and resident level, recognizing that surgeons at all levels benefit from mentorship. At the faculty level, a pilot program paired junior faculty with senior faculty mentors, trained mentors, and included structured and informal mentorship meetings [22]. Division chiefs reviewed milestones and mentorship during meetings with participants to assess satisfaction and effectiveness of the program. Of note, the authors found that nearly all junior faculty members had multiple mentors beyond those assigned to them. We'll say it one more time—having a mentorship *team* is important, no matter what your level of training is!

At the resident level, single-center mentorship initiatives have included speed dating to help junior residents identify senior resident mentors based on 90-second interactions. Despite the brief interactions, 85% of participants were satisfied with the mentorship pairings that resulted [23]. Some of these mentorship pairings might have been people meeting for the first time, while for other participants the speed dating format gave them a chance to see who was interested in being a mentor or mentee and ask for a mentorship relationship. This brings up an important point: mentors can be people you already know, and can grow out of existing relationships with a colleague or advisor. Another resident mentorship initiative paired eight residents with a main faculty mentor as well as several supplemental mentors; residents were allowed to select their mentor of choice. Again, this program recognized the importance of a mentorship team rather than relying on a single mentor. The program led to improved quality of life among residents, lower emotional scores (a measure of stress), lower levels of depersonalization by residents, and higher personal achievement, as measured by validated instruments [24]. In summary, surgeons flourish when they have a team of people supporting them, and programs that help people to create those teams benefit quality of life, personal development, and career development.

Outside of surgery, national initiatives to encourage mentorship have been created in the fields of pediatrics and pharmacology. These programs serve as examples for additional mentorship programs that could be created within surgery; national initiatives offer the benefit of pairing mentors and mentees who are not at the same institution, recognizing that the best mentor for someone based on their interest and goals might not be within their own center. The New Century Scholars program in pediatrics pairs residents interested in academic careers with both junior and senior faculty mentors and provides travel grant funding; this resulted in 63% of participants pursuing academic careers [25]. Programs that fund mentees' travel to national meetings offer opportunities for mentors to step into the role of sponsor by introducing their mentees to their peers and collaborators, therefore helping the mentees build their own networks. In academic pharmacology, a mentorship initiative to identify the needs and responsibilities for mentors at all levels resulted in the creation of the PAIRS Faculty Mentorship Checklist. This checklist can be filled out by mentors and mentees and provides formal recognition of mentorship responsibilities and a tool for mentor-mentee dialogues and mentor self-reflection [26]. It can lend structure to conversations about how mentorship relationships are functioning and help mentors and mentees to identify ways to improve their mentorship. Finally, a center-level study of mentorship in Emergency Medicine evaluated the impact of a tiered mentorship program that extended formal mentorship to medical students. The program resulted in a dramatic increase in the percentage of students who identified a mentor within that field. Notably, the junior medical students highly valued the mentorship of not only the faculty mentors but also of the more senior medical students, demonstrating that medical professionals at all levels can act as both mentors and mentees. This is an important reminder to surgeons at all levels that even as you are looking for a mentor, people may be looking for your mentorship.

26.7 Are You Receiving Effective Mentorship?

You've found a mentor, and you're excited about your career trajectory. But along the way at some point, there are challenges in your mentor-mentee relationship and you wonder if you found a good mentor after all. Quality of mentorship and consistency of mentorship are major challenges for formalized mentorship programs, which has inspired questions of whether mentors should receive training in mentorship [4] and how we should measure whether mentorship relationships are effective. How, then, are you supposed to figure out whether your mentor is a good mentor for you? For department chairs, how do you figure out who is providing good mentorship, and how do you support mentorship in your department?

As mentorship research has matured in the surgical literature and across other industries, tools that have been developed to evaluate mentorship have started to provide insights into characteristics of good mentors and of the programs that support good mentorship. These tools include the Mentorship Skills Self-Assessment Survey, which assesses mentor-specific and mentee-specific skills and evaluates mentorship relationships in an academic setting [27], and numerous mentorship measurement tools in other industries [28] that might, with modifications, provide useful tools for assessment of mentorship quality and success in academic surgery. Early research in academia has found that people who perceive that their mentor is doing a good job are also those who are enjoying greater academic success [27]. This is not particularly surprising; if your career is taking off, wouldn't you be more likely to rate your mentors as doing a great job? Still, it is far from demonstrating a causal relationship and shows that we have a lot left to learn about mentorship in academia and, specifically, academic surgery.

A second challenge is how to ensure that good mentorship is valued, so that good mentors continue to have time and resources to dedicate to their mentorship. After all, time is often in short supply in academic surgery; lack of time and compensation for mentorship activities have been reported as major barriers to effective mentorship [16]. In light of this, have departments of surgery moved to compensate surgeons for mentorship efforts? A study in 2016 found that only 3% of departments of surgery provided economic support for mentoring programs [19], so we have a long way to go until mentorship activities are rewarded in academic surgery. This lack of financial support might be viewed as a reflection of the lack of emphasis and importance placed on mentorship within departments and could highlight a need for cultural change within academia and within surgery to value time spent on mentorship more highly. As a mentee, what does this mean for you? It means that as you search for a mentor, you should be conscious that the valuable services a mentor provides are done without any compensation and with an opportunity cost of how that time could otherwise be spent. Be sure to tell your division and department chair how important mentorship is to you, and be a part of the cultural change within academic surgery to recognize and reward mentorship.

26.8 Benefits to the Mentor

Though the effects of mentorship are often measured by the success of the mentee, we believe that mentorship also offers benefits to the mentor. Mentorship creates connections between the mentor and the next generation of leadership in their field, can change the culture and face of the field, and can generate relationships that lead to clinical or research collaborations. While such benefits can be difficult to measure, metrics including "collaborative distance" have been developed to quantify the degree to which two investigators are related as coauthors and may partially capture the effects of mentor-mentee relationships. Within surgery, this principle has been applied to John Cameron, the former Chair of the Department of Surgery at Johns Hopkins Hospital. Cerullo et al. found that 84% of the trainees under Dr. Cameron's leadership pursued academic careers, with 72% having achieved the rank of Associate Professor or above and 53% having achieved the rank of full professor by 2015 [29]. Additionally, 21 of his trainees had served as program directors, division chiefs, or department chairs at academic centers, and 16 had served in leadership on national or international surgical or multidisciplinary committees [30]. These numbers underscore the strength of networks that can result from strong mentorship, and they serve as an important reminder that mentees become a part of their mentors' network and can be a valuable asset from a clinical, research, and/or personal development standpoint.

26.9 What to Do When Your Mentor-Mentee Relationship Isn't Working

Even if your mentor is perfect for you on paper, that doesn't mean that you two will "click" or that things will go smoothly. One study of clinician-investigator trainees (not limited to surgery) found that over 60% of mentees expressed some dissatisfaction with the mentorship they received [2]. This could be because of a toxic mentor, but it could also be because your mentor was promoted and their new responsibilities have given them less time to provide effective mentorship. Navigating the termination of these mentorship relationships can be difficult for both the mentor and the mentee, and there is a need for transparent pathways for problem escalation in mentorship relationships [15].

Research on mentorship relationship termination is limited, particularly in the literature on academic surgery. Therefore, we have pulled guiding principles from research on ending such relationships across industries and types of mentorship. First, you and your mentor should discuss the fact that your mentorship relationship might end at some point. This conversation should ideally occur early in the mentorship relationship, so you can set the expectation that it's okay to end a mentorship relationship if it's not working. This type of no-fault termination could be due to personal incompatibility between mentor-mentee, misalignment of goals, or other reasons [31]. If you do decide to end the relationship, give your mentor ample warning—often a few weeks' notice—to allow you to wrap up any collaborations and reach some sort of closure. Be straightforward; this has been shown to make the termination less complicated and hurtful [32]. If your mentor terminates the relationship, recognize that it was likely for the best, and that they did this because they felt that you would be better served by a different mentorship relationship.

If you are terminating a mentorship relationship for another reason, such as unacceptable behavior (e.g. breach of confidentiality, lack of actual mentorship, lack of trust, romantic issues), it might be beneficial to have a neutral party involved in the termination process [31]. This neutral party could be a program director, other departmental leader, or a department or hospital ombudsman. They can help mediate the conversation, document the reasons for relationship termination, and prevent retaliation (if that is a concern). It is important to explicitly state the reasons for terminating the relationship so that the mentor can understand how their behavior made you feel or so that you can understand how you made your mentor feel. This can also inform the conversations that the mentee and mentor have in their next mentorship relationships. As mentorship in surgery develops, we will need to study not only how to create optimal mentorship relationships, but how best to terminate mentorship relationships that aren't working.

26.10 Towards a Better Understanding of Mentorship

One barrier to successful mentorship is our incomplete understanding on what makes mentors, mentees, and the mentorship relationship successful. While the studies discussed in this chapter provide insights into what mentors and mentees view as important and their views on the importance of mentorship, many of the studies referenced are small and used surveys to assess participants' perceptions instead of objective outcomes. Additionally, some of these studies were drawn from academic medical fields outside of surgery, which may limit their generalizability to our unique experiences and needs as surgeons. However, we present them because we feel that many of the characteristics and challenges of mentorship in the academic setting are likely universal. As the field of academic surgical mentorship research matures, the development of objective assessment tools to consistently measure outcomes of mentorship in academic surgery will allow comparisons of the effectiveness of interventions and identification of best practices. We hope to see mentorship become more rigorously studied, formally recognized, and highly valued in academic surgery. In the meantime, we hope that you find a great mentorship team.

References

1. Sambunjak D, Straus SE, Marusic A. Mentoring in academic medicine: a systematic review. JAMA. 2006;296(9):1103–15.
2. Ng E, Wang X, Keow J, Yoon JY. Fostering mentorship for clinician-investigator trainees: overview and recommendations. Clin Invest Med. 2015;38(1):E1–e10.
3. Kao LS. Real world versus ivory tower: the challenge for academic surgery. J Surg Res. 2014;190(1):9–15.
4. Franzblau LE, Kotsis SV, Chung KC. Mentorship: concepts and application to plastic surgery training programs. Plast Reconstr Surg. 2013;131(5):837e–43e.
5. Nellis JC, Eisele DW, Francis HW, Hillel AT, Lin SY. Impact of a mentored student clerkship on underrepresented minority diversity in otolaryngology-head and neck surgery. Laryngoscope. 2016;126(12):2684–8.
6. Reid MB, Misky GJ, Harrison RA, Sharpe B, Auerbach A, Glasheen JJ. Mentorship, productivity, and promotion among academic hospitalists. J Gen Intern Med. 2012;27(1):23–7.
7. Bhatia K, Takayesu JK, Arbelaez C, Peak D, Nadel ES. An innovative educational and mentorship program for emergency medicine women residents to enhance academic development and retention. CJEM. 2015;17(6):685–8.
8. Sonnad SS, Colletti LM. Issues in the recruitment and success of women in academic surgery. Surgery. 2002;132(2):415–9.
9. Zhuge Y, Kaufman J, Simeone DM, Chen H, Velazquez OC. Is there still a glass ceiling for women in academic surgery? Ann Surg. 2011;253(4):637–43.
10. Cochran A, Elder WB, Crandall M, Brasel K, Hauschild T, Neumayer L. Barriers to advancement in academic surgery: views of senior residents and early career faculty. Am J Surg. 2013;206(5):661–6.
11. Janis JE, Barker JC. Medical student mentorship in plastic surgery: the Mentor's perspective. Plast Reconstr Surg. 2016;138(5):925e–35e.
12. Yehia BR, Cronholm PF, Wilson N, Palmer SC, Sisson SD, Guilliames CE, et al. Mentorship and pursuit of academic medicine careers: a mixed methods study of residents from diverse backgrounds. BMC Med Educ. 2014;14:26.
13. Smeds MR, Huynh C, Thrush CR, Moursi MM, Amankwah KS. Effects of mentorship on graduating vascular surgery trainees. Ann Vasc Surg. 2017;44:234–40.
14. Rosengart TK, Kent KC, Bland KI, Britt LD, Eberlein TJ, Gewertz BL, et al. Key tenets of effective surgery leadership: perspectives from the Society of Surgical Chairs Mentorship Sessions. JAMA Surg. 2016;151(8):768–70.

15. Collins K, Oliver SW. Mentoring: what matters most? Clin Teach. 2017;14(4):298–300.
16. Leary JC, Schainker EG, Leyenaar JK. The unwritten rules of mentorship: facilitators of and barriers to effective mentorship in pediatric hospital medicine. Hosp Pediatr. 2016;6(4):219–25.
17. Pentland L. Mentorship vs. sponsorship, and how to maximize both. Forbes. 2015 2, 2015.
18. Cochran A, Elder WB, Neumayer LA. Characteristics of effective mentorship for academic surgeons: a grounded theory model. Ann Surg. 2019;269(2):269–74.
19. Kibbe MR, Pellegrini CA, Townsend CM Jr, Helenowski IB, Patti MG. Characterization of mentorship programs in departments of surgery in the United States. JAMA Surg. 2016;151(10):900–6.
20. Butler PD, Britt LD, Green ML Jr, Longaker MT, Geis WP, Franklin ME Jr, et al. The diverse surgeons initiative: an effective method for increasing the number of under-represented minorities in academic surgery. J Am Coll Surg. 2010;211(4):561–6.
21. Butler PD, Britt LD, Richard CE, Chang B, Serletti JM, Green ML Jr, et al. The diverse surgeons' initiative: longitudinal assessment of a successful national program. J Am Coll Surg. 2015;220(3):362–9.
22. Phitayakorn R, Petrusa E, Hodin RA. Development and initial results of a mandatory department of surgery faculty mentoring pilot program. J Surg Res. 2016;205(1):234–7.
23. Caine AD, Schwartzman J, Kunac A. Speed dating for mentors: a novel approach to mentor/mentee pairing in surgical residency. J Surg Res. 2017;214:57–61.
24. Zhang H, Isaac A, Wright ED, Alrajhi Y, Seikaly H. Formal mentorship in a surgical residency training program: a prospective interventional study. J Otolaryngol Head Neck Surg. 2017;46(1):13.
25. Pachter LM, Kodjo C. New century scholars: a mentorship program to increase workforce diversity in academic pediatrics. Acad Med. 2015;90(7):881–7.
26. Law AV, Bottenberg MM, Brozick AH, Currie JD, DiVall MV, Haines ST, et al. A checklist for the development of faculty mentorship programs. Am J Pharm Educ. 2014;78(5):98.
27. Athanasiou T, Patel V, Garas G, Ashrafian H, Shetty K, Sevdalis N, et al. Mentoring perception and academic performance: an academic health science centre survey. Postgrad Med J. 2016;92(1092):597–602.
28. Chen Y, Watson R, Hilton A. A review of mentorship measurement tools. Nurse Educ Today. 2016;40:20–8.
29. Cerullo M, Lipsett PA. Assessing the magnitude of a surgical career through his trainees: the John L. Cameron legacy factor. Ann Surg. 2017;265(5):866–8.
30. Lipsett PA. The John L. Cameron legacy factor: further quantifying his impact. Ann Surg. 2018;267(2S Suppl 2):S10–s12.
31. Association AAoD. Ending the mentorship. 2018. https://www.aad.org/members/leadership-institute/mentoring/ending-the-mentorship
32. Morrow K, Styles M. Building relationships with youth in program settings. Philadelphia, PA: Public/Private Ventures; 1995.

What Every Outcomes Research Fellow Should Learn

27

Mohammad Ali Abbass and Christy Elaine Cauley

Abstract

Many general surgery residents dedicate time during residency training to research. Being successful during your research time is dependent upon your own personal and professional goals. Once you determine your goals, you should surround yourself with the mentors and resources you need to achieve them. By creating a realistic plan for your time and making the most of available opportunities you can develop the skills and relationships you need to build a successful career.

27.1 Introduction

Many general surgery residents dedicate time during residency training to research. The majority of residents spend two or more years focusing on research, and 70% of residents perform clinical research as their main focus or in combination with basic science research [1–3]. This break in clinical training provides residents with the opportunity to gain a broader perspective for clinical questions, teaches them how to develop research ideas and enhances their ability to evaluate scientific literature [1]. Making the most of your research time can prepare you for the career and life you are working towards.

M. A. Abbass
Department of Surgery, Cleveland Clinic, Cleveland, OH, USA

C. E. Cauley (✉)
Department of Surgery, Massachusetts General Hospital, Boston, MA, USA
e-mail: ccauley@partners.org

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_27

27.2 Identify your Goals

Identifying your immediate and long term career goals will help you develop a plan for your research time. The most common reasons general surgery residents take time for research are “to help obtain a fellowship” and “to initiate an academic career” [1]. When considering your goals, identify what skills or experiences will make you the best candidate for your chosen fellowship and career. Some goals for your research time might include: improving your communication or analytic skills, preparing for exams, learning to write grants, obtaining additional degrees, or spending more time with family and friends (see Fig. 27.1). Once you define your goals you can focus on being productive in the areas that matter most to you.

If your goal is to help obtain a fellowship, determine what opportunities will make you a desirable candidate in your chosen field. One important way to improve your ability to obtain a fellowship is to show that you are a productive researcher. There are many ways to show your productivity including developing research questions and hypotheses, overcoming difficulties in completing your projects, and sharing your work through presentations and publications. While your research interest might not be related to your chosen fellowship, you should devote some of your time to read papers and attend meetings in your chosen specialty. Networking with future colleagues and mentors could help you learn about fellowship and career opportunities.

When looking forward to your career in academic surgery, there are some additional skills that might help you achieve your future goals. Becoming a funded researcher is difficult with few National Institutes of Health grants being funded each year [4]. Writing or reviewing a grant or mock grant during your research

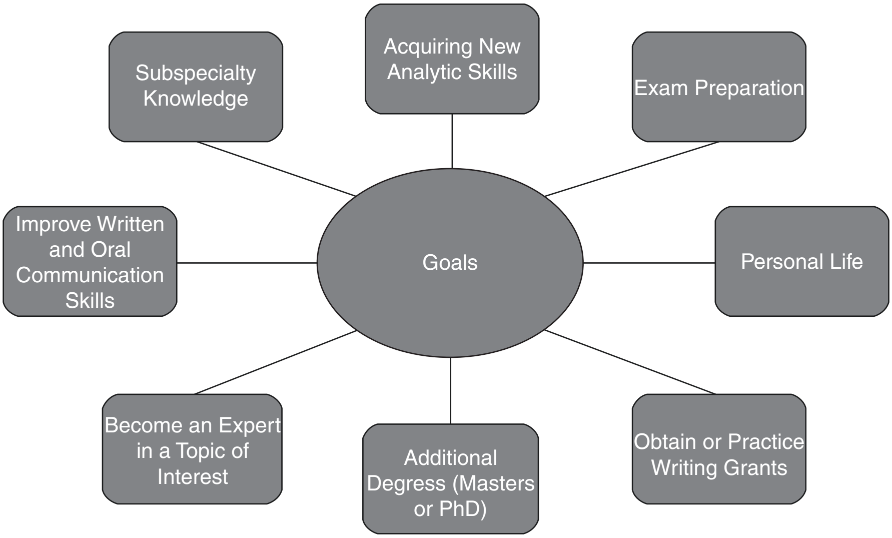

Subspecialty Knowledge
Acquiring New Analytic Skills
Exam Preparation
Improve Written and Oral Communication Skills
Goals
Personal Life
Become an Expert in a Topic of Interest
Additional Degress (Masters or PhD)
Obtain or Practice Writing Grants

Fig. 27.1 Identifying your research goals

time will help you better understand the time and support this requires. Masters and doctorate level course work could help you develop an expertise in quantitative and qualitative methods that might benefit your career long term. Health services researchers might consider course work in study design and specific analytic methods. Propensity score development, regression analysis (linear, logisitic, and longitudinal), survey design, and decision analysis are examples of topics that might improve your ability to perform and explain health services research. Discussing the support and resources your mentors found helpful when starting their careers can help you succeed when beginning your own independent research career.

Finally, taking time for your personal goals during this break in your clinical schedule is essential. Deciding how you are going to stay healthy in your personal life while balancing a busy career is difficult. Making time for family, friends, and your personal life might improve your ability to focus when working on your professional goals and work satisfaction, as well [5].

27.3 Identify Mentors Who Can Support Your Goals

When you meet with a potential mentor you should discuss your career and research interests. Taking the time to meet with several potential mentors will help you find someone who is the best fit for your goals, interests and personality. Some considerations when identifying your mentor might be: their technical or analytics background, their area of research or clinical interest, and their experience in guiding other mentees along a successful career path.

27.3.1 Current and Past Trainees

When you are considering joining a research group or laboratory, take time to talk to current or past mentees of that mentor. They can give you invaluable information by describing their experience. Some useful information to ask includes what meetings they attended, presentations they gave, number of manuscripts they plan to write, and what support they had from the research group.

27.3.2 Resources

When considering where you are going to perform your research, you want to evaluate what resources will be available to you. Understanding the expectations of your mentor will also help you set more realistic goals. In health services research you should consider:

- Databases available and how they are accessed
- Expectations for performing data entry (i.e. retrospective chart reviews)

- Assistance with complicated datasets (i.e. availability of statisticians or personnel for writing computer code)
- Statistical software availability

27.4 Deciding on a Research Topic

When deciding on a research topic you should consider topics in which you want to become an expert. Consider topics you are passionate about, those relevant to your future career, and current topics of interest. Having a personal investment in the projects you chose will press you to learn more about the topic and share your findings with others. When trying to find inspiration for your research, ask yourself what experiences in your clinical work interest you and can impact patient care. For example, you might be interested in evaluating the impact of different health care policies, patient socioeconomic status, or other cultural issues that affect a particular patient population.

Once you decide your area of interest, invest time in reading about the topic. Reading about a topic will help you formulate pointed research questions to improve upon current knowledge in that area. It is also important to understand if the findings from your research support or contradict previously published work. Becoming an expert in a particular topic requires time and focus, and will highlight your depth of interest in that specific topic.

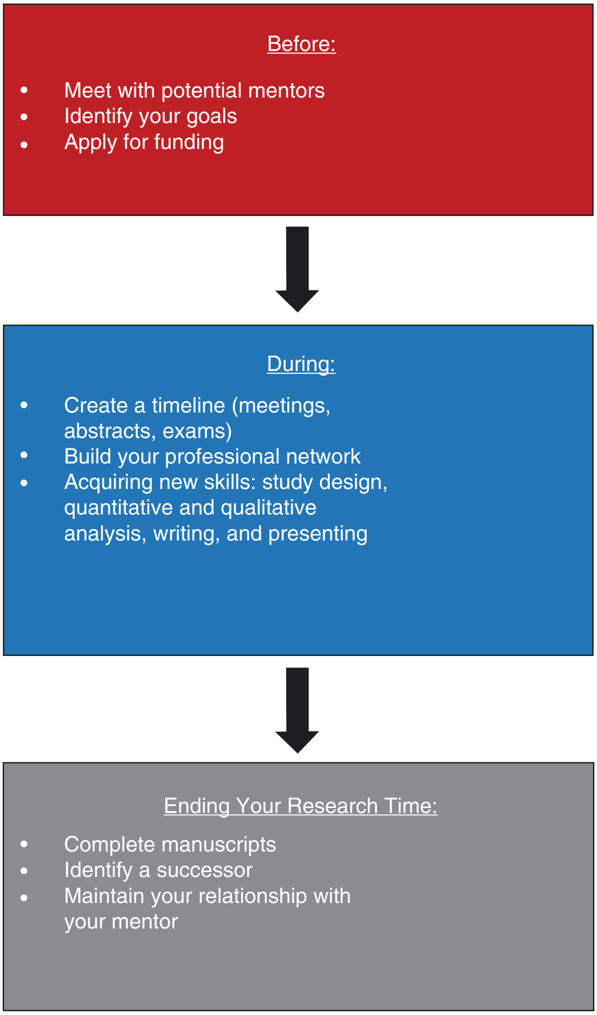
27.5 Completing a Research Project

Taking a research project from an idea to a published manuscript can be a daunting task. Information about specific study design and statistical methods for analysis are outside the scope of this chapter; however, there are a few key elements you should learn about when planning a research project. The first step is developing a hypothesis that can be tested. Once you have perfected your research question, you need to consider what study design is feasible to test your hypothesis. Consider what databases already exist that might answer your research question. If no data currently exists, evaluate the best study design to obtain your data with the time, funding, and resources you have available.

Once you obtain your data, take the time to review it for missing values and validity before you start any analysis. Talk to a statistician if you are unsure what tests are best for analyzing your data. After you run your analysis, review the results with your research group. Ensure your findings make sense both statistically and clinically. Be clear and concise when you state your findings and do not overstate your conclusions.

27.6 Create a Timeline

The time you spend focusing on research is likely to be less structured than your busy clinical days of surgical residency. Making a plan for your research time will keep you focused on achieving your goals. When planning your research time you should consider three time points: (1) before your research time begins, (2) during your research time, and (3) leaving your research time (see Fig. 27.2).



Before:

- Meet with potential mentors
- Identify your goals
- Apply for funding

During:

- Create a timeline (meetings, abstracts, exams)
- Build your professional network
- Acquiring new skills: study design, quantitative and qualitative analysis, writing, and presenting

Ending Your Research Time:

- Complete manuscripts
- Identify a successor
- Maintain your relationship with your mentor

Fig. 27.2 Planning your research time

27.6.1 Before You Start Your Research Time

Meeting with potential mentors at least a year before you enter your research time will allow you the time to identify a mentor that is right for you. While a year in advance might seem early, this timing is important if you plan to apply for funding. Establishing funding for yourself helps to offset the costs to your department or mentor and is also a great addition to your resume. There are several funding sources that award fellowships and grants for resident research. Applications are often due months before the resident's planned start date. These applications often require identification of a topic of research and letters of support from supervising mentors. By identifying a research project early you can begin reading about your area of interest. This will help you understand your project better and ease your transition.

27.6.2 During Research Time

Research requires patience and persistence. To make the most of their research time, many residents aim to work on a variety of projects. However, it is important to consider what you can complete and your personal priorities. Creating a plan for your weeks, months, and year(s) in research will help you focus on developing the skills you need to achieve your goals. When considering your schedule, you should pay attention to the timing of conferences, abstract deadlines, in service exams, and vacations.

Be realistic about what you can complete with the time, resources, and experience you have. While 69% of general surgery residents publish at least 1 paper during their research time, only 18% publish 3 or more papers [1]. Setting deadlines for sharing your work through conferences or research meetings can provide a useful timeline. Share your work with your mentor well in advance of deadlines to ensure they have time to provide you feedback for revisions, as well.

Residents with limited experience in research should focus on completing one or two projects rather than working on several projects that might never be completed. A more experienced resident researcher will likely work on several research projects during their research time. Keeping a project list can help organize multiple projects and ensure that deadlines are met. For researchers who plan to take course work, consider planning projects that coincide with techniques you will learn throughout your course work.

Consistently evaluating and analyzing journal articles and presenting your own work will improve the quality of the research you produce. Before agreeing to a new research project, you might consider how the project can develop your fund of knowledge and experience. Working on diverse projects with different study designs or statistical analysis can improve your understanding of research and medical literature. If a mentor asks you to work on a project that you do not have time to complete, consider suggesting a colleague who could benefit from working on that project.

27.6.3 Ending Your Time with a Research Group

As the end of your research time approaches, you should consider what projects will be left incomplete. The rigorous schedule of clinical residency makes it difficult to continue working on research, even for highly motivated residents. Consider what projects you can complete and those you plan to pass along to residents entering or remaining in your research group. By identifying a research successor you can help them transition into their research role as your research time comes to an end. When you leave a research group, ensure that your colleagues have your new contact information and stay in touch. Often, projects will continue or manuscript reviews will return with requests for revisions. Any data or information that you might need for completing this work should be accessible to yourself or your colleagues to ensure this work can be completed in a timely manner.

27.7 Special Considerations for International Graduates

International graduates who are planning to join a residency or obtain further training can benefit greatly from research experience. Devoting time to research shows your interest in the medical field, your initiative in working on projects and following them through to completion, and provides you with contacts that can provide valuable letters of recommendation. When identifying a potential research experience make sure you can articulate your current skills, such as medical record review, statistics training, basic science/experiment experience, manuscript preparation, or presentations. As discussed above, you should identify your goals for your research time and share these with your mentor. Consider what time you will need to study for certification examinations and set a realistic timeline. By making your mentors aware of your goals, they can be understanding and supportive of your other time commitments.

27.8 Conclusion

Being successful during your research time depends on your own personal and professional goals. Once you determine your goals, surround yourself with the mentors and resources you need to achieve them. By creating a realistic plan for your time and making the most of available opportunities you can develop the skills and relationships you need to build a successful career.

References

1. Stewart RD, Doyle J, Lollis SS, et al. Surgical resident research in New England. Arch Surg. 2000;135(4):439–44.
2. Robertson CM, Klingensmith ME, Coopersmith CM. Prevalence and cost of full-time research fellowships during general surgery residency: a national survey. Ann Surg. 2009;249(1):155–61.

3. Elliott ST, Lee ES. Surgical resident research productivity over 16 years. J Surg Res. 2009;153(1):148–51.
4. Rockey S. What are the Chances of Getting Funded? https://nexus.od.nih.gov/all/2015/06/29/what-are-the-chances-of-getting-funded/; National Institutes of Health: Extramural Nexus; 2015 [cited 2019 January 28, 2019]. https://nexus.od.nih.gov/all/2015/06/29/what-are-the-chances-of-getting-funded/
5. Sullivan MC, Yeo H, Roman SA, et al. Striving for work-life balance: effect of marriage and children on the experience of 4402 US general surgery residents. Ann Surg. 2013 Mar;257(3):571–6.

Funding Opportunities for Outcomes Research

28

Dorry Segev

Abstract

"It's All About the Benjamins"

(Sean Combs a.k.a. Puff Daddy, June 30, 1997)

It is critical for any discussion of research to address funding; four major reasons come to mind. First, high-quality outcomes research is not free. Contrary to the popular belief, outcomes research is not something you do on your laptop while watching Netflix; making a tangible difference in policy or patient care is a complex endeavor that requires time, expertise, collaborators, data, computing, and often patient engagement, and all of these things cost money. Second, few of us work in an environment where our clinical margin can finance our research, so money from outside of our clinical practice is necessary. Third, funded research is highly valuable to institutions, mostly because of prestige and indirects (the 25–75 cents or so in facilities and administrative fees that are paid to the institution for every dollar of research funds awarded to the PI). Fourth (likely as a result of third), funded researchers are highly respected in academic institutions; in fact, research funding is often a criterion for promotion, bonuses, etc.

This chapter will address what line items are commonly found in budgets of outcomes research grants (What Costs Money), various sources of research funding including government, societies, foundations, and other less-traditional sources (Who Has the Money), the types of grants that are funded, such as career development versus research grants, and their target audiences (Who Gets the Money), an overview of the NIH grant review and funding process (The Road to Riches), and some grant writing advice (Selling the Drama). Clearly, a handful

D. Segev (✉)
Department of Epidemiology, Johns Hopkins Medical Institutions,
Baltimore, MD, USA

Department of Surgery, Johns Hopkins Medical Institutions, Baltimore, MD, USA
e-mail: dorry@jhmi.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_28

of pages cannot begin to cover all of the details and advice that an investigator needs to be the Puff Daddy of research funding; but hopefully this will serve as a starting point, beyond which the reader is advised to identify one or more well-funded investigators with a track record of facilitating this process for their mentees.

28.1 What Costs Money

Young outcomes researchers often ask me if "the kind of work we do is even fundable." In other words, it seems intuitive to anyone that laboratory science would need funding: you need reagents, lab equipment, technicians, mice, cells, etc. It might seem that "just crunching data" would not follow this model and, as such, would not lend itself to traditional funding mechanisms. However, an overview of the typical expenses found in outcomes research (and reassurance that, yes, the NIH considers these to be "viable expenses") reminds us that the laboratory funding model is not so different from outcomes research.

Even if the research does not involve patient interactions, "crunching data" is expensive. Even the data themselves can be expensive, and funding agencies do not expect that you have already paid for data. Some datasets (such as claims or pharmacy data) can cost hundreds of thousands of dollars (although hopefully you can negotiate a more reasonable academic cost, often by sparking the intellectual interests of those selling the data). Additionally, dataset linkages are also time-consuming and fundable. Demonstrating feasibility and potential effects in either a small subset of the main data, or in data from a different source with similar structure, can suffice for preliminary data; there is certainly no requirement to have conducted the entire study before applying for the grant. No matter the source, data arrive dirty, and require cleaning and extensive exploration to ensure high-quality data prior to the primary analysis; again, pilot data can be derived from a subset or a different source, and in either case the main data still require this work, which requires the time and effort of a research assistant or an analyst (or both).

The analysis itself is often complex, requires computers and statistical software (you might have the software, but the licenses might require updating) and, most importantly, personnel. For an analysis to be reliable, redundancy is likely required. To ensure the best methods are used, analysts are required but not sufficient; faculty (typically from departments of biostatistics and/or epidemiology) with extensive, published experience in methods relevant to the science are critical, and must contribute enough effort (at least 1.2 calendar-months per year) to demonstrate full engagement with the research team. Since a clinical understanding is required to inform exploration of the data and the analytical approach, substantial effort from the PI and/or clinical experts is also required.

If patients are involved, expenses add up quite quickly. These can include patient incentives to participate in the research (gift cards, other tokens, meals) as well as expense reimbursement (travel, parking). Research assistants are often required to collect the data directly from the patients, abstract data from medical records, and enter data into whatever computing system has been established for capturing the data. The data collection system is also an expense of the research; often, the pilot study uses a more rudimentary data collection system, and this is expanded once funding is secured. As with data analysis, data collection does not just involve those collecting the data, but requires supervision, redundancy, and faculty collaborators with extensive experience in conducting human subjects research contributing enough effort to demonstrate engagement in the process of data collection design, subject recruitment and retention, and protection of human subjects.

28.2 Who Has the Money

The holy grail of medical research support is the NIH. Readers to whom this statement is a surprise are encouraged to rethink their interest in research. Not only does the NIH have the largest budget for medical research ($39 billion in FY2019), it has the process of grant review widely considered to be the most robust; as such, achieving NIH funding, serving on an NIH study section, or even having the letters "NIH" on your license plate are considered prestigious in the academic community and highly valuable to promotions committees. NIH publishes a Weekly NIH Funding Opportunities and Notices newsletter that is well worth receiving through their free email subscription. The general type of grants offered by the NIH (Who Gets the Money) and the NIH grant review process (The Road to Riches) are discussed in more detail below. One major advantage of the NIH is that, in general, for the more well-known grants (such as K and parent R01), there is an established payline, and if you get the score you get the money, independent of subjective "areas of interest" that play a more dominant role in other funding sources.

However, it is also important to be aware of (and seek) alternative (or supplementary) sources of research funding. The following list illustrates a broad range of non-NIH funding opportunities but is undoubtedly incomplete, and the reader is encouraged to explore funding opportunities independently and more tailored to the individual project.

Three major government-based funding sources specific to medical research are the Agency for Healthcare Research and Quality (AHRQ), the Patient-Centered Outcomes Research Institute (PCORI), and the Health Resources and Services Administration (HRSA). AHRQ functions basically like a "mini-NIH" in terms of funding opportunities, grant review, and grant structure; the other agencies use their own mechanisms. The Centers for Disease Control and Prevention (CDC) also offer occasional disease-specific grant or contract opportunities, and the National Science Foundation (NSF) is a staple for science and engineering funding that might align well with engineering/medicine collaborations. Recently, the Center for Medicare and Medicaid Services (CMS) has introduced some interesting outcomes research and innovation funding opportunities as well. Finally, the Department of Defense funds a Congressionally Directed Medical Research Program for areas of medicine directly relevant to service members, their families, and other military beneficiaries.

Investigators who practice in the VA system (US Department of Veterans Affairs) have access to a rich source of intramural VA grants ranging from career development awards to large research awards. Those in the VA system are strongly encouraged to pursue these funding opportunities, but those outside of the VA system unfortunately need to look elsewhere.

A number of foundations offer research funding, mostly in the form of career development awards, but some also offer larger so-called “R01-equivalent” grants. These include, but are certainly not limited to, the Doris Duke Charitable Foundation, Burroughs Wellcome Fund, Robert Wood Johnson Foundation, and the American Federation for Aging Research. Disease-specific foundations also include the American Heart Association, American Stroke Association, Cancer Research Institute, American Gastroenterological Association, Juvenile Diabetes Research Foundation, National Pancreas Foundation, American Diabetes Association, and National Kidney Foundation. Remember to think “outside of the surgical box” about the diseases that require surgical intervention, and consider pursuing funding from groups that seek to better understand and treat these diseases. While some of these foundations might seem at first glance to be more “medical” (as opposed to surgical) or “basic science” (as opposed to patient-oriented research), there are many examples of surgical investigators funded by these foundations and associations.

Many professional societies also offer funding opportunities to their members; as above, some societies that might at first glance seem “medical” have funded surgical investigators and are well worth pursuing. Surgical societies include the American College of Surgeons, Association for Academic Surgery, and Society of University Surgeons; specialty surgical societies include the Society for Vascular Surgery, American Society of Transplant Surgeons, Society of Surgical Oncology, American Pediatric Surgical Association, and many others. Societies that might overlap with surgical research also include the American Cancer Society, American Geriatrics Society, American Society of Nephrology, American Association for the Study of Liver Diseases, and others.

Other funding sources certainly exist but might be less formalized; the lack of formality should not discourage the enthusiastic and creative applicant, as inspiring tales of riches and glory can be found by those who seek them. Most institutions have internal career development and seed grants which are, strangely, often well-kept secrets (hopefully less so after the writing of this chapter). Some investigators have forged successful collaborations with insurance companies, state Medicaid, or the “hospital side” of institutions to conduct health services research that benefits those who pay for (or endure the cost of) medical care; the overlap between safety, quality improvement, and outcomes research can often be a strong point of leverage. Finally, grateful patients and other philanthropic efforts can support the most exciting, fast-paced, high-risk projects that are usually also the most high-impact but (unfortunately but not surprisingly) least fundable through traditional methods.

28.3 Who Gets the Money

Medical research funding begins with the larval medical/graduate student and extends through the furthest reaches of full professorship. The PI of a research group should strive for extramural funding across this spectrum; not only does a broad portfolio reflect well on a research group and develop a culture of grant writing and funding, but it also provides a track record and foundation of success for those junior investigators able to secure such funds.

Most grants fall into two major categories: career development awards, where the unit of funding is the individual (who, of course, proposes to conduct some type of research), and research awards, where the unit of funding is the proposed research (which, of course, will be conducted by a group of individuals). More technically, a proposal for a research award is typically based around a set of specific research aims, while a proposal for a career development award adds a layer of training and/or career development to these research aims. This section will explore these categories in the context of NIH awards, but similar patterns will be seen in other government, foundation, and society mechanisms.

28.3.1 F Awards

The "F" stands for *fellowship*, but this mechanism includes both predoctoral awards (such as F30 student awards) and postdoctoral awards (such as F32 postdoctoral fellowships). As a family, the F awards are also known as the National Research Service Award, or NRSAs. For medical students pursuing MD/PhDs, the F30 is an amazing mechanism that will basically pay for all of medical and graduate school tuition, provide a stipend, and even provide some research resources. For surgical residents spending "time in the lab" between clinical years, the F32 is an ideal mechanism, particularly for those seeking graduate degrees; similar to the F30, it provides a stipend (although most likely less than clinical PGY-based salary), research resources, and tuition support for a graduate degree. For those who are unsuccessful in obtaining an F32, some institutions hold T32 grants which are research area-specific and provide similar support while administered through the institution (with or without a formal application process). Experienced researchers who frequently mentor F-level trainees are encouraged to collaborate with the goal of establishing T32 mechanisms if these are not already available at the institution.

28.3.2 K Awards

The "K" stands for *career development* (apparently nobody wants "C-awards"?) and in general supports *mentored* research training at the instructor or assistant professor level (such as K01 for PhD-trained investigators, K08 for MD-trained investigators conducting non-human subjects research including laboratory science or secondary data analysis, and K23 for MD-trained investigators conducting patient-oriented research) and *mentoring* at the associate professor level (such as the K24 mid-career investigator award). Those with established, R01-funded research groups who actively mentor might consider the K24 to support this endeavor. However, "K awards" generally refer to the first step on the faculty pathway to independent research: a junior faculty member starts with some departmental or institutional funding, obtains a K08 or K23, and hopefully eventually bridges the "K to R" (see R awards below) transition for independent research funding.

28.3.3 R Awards

The "R" stands for *research*, and these independent research awards include the holy grail of all that is research funding in the universe, the R01, as well as smaller research grants (R03 and R21). The R01 is not only the holy grail because of its resources (in general, up to $500,000 per year for up to 5 years, with the possibility for competitive renewal for many years thereafter) and scope (any topic acceptable to an institute and interesting to a study section), but also because of the consistency with which funding is correlated to study section score and the advantages offered to early stage and new investigators (see The Road to Riches below).

28.3.4 P and U Awards

The "P" stands for *program project*, and these include the P01 which is basically several integrated R01-scope projects "involving a number of independent investigators who share knowledge and common resources." The "U" has an unclear etymology (at least unclear to this author), comprised mostly of the U01 Cooperative Agreements which are NIH-administered multi-center studies where each center applies for the funding required for their contribution. In general, the only readers of this chapter for whom the P and U awards are appropriate as PI are those readers already familiar with them.

28.4 The Road to Riches

The NIH pathway to funding is complex but well worth understanding for the sake of strategy and sanity. The NIH is divided into a number of disease-specific *institutes*, such as NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases), NCI (National Cancer Institute), NIA (National Institute on Aging), and 19 others. The institutes fund the grant applications (which hereafter we will refer to as "grants" for convenience and consistency with colloquial precedent), but do not necessarily review the grants. There are hundreds of *study sections* that review grants, some of which belong to the individual institutes, some of which belong to the Center for Scientific Review (CSR, another branch of the NIH), and some of which are ad-hoc special emphasis panels.

In general, a grant is submitted through a funding opportunity announcement (FOA, more below), "accepted" by an institute -- acknowledging that the topic is within their mission and that they participate in the particular FOA (this is generally pretty straightforward) -- and reviewed and given a score (with or without a study section-specific percentile) by a study section. The institute evaluates the score and/or percentile in the context of (1) the rules of the FOA (including the institute's established payline for R01s), (2) the relative scores of other applications, (3) the financial situation of the FOA (i.e. if money was set aside for this particular opportunity), and (4) the financial situation for the institute (how much money they were appropriated and how much they have spent).

The most important distinction is between who *reviews* the grants and who *funds* the grants. The study section reviews the grant, while the institute funds the grant. In the case of most R-grant FOAs (R01s, R03s, R21s), the study section is drawn from the CSR. The investigator can request a study section, and in general if this is an appropriate request it will be granted; even if a study section request is not initially granted, there is an appeals process which is occasionally successful and well worth trying (after discussing with experienced mentors). As such, investigators applying to most R grants are strongly encouraged to research the CSR study sections both online (the topic areas and rosters are public information) and through word-of-mouth. Some R-grants, and most F/K/P/U grants, are reviewed by a study section that belongs to the institute rather than the CSR (either institute-specific standing study sections, or ad-hoc special emphasis panels); in these cases, investigator request is not applicable.

In general, the R-grants labeled "PA-xx" have no money assigned to them, are evaluated by CSR, and draw from the institute's general budget. Most institutes have "parent" FOAs for the R-grants that are not disease-specific (other than the mission of the institute) and can be used for all investigator-initiated ideas. Some PAs are specific to a disease or a type of research; these still do not have money assigned to them, but they indicate a priority for the institute; at the end of the day, this just means that if the score is borderline (i.e. not quite under the payline), they *might* be able to use discretionary money to fund it. R-grants labeled "PAR-xx" are similar to PAs except they are usually more targeted and more than often use a special emphasis panel for review. Finally, R-grants labeled "RFA-xx" have actual money assigned to them (that is separate from the institute's general budget) and are usually reviewed by a study section that belongs to (or was ad-hoc created for) the institute and/or the FOA.

The R01 has the most comprehensible and predictable (if such is possible with the NIH) funding pathway. Other than certain RFAs, PARs, or other unusual circumstances, an R01 receives a score from the study section, is assigned a percentile based on the recent distribution of scores (usually specific to that study section), and is selected for funding if the percentile is below the payline for that fiscal year. Each institute establishes and publishes a payline, so interpreting the percentile generally does not require the statistical rigor that your grant will require. Most institutes establish at least two paylines, one for established investigators and one for *new investigators*, i.e. those who have never received an R01 or R01-equivalent from the NIH. Some institutes establish a third payline for *early stage investigators* who are not only new investigators but are within 10 years of their final degree (rumor has it that clinicians can make a compelling case that residency or fellowship was part of their training, and as such the clock would start after all clinical training). Sometimes an institute will also establish a formal payline (either by percentile or score) for other mechanisms (such as R21s, K awards, etc.) but these are not nearly as consistent or predictable as the R01 paylines, so conversations with the NIH staff are required to make sense of the scores and fundability.

Two recent changes to NIH applications deserve mention. First, all grants are now divided into those involving or not involving a clinical trial; clinical trial applications require extra paperwork (because a 200 page application is not nearly enough work), and each FOA specifies if it accepts clinical trials or not. Second, investigators can no longer receive more than one grant as ESI (a method that, in the past, allowed investigators with more than one hot early-career idea and the drive to prepare more than one application at a time to take extra advantage of the ESI paylines). Of course, the NIH is always changing and it is critical to stay up-to-date on NIH policies.

28.5 "Selling the Drama"

(Ed Kowalczyk of *Live*, April 26, 1994)

It is well beyond the scope of this chapter to describe the ninja skills required to identify an important, innovative scientific endeavor and present it in a compelling way to a room full of critical, likely somewhat cynical individuals who review about 10× as many grants as are funded. The skills are also specific to the research approach: for example, review of a secondary analysis might focus on the analytical approach and data quality, while that of patient-oriented research might focus on recruitment feasibility and measurement error. A good place to start might be an institutional or society grant writing course, and careful review of many already-submitted grants from your colleagues.

In general terms, it is important to remember that most grant applications come with very detailed instructions, and it is actually important to read these instructions carefully. Violations of page limits, margins, font sizes, organization of the grant, letters of support or recommendation, and other seemingly trivial issues are easy ways to get a grant rejected before it even has the chance to take up a study section's time. Similarly, in light of the number of applications that each study section member has to read, it is important to make it easy for them to like the application, both in terms of formatting (organize thoughts into clearly labeled subsections, leave space, use figures and tables to break the monotony of text) and content (explain the importance of the research, support the feasibility and likely success of the proposed science with preliminary data, clearly describe the approach so that the reviewers can actually imagine how the research will be conducted, and identify potential problems that might arise and how they might be addressed). Finally, the overview (the specific aims page of an NIH grant, for example) is critical and must "sell" the grant; many reviewers will not read past a poor overview (usually because they have 50 other grants to read), and those who do will already have a bad impression of the grant which is likely not recoverable.

A research grant (R-grant in NIH terms) involves the identification of a significant knowledge gap in the field, an approach likely to address this knowledge gap, and a prediction of how the findings might affect patient care (or policy). A career development grant (K-grant in NIH terms) involves these elements but with a (very important) parallel layer for the investigator in training: the significant knowledge gap in the field must parallel a knowledge (training) gap in the investigator, the approach likely to address the knowledge gap must also teach the investigator a new set of skills, and the prediction of how the findings might affect patient care must parallel short-term and long-term career goals for the investigator. The career development and training layer of a career development grant should *not* be taken lightly, and is more often than not the Achilles' heel of the grant.

Examples of successful grants abound, and investigators are encouraged to learn from these examples. Almost every funding source publishes a list of previous recipients, and many grant recipients (particularly those in one's own institution, field, or collaborative network) are willing to share parts of their applications. The NIH goes one step further with NIH RePORTER (Research Portfolio Online Reporting Tools), a database of all funded grants, searchable by name, institution, department, topic, mechanism, and other characteristics; most importantly, the database also includes the narrative for each application which usually lists the specific aims and other critical elements. A read through NIH Reporter is a worthwhile lesson in what topics and approaches are "fundable", how research can be framed in a compelling manner, and how to align a work scope with a funding mechanism. Furthermore, a creative use for this resource is the identification of mentors and collaborators; for example, those writing K-grants can search their own institution for R01-funded investigators with expertise in the areas where they are hoping to train.

28.6 "You See What You Look For"

(Stephen Sondheim, April 26, 1970)

The research funding environment is increasingly competitive and frustrating. That said, there are still billions of dollars out there for research funding (that's a lot of Benjamins), and the emphasis on research directly applicable to patient care and policy (such as outcomes and other health services research) is growing. The classic NIH pathway of F32 to K23 to R01 is still very feasible for those with persistence, mentorship, institutional support, and a research trajectory that lends itself to NIH funding. Others will find success through foundations, philanthropy, and other creative funding pursuits. Most everyone will have more failures than successes, most grants will require resubmission, and most successfully funded researchers submit multiple grants every year and seem almost more proud of their failures than their successes. In the context of these somewhat painful but still hopeful realities, let us never forget that the only grant with no chance of success is the grant not submitted.

Principles of Writing K and R Grants

29

Rachel E. Patzer

Abstract

The ability to obtain extramural funding, including National Institutes of Health (NIH) grants, is critical for any health services researcher. Developing and then maintaining a health services research laboratory requires funding to conduct the research and/or the hiring of core team members who will help to conduct the research. Grant writing is an important skill. Researchers must not only have novel and innovative ideas, but also communicate these ideas in a clear and concise manner while following a particular format that grant reviewers expect.

The purpose of this chapter is to summarize the important considerations for writing career development (K-level) and research project (R-level) grants for federal institutes relevant to health services research, with a focus on the NIH. Sections will explore tips for obtaining health services research funding at the NIH, and selecting HSR-related study sections. The chapter will also detail the specific sections that must be written for K and R grants (e.g., specific aims, significance, innovation, approach) and sections unique to K grants (e.g., candidate background, mentor statement, etc.). This chapter will also provide tips for addressing common pitfalls of health services research grants and general strategies for grant writing.

29.1 What Federal Institutes Fund Health Services Research?

According to the National Academy of Medicine's 2018 publication "The Future of Health Services Research" health services research funding exhibited a 4.6% compound annual growth rate between 2004 and 2011; this funding is aggregated from

R. E. Patzer (✉)
Departments of Surgery & Medicine, Health Services Research Center, Emory University School of Medicine, Atlanta, GA, USA
e-mail: rpatzer@emory.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_29

sources including the National Institutes of Health (NIH), health services industry, the Agency for Healthcare Research and Quality (AHRQ), foundations, and other federal funding sources. The total annual funding estimate for health services research was ~$2.5 billion in fiscal year 2017, with NIH funding representing approximately two-thirds of total funding. However, this is still a minority of the ~$37 billion funded by NIH annually. AHRQ, an NIH-independent federal agency within the Department of Health and Human Services, is typically considered the home of health services research. However, AHRQ's funding allotment has declined across its 30+ year history and remains miniscule (~$416.6 million in 2017) compared to the NIH [1]. In addition, several federal proposals have been presented to the Department of Health and Human Services to eliminate AHRQ or consolidate the agency within the NIH or other institutes. For example, upon passing of the Affordable Care Act, AHRQ's prior responsibility for conducting comparative effectiveness research was reassigned to the newly formed Patient Centered Outcomes Research Institute (PCORI) in 2012 [2]. This chapter primarily focuses on NIH grant funding because it is the primary funder of health services research, but many of the same tips will apply to other funding sources.

While the majority of the 27 institutes comprising the NIH fund health services research proposals, some institutes are more engaged in funding health services research than others. Each institute has a strategic plan, and institutes with plans that specifically mention health services research are the most promising to target for the purposes of grant funding. For example, the National Institutes of Minority Health and Health Disparities (NIMHD) includes a vision to raise "national awareness about the prevalence and impact of health disparities and disseminate effective individual-, community-, and population-level interventions to reduce and encourage elimination of health disparities," which emphasizes health services research approaches to addressing health disparities. Other institutes also have a portfolio of funded research that includes health services research [3].

Selecting an NIH institute may be simple if you are targeting a specific patient population (e.g., cancer—submit to the National Cancer Institute (NCI); kidney disease—submit to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), etc.). However, if you are targeting a health system (e.g., hospitals, emergency rooms, etc.) that treat a diverse group of patients, you will need to put more thought into which institute may be the most appropriate. Start by reviewing the various missions of each institute and potentially target one that is not disease-specific. For example, despite a broad population, you could narrow your research focus to an older population and target the National Institute of Aging. It is also a good idea to examine funding paylines (funding cutoff points for grant applications) by each institute, grant type (e.g. K01 vs. R01), and the number of awarded grants per institute. While not all institutes publish paylines, it is a component worth exploring because substantial variability exists in funding competitiveness by institute (e.g., in FY2016 the National Heart, Lung, and Blood Institute (NHLBI) funded 48% of K01 grants vs. 19.2% at NCI). It may be in your best interest to select the institute with the best payline to increase your odds of funding. It is also important to note that success rates vary from year to year, so applicants may want to exercise caution in solely using these data to drive their submissions.

29.2 Selecting a Study Section

Regardless of the NIH institute you target, it is worthwhile to request your preferred study section in your application. You will want to request that your grant application is reviewed by health services researchers who understand and value the importance of the field. In the cover letter to your grant submission, you can specifically request the study section of your choice. Although there is no guarantee your request will be honored, it may be particularly important to request a study section if you are submitting to an institute that typically only funds a small portion of health services research or whose mission does not specifically state its commitment to the field. If the methodology in your proposal is highly specialized, such as microsimulation modeling, you can also request reviewers with that particular expertise.

A complete list of NIH study sections is available on the Center for Scientific Review website (in addition to other useful resources about the grant review process) [4]. There are a variety of relevant health services research study sections, which are primarily clustered under the Healthcare Delivery and Methodologies (HDM) Integrated Review Group (IRG). The HDM IRG is a multidisciplinary review group that typically consists of a wide range of clinical, community, and population research studies that "examine the sociocultural, economic, environmental, and infrastructure factors that influence the delivery of health care, patient outcomes, population health, and health policies." The topics vary from acute to chronic health issues, and may focus on the individual patient, communities, populations, families, or health systems. The HDM study sections also review studies that propose new research methodology, such as novel analytic approaches relevant to health services, behavioral, and social science research. Within HDM, study sections include those detailed in Fig. 29.1. A complete list and description of study sections can be found online (https://public.csr.nih.gov/StudySections).

29.3 Identifying the Appropriate Funding Opportunity

Once you identify the appropriate NIH institute (or institutes) to target your proposed research, you will need to identify which grant mechanism is the most appropriate for you. This depends on the stage of career you are in (e.g., early, mid, or senior-level) as well as the stage of your research (i.e. how much preliminary data you have). There are NIH awards available for each stage in your career (T32, F32 for pre- and post-doctoral fellows, K-level mentored career development awards, mid-career transition awards, and investigator-initiated R-level grants) (Fig. 29.2). Each institute will have a list of current funding opportunities posted to their website. These will likely include program announcements (PA) that designate priority areas for particular funding mechanisms in specific areas of science and research funding announcements (RFAs)

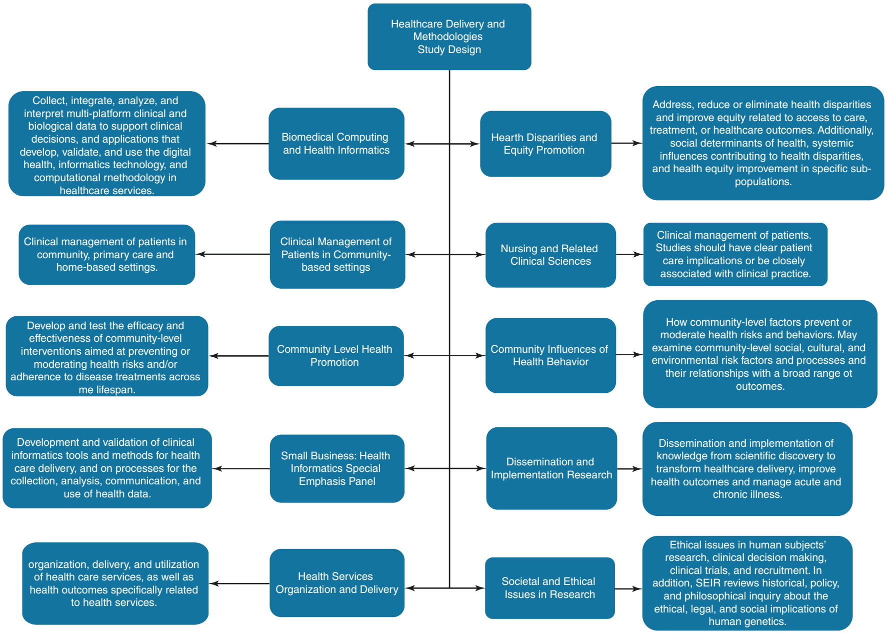

Healthcare Delivery and Methodologies Study Design
Collect, integrate, analyze, and interpret multi-platform clinical and biological data to support clinical decisions, and applications that develop, validate, and use the digital health, informatics technology, and computational rnethodology in healthcare services.
Biomedical Computing and Health Informatics
Hearth Disparities and Equity Promotion
Address, reduce or eliminate health disparities and improve equity related to access to care, treatment, or healthcare outcomes. Additionally, social determinants of health, systemic influences contributing to health disparities, and health equity improvement in specific sub-populations.
Clinical management of patients in community, primary care and home-based settings.
Clinical Management of Patients in Community-based settings
Nursing and Related Clinical Sciences
Clinical management of patients. Studies should have clear patient care implications or be closely associated with clinical practice.
Develop and test the efficacy and effectiveness of community-level interventions aimed at preventing or moderating health risks and/or adherence to disease treatments across me lifespan.
Community Level Health Promotion
Community Influences of Health Behavior
How community-level factors prevent or moderate health risks and behaviors. May examine community-level social, cultural, and environmental risk factors and processes and their relationships with a broad range ot outcomes.
Development and validation of clinical informatics tools and methods for health care delivery, and on processes for the collection, analysis, communication, and use of health data.
Small Business: Health Informatics Special Emphasis Panel
Dissemination and Implementation Research
Dissemination and implementation of knowledge from scientific discovery to transform healthcare delivery, improve health outcomes and manage acute and chronic illness.
organization, delivery, and utilization of health care services, as well as health outcomes specifically related to health services.
Health Services Organization and Delivery
Societal and Ethical Issues in Research
Ethical issues in human subjects' research, clinical decision making, clinical trials, and recruitment. In addition, SEIR reviews historical, policy, and philosophical inquiry about the ethical, legal, and social implications of human genetics.

Fig. 29.1 Health services research study sections

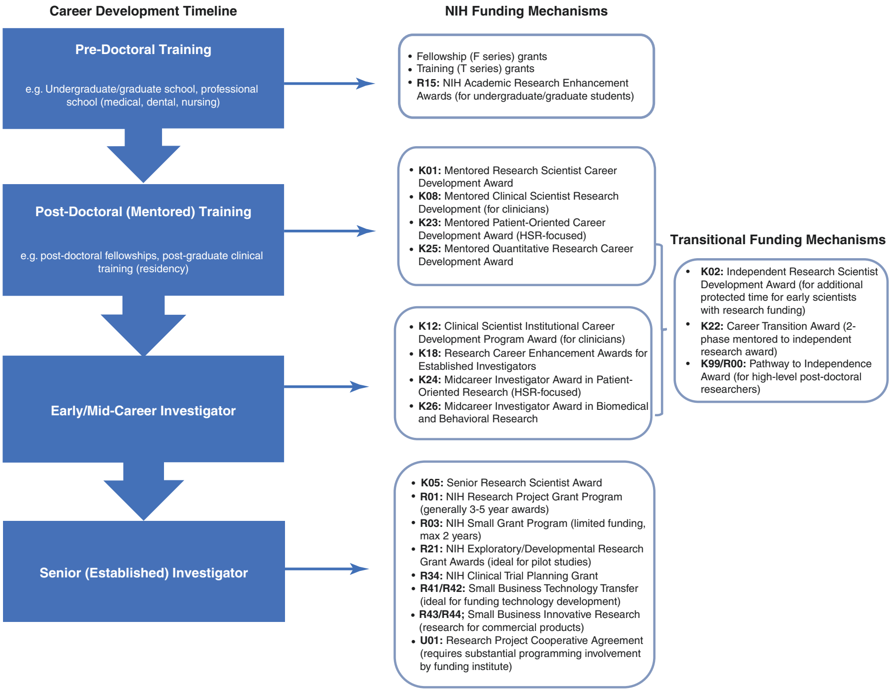

Career Development Timeline
Pre-Doctoral Training
e.g. Undergraduate/graduate school, professional school (medical, dental, nursing)
Post-Doctoral (Mentored) Training
e.g. post-doctoral fellowships, post-graduate clinical training (residency)
Early/Mid-Career Investigator
Senior (Established) Investigator
NIH Funding Mechanisms
• Fellowship (F series) grants
• Training (T series) grants
• R15: NIH Academic Research Enhancement Awards (for undergraduate/graduate students)
• K01: Mentored Research Scientist Career Development Award
• K08: Mentored Clinical Scientist Research Development (for clinicians)
• K23: Mentored Patient-Oriented Career Development Award (HSR-focused)
• K25: Mentored Quantitative Research Career Development Award
• K12: Clinical Scientist Institutional Career Development Program Award (for clinicians)
• K18: Research Career Enhancement Awards for Established Investigators
• K24: Midcareer Investigator Award in Patient-Oriented Research (HSR-focused)
• K26: Midcareer Investigator Award in Biomedical and Behavioral Research
• K05: Senior Research Scientist Award
• R01: NIH Research Project Grant Program (generally 3-5 year awards)
• R03: NIH Small Grant Program (limited funding, max 2 years)
• R21: NIH Exploratory/Developmental Research Grant Awards (ideal for pilot studies)
• R34: NIH Clinical Trial Planning Grant
• R41/R42: Small Business Technology Transfer (ideal for funding technology development)
• R43/R44; Small Business Innovative Research (research for commercial products)
• U01: Research Project Cooperative Agreement (requires substantial programming involvement by funding institute)
Transitional Funding Mechanisms
• K02: Independent Research Scientist Development Award (for additional protected time for early scientists with research funding)
• K22: Career Transition Award (2-phase mentored to independent research award)
• K99/R00: Pathway to Independence Award (for high-level post-doctoral researchers)

Fig. 29.2 Appropriate funding opportunities by career and research stage

or funding opportunity announcements (FOAs). PAs have standard due dates that are published on the NIH website, and RFAs typically have only one deadline. RFAs are ideal because they indicate that the institute has set aside a specific amount of money for work in this area. A weekly subscription to the NIH e-TOC (Table of Contents) is a useful way to stay informed about new grant opportunities and funding announcements. Always be sure to read the funding opportunity instructions carefully to determine programmatic fit and eligibility criteria. If you believe you are eligible and your proposed research is a good fit, start by sending an email to the designated program official (PO) listed on the application with a brief description of yourself and your proposed work; also request a phone call to speak with the PO. Include an updated biosketch (more below) in this email so the PO has background information about your experience. If you have draft specific aims at this point, you could also include your draft aims to help provide additional information. This conversation ideally occurs at least several months in advance of the deadline to help you determine the potential fit of your idea with the grant mechanism or opportunity. The PO can help to ensure that your topic fits within the mission of the institute, and s/he can also help to identify the appropriate study section to request and provide guidance on potential scope of the proposed work. PO's at the NIH are your advocate, are extremely experienced, and are usually very approachable and responsive.

29.4 General Strategies Before You Start Writing

Once you identify an appropriate target grant and have a deadline, contact your office of sponsored programs (OSP) or research administration services at your institution. OSP will have internal deadlines that are earlier than the stated grant mechanism deadline, and it will be helpful to work with your research administrators as early as possible to ensure you meet these deadlines. If you plan to work with other institutions or have subcontracts, your deadlines are likely even earlier (especially for finalizing the budget and personnel sections of the submission). Early on in the grant writing process, it is beneficial to set up a timeline as early as possible so that you can manage all of the grant components. This timeline should not only include a list of all main sections of the grant (e.g., Specific Aims, Significance, Approach, etc.) but also the other components that are required and can take substantial time to put together (e.g., budget, biosketches, planned enrollment tables, facilities and resources, protection of human subjects, letters of support, etc.). Note in the grant application instructions if there is a letter of intent deadline. While not always required, this should be submitted if it is recommended as it helps the institute understand the proposed work and give them sufficient time to identify qualified reviewers.

Another tip—if possible, try to identify an example of a previously funded grant to give you insight into expected format, content, and scope. For example, if you are targeting a career development award, find someone who (relatively recently) received this type of grant. It is an added bonus if the example grant is also from the same institute that you are targeting, or at least a health services research grant. The NIH has a particularly useful online resource called the Research Portfolio

Online Reporting Tools (RePORTER) that you can use to give you an idea of successfully funded grants [5]. You can even filter grants by study section or institute to see what topics were previously funded by each group. This may also be used as a mechanism to identify individuals within your own institution that have received grants from certain NIH institutes. Additionally, this tool can help you identify mechanisms that you may be able to approach to ask for a funded grant example or general advice about grant submissions.

29.5 Career Development Award—The K Grant

The purpose of the mentored career development award (e.g., NIH K grant) is to gain research expertise in a new area that will enhance your research capabilities. The career development award is intended to prepare you for an independent research career with formal guidance from a mentor and/or advisory team. NIH K awards provide protected time (typically 75%) from clinical, teaching, and/or administrative duties to focus on building your research portfolio and developing your skill sets (e.g., by learning new methods, taking coursework, etc.) to help you conduct your proposed research. It is important that you identify some kind of new training that you will receive. If you are new to the field but hope to become a health services researcher, an ideal example may be proposing to obtain a formal degree (e.g., Master's in Science of Clinical Research (MSCR) or a Master's in Public Health) during the protected time of the award to learn new skills (e.g., epidemiology, biostatistics, behavioral sciences, etc.). If you have a doctoral degree in these areas already, you could propose completing various training courses instead (e.g., NIH training institute for Dissemination and Implementation Research in Health, or a training institute on pragmatic trials to supplement other specific coursework that may be relevant to your proposed research). In your application, it is important to establish how this new training will advance your career as well as the science you propose.

The NIH website has a very helpful NIH K Research Training and Career Development component (formerly known as the K Kiosk) for identifying which K grant is appropriate given your background and career stage [6]. Typical K grants are mentored research scientist grants (K01) for those with a PhD, mentored clinician scientist (K08) grants for those with an MD, and mentored quantitative research development grants (K25) targeted for engineers or quantitative scientists. Of note for health services researchers is the K23, which is a mentored patient-oriented research career development award intended to support clinical investigators dedicated to patient-centered research. In addition, an excellent option for those who may not quite be ready for an NIH K grant is an institutional K (e.g., KL2) grant, if available at your institution. These grants are available through institutions that have a Clinical and Translational Science Institute, or CTSA, and provide 2 years of protected time and coursework (e.g. completion of an MSCR degree), with the expectation that candidates apply for an NIH career development (K) award in their second year of the program. Foundations or societies also may have career development grants targeted to junior faculty (e.g., American College of Surgeons research career development award, Society of University Surgeons, etc).

When contemplating whether a career development award is right for you, it is essential to ensure you meet eligibility criteria and that you have the support of your department. This is particularly important for surgeons who are 100% clinical and already in a faculty position and for whom department chairs must arrange a necessary cost share. Ideally, research time and intentions for an NIH K-grant are discussed prior to beginning a faculty position. Institutional support of the candidate, a scorable criterion for review, is important to show to reviewers. A letter of support from your department chair expressing commitment to providing protected time for a career development award is vital to your application.

There are restrictions in OSP to ensure eligibility. For example, some career development awards are restricted to US citizens or permanent residents, and many have explicit requirements on the maximum number of years from completion of post-doctoral training. Each NIH institute will have more details about maximum budget, eligibility, and instructions listed on the PA for each K mechanism. Other "requirements" for K awards may not be explicitly stated in the application but will most likely be critiqued in a review.

The scorable sections of most NIH career development grants include the Candidate, Career Development Plan/Career Goals & Objectives, the Research Plan, the Mentor(s), and the Environment and Institutional Commitment to the Candidate (Table 29.1). Many foundation or society career development awards require similar sections. It is important to gauge from your mentor and other experienced researchers in the field (preferably those who have reviewed career development awards) whether you are competitive for a K. For example, some K awards favor candidates who are in faculty positions. If you are a post-doctoral fellow applying for a K grant, most reviewers will look for a letter of support from your institution stating that you will have a faculty position that is ***not*** contingent upon receiving this grant. Positions such as Assistant Professor will be viewed as more competitive than staff scientist, even if you are technically eligible with either title. Your publication record will also be scrutinized; for example, reviewers will look to see if you have at least one publication on the topic you propose in your application and at least one publication with your proposed mentor and/or mentoring team.

Table 29.1 NIH review criteria for K vs. R grants

Career development (K-type) grant	Research project (R-level) grant
Candidate	Significance
Career Development Plan/Career Goals	Investigator(s)
Research Plan	Innovation
Mentors, Co-mentors, Collaborators, Consultants	Approach
Environment & Institutional Commitment to Candidate	Environment

29.6 Research Project Grant—The R01 Grant

The traditional path for researchers is to receive a K, and then an R-level grant. Typically, those with a career development award will apply for an R-level grant a few years into the career development award (or once they have published results from the K grant). Those who have completed their terminal research degree or post-graduate clinical training within the past 10 years and who have not previously competed successfully as a PI for an R01 or an R01 equivalent can qualify for early stage investigator (ESI) status. Of note, NIH will consider requests to extend ESI status for reasons such as family care responsibilities, disability, and other cases determined at the sole discretion of the NIH. The status should appear at the time of submission, so investigators should ensure that their status is correctly marked on their ERA Commons profile. Be careful if it is your first R01 submission and you submit as multi-PI (two PI's). You will only qualify if both PIs have ESI status. Of note, if you are successful in receiving the R01, you will lose your ESI status once awarded [7–9].

29.7 Importance of the Biosketch

Your NIH biosketch is critically important for any NIH grant and is accounted for in the scoring of grants (including the candidate section for career development grants, and the investigator(s) section for R-level grants). Instructions and examples are provided on the NIH website [10], but it may help to obtain an example from someone who has been previously awarded NIH funding. If this is your first grant application, ensure that your mentor takes the time to provide edits and critical feedback on your biosketch (and other sections of the grant, of course!). Your personal statement in the biosketch should detail why you are qualified to lead the project and receive the award (state the actual mechanism to personalize the biosketch, e.g. "I plan to use the training and experience from this K01 award to establish an independent career as a health services researcher focused on implementing evidence-based interventions to improve gaps in the care of surgical patients with complications, etc."). The personal statement should also include relevant aspects of your prior training, gaps in your current training that necessitate a career development award, and your prior experience in the field. You may also use this opportunity to describe any gaps or limitations in your past productivity (e.g., family obligations, active military service, illness, etc.). You can (and should) also highlight up to four relevant publications to the proposed research and, ideally, provide examples of prior publications with your mentor and/or mentoring team to show established collaborations. The personal statement should not exceed the first page of the biosketch, should be confident but not boastful, and should be succinct.

Other sections of the biosketch are also important (Positions and Honors, Contributions to Science, and Research Support), but reviewers will likely focus on the personal statement as a way to succinctly summarize some of the key aspects that are in the rest of the biosketch. It is critically important that, as the Principal

Investigator, you ensure the biosketches of your coinvestigators (and/or mentors) have also been personalized to match your grant and that each coinvestigator and/or mentor specifically states their role on the proposed research (e.g., “I will serve as the primary mentor for the candidate and oversee all aspects of the training and career development plans, as well as the research aims”).

29.8 K Grant: Candidate Background and Goals for Career Development

The candidate’s background is the first sub-section of the candidate section. Typically written in the first person, the background is meant to be a brief (typically one or two paragraphs) overview of your research and/or clinical training experience and your motivation for continuing this training/research with the career development award. This subsection is primarily a personal essay that also includes scientific information, your prior findings, and your research goals. The candidate information section will overlap with the biosketch personal statement, but can elaborate more on your personal experiences and how they have led you to become an ideal candidate for this career development grant. Highlight your current strengths, the needs of your research, and the training gaps that must be addressed to ensure you develop expertise to close those gaps and become an independent investigator.

The second sub-section of the candidate section is the section on career goals and objectives. This section is typically about 1 page in length. The first part of this section should summarize your past scientific history and how a potential career development award can help bridge your past and future career development. What are some consistent themes you can draw from your experiences that motivates you to continue your research? The next paragraph can describe your long-term goals. For example:

- My long-term career goal is to become an independent health services research investigator performing research to improve the dissemination and implementation of evidence-based interventions to improve access and outcomes of patients with end-organ failure requiring transplant surgery.

It is important to make clear in this section that this career development award will not just be a continuation of your post-doctoral work under your mentor; provide justification for how this award can help you grow into an independent investigator. It is possible you could be criticized in the review if your research is too similar to your mentor’s research. Your short-term training goals related to this career development grant should fit within your longer-term career goals, and can be stated in relation to the goals of this proposal (e.g. “specific objectives of my career development plan are to…”). This section should provide justification for the necessary skills or training you need to achieve your career goals. Successful candidate sections often clearly list the domains in which they intend to obtain more training. The most common number of domains to propose for new training is 3–5. For those pursuing a health services research career, this could include new training

Table 29.2 Example Table from Career Development Award: Summary of goals for career development

Proposed training domain	Prior training	K award scientific and training objectives	Long-term career objectives
Behavioral research methods	Introductory behavioral sciences class in Master's degree	To study the association of hospital volume and outcomes by surveying hospital leaders and….	Establish an independent research career in health services research
Dissemination and implementation research methods	No prior training	To develop expertise in the implementation of evidence-based interventions…	To implement evidence-based interventions in clinical practice
Scientific and grant writing	Published 3 first-authored papers; informal training from mentor	To improve technical skills in scientific and grant writing	To develop an R01 on…

in qualitative or quantitative methods, training in pragmatic trials, survey design, implementation and dissemination methods, methods in health policy and economics, etc. The domains should be relevant to your proposed research plan (e.g., if you propose to conduct a survey in your research proposal, a training domain in survey research is appropriate). Beyond specific methodology, training domains on scientific and grant writing and responsible conduct of research are largely applicable to most applications and should be included. This section can also answer how the specific training and career goals for this proposal will be able to help you accomplish your research questions. It could also help to provide a table so reviewers see a clear, concise summary of your career goals for this section (Table 29.2).

The third sub-section is the candidate's plan for career development/training activities during the award period. This section should include details on your mentoring team, as well as the specific training you will receive for each training domain you have proposed in your goals section. For example, you could start with **Structured Mentorship**, and summarize in 2–3 sentences the expertise that your mentorship team brings to complement your background, career goals, and proposed research aims. Next, provide detail of each mentor, starting with the lead mentor:

- **Dr. Jane Doe, MD, PhD (Primary Mentor)—**Dr. Doe is a Professor of Surgery and Epidemiology, and the Director of Health Services and Outcomes Research for the Department of Surgery. She has an international reputation for her work in patient-reported outcomes and quality metrics. She has been the Principal Investigator of numerous R01's examining hospital quality and outcomes, and has been continuously funded by the NIH for the last 15 years. Dr. Doe has a successful track record for mentorship (primary mentor for 5 K-level awards and 7 F-level awards) as well as >15 trainees at the post-doctoral level over her career. For this proposal, Dr. Doe will provide guidance and training in grant writing, qualitative methods, and patient-reported outcomes. Dr. Doe has committed to weekly face-to-face mentorship meetings, in addition to contact by telephone and email as needed, to discuss the research protocol, review study progress and career development, plan future grant proposals, and a plan for dissemination of my research with peer-reviewed publications.

This section should complement the mentor statements of your mentoring team and/or advisory committee members. Note, anyone described in this section should also have a biosketch included in the application. Be cautious about proposing to utilize mentors that are not at your institution. This is typically not a good idea, especially as a primary mentor. If you do include a co-mentor or advisory committee member from another institution, you will need to justify to the reviewers why this is critical, and why the mentor would be willing to do this. A strong letter of support and evidence of prior collaboration would be necessary to ensure the environment is supportive.

This section should also elaborate on the specifics of your previously stated training domains that are critical to your success as an independent investigator. Within each of these training domains, state your intended method for how you will obtain this training. This could include didactic training in formal coursework, training institutes that may be on or off campus, and/or any additional training. You can provide an overview of the new methods and skills you propose to learn, and bullet or list each of the training domains to provide the details of the structured training (e.g., didactic coursework including course numbers/names, seminars, lab meetings, conferences, etc.). For example, you could describe how you will obtain specific experience on methodology of pragmatic trials by attending an NIH workshop on the topic and by attending bi-monthly lab meetings of one of your mentors who has expertise and ongoing research projects in this area). Details about the timing of these activities, including how much time and effort you will spend on each activity, should be explained. The most informative way to communicate all of this detail is in a table (e.g., Table 29.3 shows an example for the first 2 years).

Table 29.3 Example Table from Career Development Award: Career development timeline

	Year 1		Year 2	
	Q1-Q2	Q3-Q4	Q1-Q2	Q3-Q4
Formal coursework				
Behavioral research methods (goal 1)	EPI 540 MSCR 591			

Comparative effectiveness/medical decision making (goal 2)	CMS Medicare Data Workshop;
	MSCR 761;
	MSCR 509
		HPM 522		

Health information technology (goal 3)	HLTH 3100P	HLTH 3101P	NIH mHealth Institute	
Clinical trials (goal 4)		MSCR 520, MSCR 591		
Responsible conduct of research (goal 5)				

(continued)

Table 29.3 (continued)

	Year 1		Year 2	
	Q1-Q2	Q3-Q4	Q1-Q2	Q3-Q4
IRB rotation	X			
Key concepts in clinical research for investigators	X		X	
Research matters seminars (office of clinical research)	XX	XX	XX	XX
Professional development				
Local seminars	Surgery Grand Rounds (weekly), Lab Meeting (weekly), School of Medicine Health Services Research Group (monthly)			
National conferences	Academy Health, American Society of Transplantation, Surgical Outcomes Club (each venue annually)			
Research plan				
Develop predictive models (Aim 1)	**XX**			
Focus groups (aim 2)	**XX**			
Feasibility study (aim 3)		**XX**	**XX**	
Close-out				
Data analysis				**XX**
Publication			**XX**	**XX**
Grant writing		**Institutional Seed Grant Submission**		**NIH R03 submission**
Structured mentorship				
Weekly meetings with lead mentor	**XX**	**XX**	**XX**	**XX**
Monthly/bi-monthly meetings with co-mentors	**XX**	**XX**	**XX**	**XX**
Quarterly meetings with advisory committee	**XX**	**XX**	**XX**	**XX**
Clinical and teaching responsibilities (25% time)				
Weekly subspecialty clinic	**XX**	**XX**	**XX**	**XX**
Attending for weekly surgery resident clinic 4 h/week (career goal 1)	**XX**	**XX**	**XX**	**XX**

29.9 K Grant: Mentor, Co-mentor, Consultant, Collaborators Section

One of the most important aspects for mentored career development awards is finding the appropriate mentor. Ideally, this is a highly accomplished senior academic researcher with substantial experience (including a long history of NIH funding and prior experience mentoring junior faculty). Ideally, your mentor has a stable financial environment to ensure you will have the support you need. While a history of R01 funding is ideal, this is not necessarily required of a successful mentor. Most importantly, the mentor should be 100% committed to you, your proposed research, and your career development. This high level of support must be communicated clearly in the mentor statement.

The overall objective of the mentor statements section is for the mentor(s) to clearly describe how the candidate's research career will be enhanced by this career development award. The mentor should detail the plan for the candidate's training and career development, the source and support for research, the nature and extent of supervision, and the mentor's commitment to the candidate's development, teaching and research load. It is also important to emphasize other responsibilities of the candidate (to ensure they will have 75% protected time), and the plan for transitioning the candidate to career independence. The mentor statement should also describe the mentor's previous experience as a mentor (e.g., number of prior trainees, number of successful career grant recipients, etc.).

It is important that this section complement the candidate's background and career goals sections. After discussing the specifics with your mentoring team, it may help to provide a detailed outline with the specifics of what you agreed on in the mentoring plan to ensure consistency across the grant application. If you have additional mentors, they can include additional material as needed to the same document (max of 6 pages). Of note, letterhead is not necessarily required for the mentor statement, but is required for consultants, collaborators, or contributors.

29.10 K and R Grant—Research Plan

The research plan for a K grant can follow the same advice as for R-level grants, but the scope should be substantially smaller and allow for the time you need to obtain appropriate training to conduct the proposed research. This section consists of specific aims, significance, innovation, and approach. These are uploaded as two separate files: specific aims (1 page), and research strategy (12 pages).

29.10.1 K and R Grant—Specific Aims

Hopefully you have reviewed several examples of successfully funded grants and their specific aims, as this is the most helpful way to get started with these sections. The importance of the specific aims page cannot be overstated. It is the most important document of your entire grant application, so it needs to make a big impact. If you are fortunate enough to have your grant discussed at study section, this may be the only document (perhaps with your biosketch) that other reviewers will read while your proposal is being discussed. You should start drafting your aims page early and perfect it (after obtaining input from your mentoring team and the program officer) before continuing with the rest of the research plan. Remember that the research you propose in your specific aims should represent a new and expanded direction in your work. You should have a good amount of preliminary scientific support from other researchers as well as from your previous work.

Generally, the specific aims page should start with a brief paragraph establishing the problem. To address the population health impact, use absolute (not just relative) numbers (e.g., epidemiology of the disease or condition should be detailed using the incidence or prevalence of the disease). Be specific—use numbers—and cite your examples. The second paragraph should detail the knowledge gap in the problem you discussed in the first paragraph. Cite any sentinel studies as well as any of your own studies that may be relevant and can help establish feasibility and/or expertise in this area, and then state what is known and not known. This is where you will set up the third paragraph of how your study will address these important knowledge gaps, and answer a critical question that is important and novel for the field. You can preview your innovative approach here about how you will address these important knowledge gaps. It is ideal if you can also bring this back to the FOA or the institute you are applying for (you could even quote here how this is relevant to the specific funding announcement). Be sure to state somewhere in these paragraphs your long-term objective, as well as the objective of this grant. Next, you will list your specific aims (typically 2–3 aims), which should be as brief and specific as possible and will convey your outcomes and your hypotheses. Specify the aim in bold, and then detail the methodology that will be used to answer the aim (e.g., data sources, total number of patients/hospitals/clinicians you will study, the study design, time period, outcome(s)). The primary outcome must be crystal clear. Make sure your hypothesis is clearly stated. Some critical problems with aims: too many aims that make the work seem infeasible, subsequent aims are dependent on the prior aims, aims that are overly ambitious, and aims that have very little to do with one another to fit within the overall objective of the proposal.

The last paragraph (can be 1–2 sentences) should summarize how these results will have a real-world impact on population health, clinical medicine, health systems, etc. For a career development grant, many reviewers will also look for the next step of this research, e.g. “Upon completion of these K01 aims, we expect to have identified the hospitals with the worst post-surgical outcomes in our region and the patient and health system characteristics driving poor outcomes. These results will provide critical pilot data to inform our future R01 grant submission to implement existing evidence-based interventions to improve surgical outcomes.” This is a lot of content for one page, so ensuring it is clear and succinct should be a top priority!

29.10.2 K and R Grant—Writing the Significance Section

The purpose of the significance section is to explain the importance of the problem your proposal addresses and how, if successfully achieved, your proposed aims will address the critical gaps in the field. This section details how your proposed research will improve scientific knowledge, technical capability, and/or clinical practice in the field. This section is NOT a literature review. The significance section should answer the “why” and the “so what?”—what is the compelling rationale for this research and how will your proposed work advance the science?

The significance section is typically no more than two pages (4–6 detailed paragraphs). It is extremely helpful to reviewers if you start each paragraph with a bolded summary statement or heading for the topic. The first statement is typically an overview of the epidemiology of the disease/condition/problem to provide a summary of the population health impact and scope of the problem. Subsequent sections should set up the gaps in the literature and the rationale of your proposed research. For example, in a proposal that plans to pilot test a shared decision aid for surgical patients, one paragraph in the significance section could include background about the motivation for why a decision aid may be important in this population, e.g. "Critically important treatment decisions are often made without evidence-based information about a patient's prognosis." This paragraph would then be followed by a summary of what is known in the field about how patients currently make treatment decisions and why this may be a problem. The final sentence of this paragraph should tie the gaps you have acknowledged back to your proposal, e.g. "Through surveys and focus groups, this proposal seeks to identify the most important information that patients and providers would like when considering transplant surgery and then develop and pilot the use of a simple, decision support tool to help patients make treatment decisions." The significance section may also include important preliminary data (this can also go in the approach section). If you have any relevant publications of your own, incorporate these here (e.g., "We have previously shown…(citation)"). For example:

- **A.1. The Challenge of Medication Adherence in Organ Transplant.** Solid organ transplantation is a chronic illness, where transplant patients must adhere to life-long treatment. In 2011, 17,000+ individuals received a kidney transplant and more than 4000 received a liver transplant; either from a deceased or living donor [1]. Approximately one third of these organs fail within 5 years—primarily due to host rejection of the organ (graft) [2]. To prevent graft failure and life-threatening, costly complications, transplant recipients must closely adhere to immunosuppressive medication regimens that require constant monitoring of therapeutic levels and frequent dose adjustment by care providers [3, 4]. Patients are often informed of medication changes by phone without written instructions, increasing the risk of confusion as existing regimen information and prescription labels become outdated [5, 6]. Patients also take other regimens for comorbid conditions, such as diabetes and hypertension [7–9]. According to recent pilot study we conducted among 78 transplant recipients from three metropolitan transplant centers, patients take on average 10 medications [10]. Transplant recipients' medication regimens are exceedingly complex and dynamic, involving frequent physician visits, laboratory tests, and lifestyle and dietary modifications. Ongoing regimen adherence—including adjustment to frequent changes—is crucial, yet very challenging. *In this proposal, we will employ a user-centered design to develop an electronic patient portal for transplant recipients and their providers that will post real-time updates to medication changes, organize patients' entire drug regimen in the most efficient manner, provide a way to communicate concerns via instant messaging, and program daily SMS text reminders to help patients' remember when to take medicines.*

In contrast to the advice on completing the specific aims before writing anything else, it may be okay to start with bullets for main points of the significance section, and then add in more detail after the approach section is written to ensure the content is directly applicable to the methods and the approach you take to answer your proposed aims. Other important tips: bold and italicize main points (busy reviewers will appreciate this as they may just skim this section!) and include white space so the proposal is easy to read; remember that your reviewers are likely not all experts in your field, so do not assume they have any basic knowledge about the problem; avoid acronyms and common abbreviations in your field (as they may not be common in other fields); incorporate figures and tables of the main points you want to convey to the reviewers (and make sure that the tables/figures are readable and have appropriate citations so it is clear where any data come from).

29.10.3 K and R Grant—Writing the Innovation Section

The NIH definition of “innovation” is important to consider when writing this section: “Does the application challenge and seek to shift current research or clinical practice paradigms by utilizing novel theoretical concepts, approaches or methodologies, instrumentation, or interventions? Are the concepts, approaches or methodologies, instrumentation, or interventions novel to one field of research or novel in a broad sense? Is a refinement, improvement, or new application of theoretical concepts, approaches or methodologies, instrumentation, or interventions proposed?”

This section can be short (a paragraph) and is best digested by reviewers in bullet points. This is where you could highlight how your proposal is doing something that has not been done before. For example, you may be the first to test a particular research question (ever or in this particular patient population), study a particularly unique and understudied patient population, or use innovative methods, techniques, or study design to answer the important research questions. Do not go overboard and stretch the innovation if the point is not truly innovative. Three to four bullets is the norm here, and in career-development grants the bar is not set quite as high as R-level grants in terms of novelty. However, you should still be proposing something novel that has not yet been researched.

29.10.4 K and R Grant—Writing the Approach Section

The approach section is the section of the grant in which you describe your overall research strategy, planned methodology, and proposed analyses to answer the specific aims of your research. There are no required named subsections as dictated by NIH instructions, but key sub-sections within the approach include the study overview, timeline, study team, conceptual model, preliminary data (if not covered earlier), study design, population (e.g., patients or subjects, hospitals/health systems, etc. and any inclusion/exclusion criteria), data collection procedures, informed consent procedures, study measures (including outcome measures), statistical analyses, sample size/power, data management (including missing data), expected outcomes, dissemination of results, potential limitations and alternative approaches. This section is typically the longest section (8–10 pages) and contains all of the details on how you will answer your aims. Reviewers will score this section.

It helps reviewers to start with a section on study overview to remind reviewers of the overall goals and objectives and include the what, where, when, and how of the research proposed. This helps to provide context prior to the deep dive of details in the next sub-sections within the approach. The initial overview should include the actual geographic location and study sites (and potential justification for focusing on these particular regions) in the context of the overall study objective and long-term research goals.

The next paragraph may include details on the study team (including information about the PI, all co-investigators, study sites, etc.)) and whether the team of investigators has previously worked together. If so, describe (e.g., our team has previously collaborated on two prior R01 grants and several papers on this subject (cite)). For example:

- **C.2. Project Leadership and Management**. Our collaborative team for this proposed project will be led by epidemiologist and health services researcher Avedis Donabedian, MD, MPH (PI from Harvard University). Dr. Donabedian has more than 30 years' experience studying health systems management and the quality of medical care and is the PI of two large studies in this area (R01-MD123456 and R01-DK123456). Included on the scientific team will be members from our surgical collaborative, including surgeon Joseph Lister (Co-I, Professor, Harvard), epidemiologist John Snow, PhD (Co-I, Associate Professor, Harvard), behavioral scientist and qualitative researcher Dr. Everette M. Rogers, PhD (Co-I; Assistant Professor, Harvard); and health economist Dr. Adam Smith, PhD (Co-I; Associate Professor, Harvard). An added strength of our proposed team is the contribution of the Massachusetts quality improvement organization (see **Letter of Support**). Our team includes experienced clinicians, epidemiologists, health services researchers, and social and applied behavioral scientists, epitomizing team-based science. All Co-I's, community partners, and patients have previously collaborated through the 10-year history of our surgical collaborative.

In R01 grants, the investigator(s) is a scorable section, and reviewers will often look to this paragraph in the approach section in addition to the submitted biosketches to get a summary of the key personnel on the grant and to evaluate whether there are any major weaknesses (e.g., missing expertise). It should be stated how often the team will meet together, and if coinvestigators or study sites are distributed geographically, how project oversight and site-standardization will occur to ensure the proposed aims are carried out successfully. This section can also help to address feasibility of the proposed research if it is clear that the team has worked together previously.

The following paragraphs and sub-sections will depend on the type of study you do. Using examples of previously funded grants can be valuable here. In addition, because the type of information you need to address may vary based on the study design for your health services research project, it is helpful to use a checklist. If you are following a particular study design (e.g., cohort study, randomized study, pragmatic trial, survey, etc.), use study design guidelines and their corresponding checklists (http://www.equator-network.org/) to ensure you have covered all of the essential elements a reviewer may raise (e.g., STROBE guidelines for observational research, CONSORT guidelines for randomized trials, SURGE guidelines for surveys, etc). In addition, NIH has a new emphasis on "rigor and reproducibility" relevant here, so include a statement somewhere in the Approach that demonstrates how your methods will be rigorous and reproducible [11]. These guidelines should be consulted with while writing the approach section to ensure each item is addressed.

For health services research grants, there are several areas that tend to most often get discussed and critiqued. For nearly all proposals, there should be a conceptual model or theoretical framework guiding the research aims and driving your hypotheses. If you are not an expert in this, find someone who can help you (e.g., a behavioral scientist could help if you are proposing a behavioral intervention). The best way to convey this model is typically using a figure or diagram to illustrate the key concepts and how they fit in with your specific aims. For example, the social ecological model is a commonly utilized framework that considers how the interaction of individual, interpersonal, community, organizational and societal factors may have an effect on patient outcomes, and may be a relevant model to use when planning and implementing health promotion interventions. It helps to present your conceptual model prior to describing your specific aims and approach (for example, when selecting what covariates to measure and/or control for in analyses, you can refer back to this model). If you plan to collect a number of study variables/measures, it may help reviewers if you have a summary table highlighting the proposed measures, how they will be collected (e.g., survey, phone call, electronic medical record, etc.) and the primary outcomes for aims. Other potential issues in health services research-related grants are similar to what you may be critiqued for in a health services research manuscript—remember to address any potential biases that may occur (e.g., selection bias due to loss to follow-up or missing data, misclassification bias, unmeasured confounding, etc.) and how you propose to address these biases. Study design guidelines can help with this, and for those seeking additional background on these topics, Rothmans' *Epidemiology: an Introduction* provides a good overview of the many types of biases in clinical research [12].

In 2015, NIH published new policies for addressing sex as a biologic variable. This should be addressed in the approach section. As stated in the NIH notice, "NIH expects that sex as a biological variable will be factored into research designs, analyses, and reporting in vertebrate animal and human studies." [13] This can be addressed in the subsection on study population that can detail the expected demographic and/or clinical characteristics of the proposed study population by describing the expected sex distribution or any pre-planned stratified analyses by sex.

29.11 K and R Grant—Other Sections of Grant

It is easy to spend months on the specific aims and research plan alone, but do not forget about all of the other important documents that are required for the grant. Some are scorable, and all are important. A missing document (e.g., human subjects section) could prevent an otherwise excellent grant from being funded. While this chapter cannot go into the details for all of these sections, the most relevant to consider include the abstract, narrative, facilities and resources, budget, statistical analysis plan (depending on whether you are proposing a clinical trial or not), planned enrollment tables, inclusion of children and minorities, resource sharing plan, data safety and monitoring plan or board, and letters of support.

Letters of support are important for reviewers, and can be used to help score the Environment section for R-level grants and the Environment & Institutional Commitment to Candidate section for K-grants (see Table 29.1). There are specific requirements for letters of support for career development awards (e.g., K01 requires three letters and a maximum of five letters, these cannot be from your primary mentor or co-mentor). Letters of support are important to determine feasibility of the study as well (e.g., if you are proposing to use a data source that is not publicly available, a letter of support detailing that you have permission to use these data is helpful). If you propose to conduct your study at study sites outside of your institution (e.g., community hospitals) that are not subcontracts on your proposal, it is also important to demonstrate feasibility of this with a letter of support. Other letters could convey general support, such as letters from relevant clinical societies or patient advocacy organizations that help to communicate the potential impact of the work you propose. While many grant writers tend to recycle facilities and resources documents from prior grants without much thought, for health services research grants (that may be more complex to describe in the approach section—e.g., the characteristics of the 10 hospitals you plan to include as study sites), this is an opportunity to provide more detail (e.g., hospital location, leadership, volume of patients, etc.) in addition to the typical material you find in this document. Typically, your research administration office can help you with the various other documents needed for the proposal.

29.12 General Tips for Writing Success

If writing grants was easy, everyone would be successful in getting funding. It takes time and hard work. Plan to start writing as early as possible. If it is your first grant submission, an ideal timeframe is 4–6 months. Start early, and write often. Create a "to-do" list that is more than just "write K grant" but breaks this down into subsections so each section is easily tackled. If you have not yet had the experience, review someone else's grant either informally or through an institutional study section. Reviewing other people's grants will allow you to put yourself in the mind of the reviewer and can improve your grant writing skills. Always remember that your reviewers are people that are very busy, and are responsible for reading many other grants in addition to yours (often 10 or more at a time) amidst their own research and/or academic responsibilities. Writing clearly and concisely and formatting your grant so it is visually appealing is important. Remember to avoid abbreviations and acronyms if at all possible, and do not assume that your reviewer is in your field. Proofread several times, as typos will imply to reviewers that your research is also sloppy. Prior to submitting your final grant product, utilize internal reviewers at your institution, if possible. Find a trusted colleague and/or mentor who will give you critical feedback early. Then, submit and submit often!

References

1. Whicher D, Rosengren K, Siddiqi S, Simpson L, editors. The future of health services research: advancing health systems research and practice in the United States. Washington, DC: National Academy of Medicine; 2018.
2. Bindman A. Moving AHRQ Into The NIH: New Beginning Or Beginning Of The End? https://www.healthaffairs.org/do/10.1377/hblog20170327.059384/full/. Accessed 22 Jan 2019.
3. U.S. Department of Health and Human Services NIH. Mission and Vision. https://nimhd.nih.gov/about/overview/mission-vision.html. Accessed 22 Jan 2019.
4. U.S. Department of Health and Human Services NIH. https://public.csr.nih.gov/. Accessed 22 Jan 2019.
5. U.S. Department of Health and Human Services NIH. Research Portfolio Online Reporting Tools. Version 7.37.0:https://projectreporter.nih.gov/reporter.cfm. Accessed 22 Jan 2019.
6. U.S. Department of Health and Human Services NIH. Research Career Development Awards. https://researchtraining.nih.gov/programs/career-development. Accessed 24 Jan 2019.
7. U.S. Department of Health and Human Services NIH. Policy Supporting the Next Generation Researchers Initiative. 2017. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-101.html. Accessed 24 Jan 2019.
8. U.S. Department of Health and Human Services NIH. Update on NIH's Next Generation Researchers Initiative (NGRI) Policy on Early Established Investigators (EEIs). 2018. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-18-214.html. Accessed 24 Jan 2019.
9. U.S. Department of Health and Human Services NIH. Early Stage Investigator Policies. https://grants.nih.gov/policy/early-investigators/index.htm. Accessed 24 Jan 2019.
10. U.S. Department of Health and Human Services NIH, Office of Extramural Research. Biosketch Format Pages, Instructions and Samples. https://grants.nih.gov/grants/forms/biosketch.htm. Accessed 24 Jan 2019.
11. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612–3.
12. Rothman KJ. Epidemiology: an introduction. Oxford: Oxford University Press; 2012.
13. Consideration of Sex as a Biological Variable in NIH-funded Research. In: Health NIo, ed. *NOT-OD-15-102*2015.

Choosing Your First Job as a Surgeon and Health Services Researcher

30

Scott E. Regenbogen and Alexander T. Hawkins

Abstract

An academic surgeon's first faculty job is a key determinant of future success. However, the process for finding that job is poorly defined, especially as compared with the standardization of residency and fellowship training. Securing a faculty position involves four distinct phases: identifying what type of job you are looking for, identifying potential positions, screening jobs and successfully negotiating a package that will offer you the best chance of success. At each stage, careful attention to the candidate's goals, skills, and career mission are essential. In this chapter, we will discuss some guiding principles to help find a job with greatest chance of academic success and personal and professional fulfillment.

30.1 Introduction

When aspiring academic surgeons approach the end of training, the process of finding a faculty job is often daunting. After 5–10 years of postgraduate surgical training, it has been a long time these individuals have undertaken a job search that involved anything other than a pro forma application and a computer matching algorithm. Tenure-track academic surgery positions are rarely found in the back of journals or in online job databases, so the search is driven more by networking and the influences of mentors and local department leaders than by the forces of the

S. E. Regenbogen (✉)
Department of Surgery, University of Michigan, Ann Arbor, MI, USA
e-mail: sregenbo@med.umich.edu

A. T. Hawkins
Department of Surgery, Vanderbilt University Medical Center, Nashville, TN, USA
e-mail: alex.hawkins@vumc.org

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_30

traditional job market [1]. The choices made in this process are extremely important, however, as the first faculty position is often cited by successful academics as an absolutely critical factor in the development of a productive career in academic medicine [2]. A measured, careful and structured approach to the job search process is, therefore, highly advised, yet often bypassed, in favor of familiar settings and known entities. The process of obtaining a first faculty position in academic surgery comes in three main stages, which we will consider separately: looking for jobs, screening opportunities, and crafting a position. Before embarking on any of this, though, a period of self-evaluation is required.

30.2 Determining What Job You Are Looking For

Before the job search begins, it is critical to take the time for introspection about what you want in your early career. Success at this stage will require an honest and detailed cataloguing of needs, interests, motivations and goals. It is important to identify the elements of a job that are keys to development and success.

With a mission in hand, the first major framing decision will involve the emphasis and weighting of activities. How much of the professional effort will be patient care, and how much will be scholarly? Assuming that health services research is a meaningful part of the mission, there are generally two models for the academically influential surgical health services researcher. Some will be surgeons who publish—a clinically busy surgeon with a reputation as a skilled practitioner, a regional, national, or international referral base, access to clinical outcomes data, and collaboration with researchers who provide analytic expertise and continuity to the flow of academic work. These surgeons will generally spend 60–80% of their time on clinical work, external grant funding will not be a major contributor to their salary, and their incentive compensation will be determined primarily by clinical revenue. Others will be the health services researchers' version of the surgeon-scientist—a researcher who operates. These surgeon-scientists will generally be known for their publications, presentations and external grant support. Their clinical practice will be designed to consume no more than 20–40% of their time. They will support their salaries as principal investigators with external grant funding to offset the opportunity costs of lost clinical revenue, and their incentive compensation will ideally be tied to scholarly productivity and revenue from grant funding [3].

For young surgeons nearing completion of clinical training, it may be hard to distinguish which of these roles is a better fit. After all, this degree of self-determination has been largely absent from the trainee's recent years' experience. But there are some clues you can use for self-assessment. A look at the non-negotiable elements of the mission statement will often come from the side of the equation that should be receiving more emphasis. Also, "stretch goals" will often define the area deserving of most attention. Some may imagine achieving recognition through clinical volume and reputation, others through administrative promotion and leadership, and others from a major grant or paper published in a prestigious journal. Understanding which of these achievements will be the primary motivator, and which job model is a better career fit will shape the kinds of faculty positions sought, and the ways that potential employers will evaluate your candidacy. Keep in mind that nothing is set in stone and that successful careers often take a circuitous path.

The last step in preparation is an honest self-examination, to evaluate skills, expertise, and gaps in your track record. Do you have the knowledge, experience, and tools to carry out a career plan to reach your goals? Scrutinize your publication record, previous mentor relationships, and curriculum vitae, and try to imagine how your experience will be rated by interviewers, potential mentors, and department chairs. How many first-author publications do you have to demonstrate your ability to conduct and complete a paper? Have you ever obtained research funding, such as a training grant, or local project support? Do you have the classroom didactic training to support your research plan as an independent scholar? The health services research toolkit typically involves knowledge of biostatistics, clinical epidemiology, health policy and management. If any of these are lacking, but required for future work, consider how you might obtain formal training either before or, if necessary, during your appointment [4]. Some departments may be able to support enrollment in relevant coursework, but this will need to be considered early on.

This introspective exercise should culminate in construction of a personal "mission statement" that summarizes career goals across the four components of a job in academic surgery—scholarly, administrative, didactic, and clinical. In each of these categories, you will want to identify 1, 3 and 5 year concrete goals (Table 30.1). Some goals may be subject to external forces not entirely within your locus of control (e.g. "obtain" a grant, rather than just "submit") and it is understood that success is not necessarily dependent on achievement of each of these goals on the prescribed calendar. Rather, this document will communicate your priorities to potential mentors, employers and collaborators and establish benchmark targets. Even more importantly, if properly constructed, it will be the touchstone by which you will adjudicate job choices and other critical career decisions along the way [5]. This is a task that deserves dedicated time and careful attention before the job search even begins.

Table 30.1 Strategic goal planning (with examples)

	1-year goal	3-year goal	5-year goal
Scholarly	Publish a manuscript as first year author	Obtain an internal research grant	Obtain an external research grant
Administrative	Establish a network of connections within your department	Assume a position on a hospital committee	Earn a leadership position
Mentoring	Lecture to the residents on your core clinical focus	Publish a paper with a resident in your research area	Mentor a resident in your research area
Clinical	Ramp up clinical practice to mutual targets	Learn a new clinical skill	Build out a new clinical program

30.3 Identifying Potential Positions

Once you have carefully compared your future goals and current skills, it is time to begin looking for faculty job opportunities. Depending on your clinical specialty, these jobs may be posted on institutions' human resources websites, or advertised in the classified sections of journals or at meetings. However, it is most likely that the right job will not be found this way. Most jobs in academic surgery come from networking and referrals from colleagues or mentors [6]. So, a first step is simply to talk with those around you—residency and fellowship directors, division and department chairs, and other acquaintances in your clinical and/or research specialty. In health services research, these people can be found at the Quality, Outcomes, and Cost sessions of the American College of Surgeons Clinical Congress Surgical Forum, at the Outcomes Sessions of the Academic Surgical Congress, and through the Surgical Outcomes Club meetings and website. Seek out people whose papers you read and whose talks you admire. In particular, think about young faculty surgeons you might emulate—their mentors are likely to be people who can provide wise and essential guidance toward job opportunities. At each national meeting you go to, make a "meet list" of people you want to connect with. Contact them 2 weeks out and ask for 15 min of their time. Their contact info can usually be found on line or through societies' websites. You must cast a wide net. The more you talk to people, and the more visible you make yourself, the wider the range of opportunities you will have to evaluate and consider.

An early decision to be made about the setting for the job is what type of research environment to seek out. Though surgical health services research has grown substantially in the past decade, from a sparsely populated cottage industry to a solid establishment distributed widely around the country, there are still a limited number of surgical departments that can boast a well-apportioned infrastructure in the field. One option is to seek out one of these institutions—a number of them can be found on the Surgical Outcomes Club website (http://www.surgicaloutcomesclub.com/links). These are departments in which an established senior investigator has ongoing projects and opportunities for junior faculty, residents and fellows. There may be up-and-coming protégés who have obtained career development awards with this senior mentor, or even progressed to their own major project grants under this tutelage. These institutions may have an established data management infrastructure, experienced analysts, and a core of other investigators with whom a junior faculty member might collaborate.

Beyond these surgical health services research hubs, however, there is a far wider variety of academic centers with a mature infrastructure of non-surgeons, and even non-physicians, doing this kind of work. Many of these groups would be well-served by the addition of a young surgical investigator among their ranks. Often, their work will extend to areas of clinical expertise for which a surgeon can offer important practical insights. Senior investigators in these settings can often provide very good mentorship and development opportunities for junior faculty. They may be somewhat unfamiliar, however, with the clinical demands and expectations placed on junior surgical faculty. The academic medical doctor will often do just one half-day outpatient clinic per week, or attend an inpatient service several weeks per year. They have large blocks of purely non-clinical time, whereas academic surgeons will typically have weekly clinic, operations, and a steady stream of patients in the hospital, requiring academic and clinical work to proceed in parallel. So, if you plan to be the surgeon in a medical health services research group, you will need to establish an understanding with your mentor and with your clinical chief to establish a productive mentoring and organizational structure and career development plan that works in this setting.

The third option for research setting is to "go it alone" as the pioneer surgical health services researcher in a department seeking to expand its research domain. Many surgery departments' traditional focus on basic science research has broadened, as the advent of outcomes benchmarking, public reporting, pay-for-performance, and reimbursement reform have increased recognition of health services research as an essential component of surgical care improvement. These institutions may be interested in recruiting young faculty with training and experience in epidemiology, biostatistics, program evaluation, and health economics to found a home-grown surgical health services research group and provide training opportunities for residents and scholarly support to other faculty. Or they may be interested in applied health services research techniques, looking for a surgeon to lead internal quality improvement and clinical innovation efforts. These settings provide a great opportunity for early independence. Essential to this approach is a highly-motivated department chair committing meaningful resources and influence to the effort. In this scenario, you need to be very clear about the staff and finances you will be provided as this will be a key indicator of your success. Also make sure that there exists some level of higher mentorship at the institution. This may be found outside of the surgery department, depending on your particular interests and goals. There are real challenges for the junior faculty surgeon trying to establish both a clinical practice and a research infrastructure at the same time. And there is a very real risk of intellectual isolation without other like-minded faculty around. A young investigator going it alone will be well-served, therefore, to maintain close ties and collaborations with former mentors or other allied colleagues in other institutions.

30.4 Screening Jobs

Understanding the advantages and limitations associated with each of these settings, the next step will be to field recruiting offers and reach out to institutions that may be of interest. As these communications proceed, you will need to decide which ones are potentially viable opportunities worth investigating with a visit and interview. This is a good time to revisit that mission statement, refresh and update your priorities and goals. They will be used as a checklist by which to evaluate each institution and the jobs they propose (Table 30.2).

A useful first screen will be to evaluate the success of academic junior faculty they have hired in the past 5 years or so. The experience of recent hires is a very

Table 30.2 Checklist for assessing potential positions

Research	Do your goals align with those of the chair? What kind of support will you receive- monetary, biostatistics, research assistants, project managers?
Mentors	Who will be your mentors? There should ideally be a panel comprised of both senior mentors who can help with resources as well as mid-level faculty who can help more with the execution of your work. Methodology is usually more important than subject content.
Geography	Are you (and your family) going to be happy in the location? Are you close to extended family (if you want to be)?
Partners	Do you get along well with your potential partners? Do they seem happy? Are they supportive of one another? Has there been a lot of recent turnover?
Compensation	Is the salary competitive? Are you going to be able to meet your expected expenditures?
Protected time	Will you be provided with explicit protected time to pursue your goals? Make sure you understand your block time as this is an essential foundation of success.
Clinical work	Are you going to be expected to provide clinical services that fit with your training and desires?
Call	How often will you be on call? What are your explicit responsibilities while on call?

good predictor for how things might be if you join the department. Read their faculty profiles on the department's website, search PubMed (http://www.ncbi.nlm.nih.gov/pubmed) for their publication records and the NIH RePORTER (http://projectreporter.nih.gov/reporter.cfm) for their federal funding results. How many of them have obtained career development awards or society grant funding to support their startup? Are they publishing successfully? When an interview visit is planned, be sure that some of these young faculty members are included in your itinerary as they will be a very important inside information source about the viability of the clinical/academic balance in the institution. If they have been well supported and positioned to succeed, there is a good chance that you can follow their lead.

During the first interview visit, there is a great amount of detailed information to be obtained. Think of the first interview as a first chance to get to know the personality of the institution, the department, and the faculty. Is this an environment in which you can see yourself being successful? It is essential to get the "30,000 foot view" of what is valued and rewarded in the department. What is the leadership's vision for the department and its faculty? Do they want to be the dominant clinical referral center? How do they view their role in the surgical community? What do they want to be known for? And what is the salary and incentive structure for faculty? Most departments will award financial bonuses for clinical productivity, but there are also some that offer equivalent or alternate rewards for academic success [5, 7]. These allow faculty to forgo additional clinical referral in favor of scholarly work, without sacrificing their income. Even more importantly, however, the presence of these incentives demonstrates a commitment to academic activity and anchors the goals of the institution by putting money on the table. At the most basic level, try to figure out whether the priorities and vision of the department align with your personal ambitions, and whether your most important goals will be supported? [8].

If the department's mission and goals seem a reasonable match, try to define and imagine yourself in one or two very specific jobs within the department, both the clinical part and the research part. On the clinical side, assess the demand—is there a need for someone with your training? If you can bring a needed set of skills to an environment in which referrals are directed to the health system, rather than to particular surgeons, it will take far less effort to get a clinical practice started. Young surgeons who need to generate personal referrals will be more likely to sacrifice research time in order to respond to unplanned referrals and generate good will from their referring doctors. At the other extreme, closely evaluate your potential partners. Be sure that there is some opportunity to share clinical responsibilities with others, so that the burdens of urgent referrals will not too often interfere with planned academic time. Do you get along well with your potential partners? Has there been frequent turnover? It is essential to talk not only with faculty that are at the institution, but also those who have left. Overall, there should be general agreement between how busy you want to be, and how busy you will need to be in the particular setting being considered.

Probably the most important element to investigate, however, is the availability of adequate mentorship [6, 9]. Mentors will come in many forms. You should make contact both with potential senior and mid-level mentors. An effective senior mentor will guide career development training, ensure effective navigation of local politics, shape and optimize the quality and viability of manuscripts and grant applications, and guide involvement in key committees and specialty organizations [10]. An independently successful mentor can also provide research resources, as a mentee can often make secondary use of the data, equipment, and collaborators already present in the mentor's organization. A senior mentor with a well-established and productive research infrastructure may provide even more resources in-kind than could be obtained with departmental start-up funds. And the academic mentor can be an essential line of defense against competing demands from clinical superiors. Mid-level mentors are also essential to helping with "boots on the ground" execution of research and navigation of institutional culture. These are the faculty members that you aspire to emulate in 5–10 years. Finally, for women and minorities, identifying mentors that you will be able to work with is even more critical. Women and minorities may face an additional set of challenges when starting an academic career [11]. Identifying mentors who can help you address these challenges is paramount.

Finally, think about the physical space. Is the hospital physically separate from the university and academic center? As a surgeon, rounding on postoperative patients, attending departmental meetings and other clinical responsibilities may leave relatively few days with no clinical commitments at all, so the time spent in transit can add up, unless travel between the clinical and research sites is relatively easy. Some distance, however, may be helpful, as physical departure from the clinical environment allows more complete separation of time commitments.

30.5 Negotiating the Job

If after the first interview, you are considered a good candidate for the position, a second visit will typically follow. The true negotiation begins with an offer letter from the department. This may come before or after your second interview. The offer letter should detail salary and other compensation, clinical responsibilities, teaching roles, administrative support, academic expectations, research start-up funds and other resources. Some institutions will even include explicit academic and clinical mentoring relationships in the letter as well. Regardless of the timing of the offer letter, the second interview offers the chance to ask more in-depth questions about what your job will look like, including the above themes.

On the clinical side, you should aim to define what proportion of your effort is supposed to be spent on patient care, and how that effort might be adjusted if research funding is obtained. If a career development award, such as federal "K" grant, is part of your academic plan, be aware that they typically require 75% effort dedicated to research, and ask whether this sort of funding mechanism could be accommodated. Some will make protected time for academic work explicit, whereas others may define effort allocations more conceptually. Simple measures of clinical expectation might be how many days per week you will operate, and how often you will be on call. Some departments will have minimum productivity expectations, measured in dollars or work RVUs. Others may treat the individual surgeon as a cost center and expect faculty to generate revenue to cover their costs and salary. At the most basic level, will your regular paycheck be determined by your clinical productivity? Regardless of the accounting method, it will be important to understand how clinical volume will be measured and at what point a new faculty member will be held accountable. Some will offer 2 or 3 years of allowance to grow clinical practice volume, but this is not universal.

The letter should also define what resources will be available to support clinical work. Will operating room block time be allocated to you directly? Will clinic space, medical assistants, and clinic support staff be available? Are there mid-level providers or clinical care support staff to answer patients' phone calls, obtain records, and help with the logistics of clinical care?

On the research side, the offer letter should state general expectations and metrics for success in scholarly work. Physical space for conducting research should be explicit. If you are joining a well-established group, make sure that you have an office, cubicle, or at least a dedicated seat alongside your collaborators and mentors. If the group already has data managers, programmers, analysts, and/or statisticians, your start-up financial contribution to the group could be to have the department pay for a time-share of one or more of these people—perhaps half a data analyst for 2 or 3 years. Remember that this should include support for both salary and benefits for this individual. Otherwise, consider requesting salary and benefit support to hire a research assistant who can perform some of these tasks. Some start-up discretionary cash should be included as well, to allow the purchase of computer equipment, software, databases—the typical health services research needs. Although the cash support ("hard money") needed for health services research may be less than is needed to run a basic science lab, it will still be needed, especially if not joining a group with established data sources and analytic support.

Some department are reticent to put fine levels of details in your contract. A white paper that summarizes many of the above issues, as well as including your own goals from Table 30.1 can be very useful in making sure that both you and your chair are on the same page. This document will detail your own goals and the resources (time, staff, training, etc.) that the department will commit to help you achieving them.

Surgical training provides us with an enormous skill set, but negotiating a job offer is not one of them. When you enter this phase of securing a faculty position, it is important to leverage additional help. A wonderful and easily accessible book for this task is the book *Getting to Yes* by Roger Fisher, Bruce Patton and William Ury [12]. It introduces the concept that instead of a confrontation, the exercise of negotiation should be an attempt to always reach a "Win-Win" agreement. Your chair will have finite resources to dedicate and it is essential that you develop clear mutual goals. There will be gives and takes throughout the process. That why it is so important from the outset that you think about elements of a job that you must have and those that you cannot live without. Involve department heads and chairs of the centers that you trained at during this process to ensure that you are getting a fair offer. Consider having your contract read by an attorney with experience in employment contracting.

In the end however, the offer letter and subsequent negotiations are just a starting point. This conversation will define the essential details of the job, but the realities of the work will continue to evolve even after the final signed version of the contract [8]. Upon arrival in the job, start immediately setting consistent patterns for your involvement in academic work. Put boundaries on clinical time. Establish your limits for unplanned, urgent consultations. Make sure time is blocked on your calendar for research meetings, reading, thinking, and writing, and treat these time blocks like mandatory meetings. And share your research progress with your clinical team, to help them understand the value and importance of both sides of your professional life. The precedents set in the first few months of the job will be hard to alter later, and will be more rigid determinants of the realities of the job than the paper contract signed in advance.

30.6 Summary

The task of finding a first faculty job in academic surgery can be daunting, but a few guiding principles can keep the process on track. First, go back to your mission statement often. Remind yourself why you chose to pursue this path in life, and why it is important to you and others. There is one job that would be right for everyone, so the requirement is to find the particular job that is right for you. Second, keep a long view. A career in surgery is likely to last three decades, so think about long research arcs and their impact. Choose a line of inquiry that motivate you to get out of bed in the morning, and feels important enough to counterbalance the very compelling demands of patient care. And finally, take a great opportunity over great money. Mentorship, collaborators, protected time, institutional support, and successful work/life balance will be far greater determinants of success and satisfaction than salary level, start-up funds and other compensation.

The process is poorly defined and the stakes are high—the first job will likely set the course for the rest of your academic career. Often the process is undertaken during a time of high mental and physical workload in clinical training, with less time for careful consideration and negotiation than is deserved. The needs and interests of the academic surgeon—challenging clinical work, competition for research funding, and the drive to scholarly discovery—are complex, and often competing. But the opportunity and the privilege to make important contributions to patient care and public health as a surgical health services researcher are very real.

References

1. Skitzki J, Reynolds HL, Delaney CP. Academic university practice: program selection and the interview process. Clin Colon Rectal Surg. 2006;19(03):139–42.
2. Nelson PR. Timeline for promotion/overview of an academic career. In: Chen H, Kao LS, editors. Success in academic surgery. London: Springer; 2011. p. 11–30.
3. Staveley-O'Carroll K, Pan M, Meier A, Han D, McFadden D, Souba W. Developing the young academic surgeon. J Surg Res. 2005;128(2):238–42.
4. Kuy S, Greenberg CC, Gusani NJ, Dimick JB, Kao LS, Brasel KJ. Health services research resources for surgeons. J Surg Res. 2011;171(1):e69–73.
5. Souba WW, Gamelli RL, Lorber MI, Thompson JS, Kron IL, Tompkins RG, et al. Strategies for success in academic surgery. Surgery. 1995;117(1):90–5.
6. Ghobrial IM, Laubach JP, Soiffer RJ. Finding the right academic job. Hematology/the Education Program of the American Society of Hematology American Society of Hematology Education Program 2009:729–733.
7. Poritz LS. Research in academic colon and rectal surgery: keys to success. Clin Colon Rectal Surg. 2006;19(3):148–55.
8. Schulick RD. Young academic surgeons participating in laboratory and translational research. Arch Surg. 2007;142(4):319–20.
9. Lillemoe KD. Surgical mentorship: a great tradition, but can we do better for the next generation? Ann Surg. 2017;266(3):401–10.
10. Sosa JA. Choosing, an being, a good mentor. In: Chen H, Kao LS, editors. Success in academic surgery. London: Springer; 2011. p. 169–80.
11. Greenberg CC. Association for Academic Surgery presidential address: sticky floors and glass ceilings. J Surg Res. 2017;219:ix–xviii.
12. Fisher R, Ury WL, Patton B. Getting to yes: negotiating agreement without giving in. Rev. ed. Boston, MA: Houghton-Mifflin; 2011.

Building a Health Services Research Program

31

Benjamin S. Brooke and Samuel R. G. Finlayson

Abstract

The ability for surgeons to carry out high-quality health services and outcomes research requires organized efforts and infrastructure at both the departmental and institutional level. Beyond having well-trained investigators, programs dedicated to surgical health services research must possess well-managed resources along with a clear mission and vision. This chapter details the important elements that are needed to build a successful surgical health services research program.

31.1 Introduction

A growing number of surgery departments in academic medical centers are striving to develop capacity in surgical outcomes and health services research (HSR). While some surgeons still harbor the misconception that surgical HSR is something that can be done on nights and weekends with data analyzed on a personal computer, there is increasing recognition that meaningful surgical outcomes research related to policy and clinical practice requires substantial commitment and infrastructure. This chapter will outline the important components of surgical HSR programs, and provide suggestions for program building based on the authors' experiences with the Utah Interventional Quality and Implementation Research (U-INQUIRE) Group at the University of Utah, and Center for Surgery and Public Health at Brigham and Women's Hospital.

B. S. Brooke (✉) · S. R. G. Finlayson (✉)
Utah Interventional Quality and Implementation Research (U-INQUIRE) Group, Department of Surgery, University of Utah School of Medicine, Salt Lake City, UT, USA
e-mail: Benjamin.Brooke@hsc.utah.edu; Sam.Finlayson@hsc.utah.edu

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5_31

31.2 Mission, Vision, and Goals

Surgical HSR programs function most effectively when they are mission-driven. A mission represents the shared purpose of the members of the research group, and helps to guide decision-making and resource allocation. HSR is very broad in scope, and spreading effort and resources across a range of disparate purposes can weaken a research program, just as spreading burning coals can extinguish a fire. Successful surgical HSR programs have typically started by focusing on specific topics that can be studied across different specialties, such as quality of care, system innovation, or regional collaboratives.

The first step in developing a surgical HSR program is to articulate thoughtful mission and vision statements. A mission statement outlines the program's aims, identifies the constituencies that the program serves, and describes how the program is uniquely suited to making its intended contribution. In essence, the mission statement describes why the HSR program exists and what it can do for the department and institution. The value of a mission statement is in its ability to guide resource and effort allocation, and align them with specific program goals. The mission statement also articulates the conceptual framework within which the HSR program functions. In comparison, a vision statement describes the ideal to which the program aspires, and is intended to inspire efforts to meet these objectives.

The mission and vision of the program are ideally translated into actionable strategic and tactical goals. Strategic goals describe broadly the successes that the HSR program would like to attain, such as changing practice across a clinical collaborative, achieving high levels of external funding, or creating a strong surgical research analytic core. Tactical goals describe more specific, easily measured tasks that lead to the achievement of strategic goals, such as successfully competing for a program grant, hiring a talented data analyst, or gaining access to a unique source of data.

A surgical HSR program's mission, vision, and goals should be periodically revisited and revised based on the successes, failures, and evolving strengths of the program. Successful programs are able to adapt to changing circumstances, both inside the organization (e.g. faculty turnover, access to new data sources, etc.) and outside of the organization (e.g. new extramural funding opportunities). Moreover, it is important to be able to periodically shift the program's focus to new opportunities without disrupting existing projects.

31.3 Organization and Governance

The organizational structure of a surgical HSR program is established to allow direct control over critical program assets, such as databases, servers, and the work priorities of staff members. The leader of the HSR program should be given a sufficient budget to hire key personnel, obtain access to computing resources and secure access to data for surgical investigators. Program leaders who are just starting out and or have limited funding are often forced to rely on data sources belonging to other groups, or analyzed by programmers who report primarily to investigators outside the program. This is a situation that should be escaped as soon as possible. When funding is limited to internal sources such as the department, program leaders should find low cost sources of data (see research platforms below), but still hire part or full time analytic and/or project support. Program leaders without their own data and personnel reporting directly to them will often find their projects at the end of someone else's queue.

In addition to a core of faculty investigators, a surgical HSR program will ideally include administrators, project managers, programmers, masters-level data analysts, and doctoral-level biostatistical support. The number of individuals within these categories depends on the size of the program's research portfolio. Specialized talent – such as systems engineers, decision analysts, or clinical coordinators – may also be needed depending on the type of research pursued within the program. Where there are adequate resources and faculty mentorship, research trainees within different doctorate or master's level graduate training programs can add significantly to a surgical HSR program. Trainees often bring new energy and perspective to projects as well as some level of programmatic and analytic support.

Surgical HSR program leaders should strategically direct the allocation of assets and resources, the most valuable of which is their key personnel's time. Leaders should carefully oversee how much time analysts, project managers, and other personnel spend on each research project, and direct these personnel to give greatest priority to the work that is most in line with the mission and goals of the program. Often, projects need to be prioritized based on whether they are externally-funded or simply investigator-initiated by faculty members in the Department of Surgery. Extramural funded research projects are given the highest priority, as both an obligation to the sponsor and to provide the best possible service to surgical investigators who bring funding to the program. But priority should also be given to investigator-initiated projects that have a high likelihood of leading to future extramural grant funding. Often, this is work in which junior faculty members are engaged under the mentorship of program leadership, which will lead to career development awards. This form of department-sponsored research usually represents obligations that the Department of Surgery has made to specific faculty as part of a hiring package.

To manage analytic assets of the HSR program, the project portfolio of each analyst and project manager should be reviewed weekly. This includes reviewing percent time spent on each project for the preceding weeks, providing guidance when necessary to ensure that the highest priority projects got the attention they need, and setting short-term actionable goals (see Table 31.1). In practical terms, available resources (including time) need to be viewed as an investment, and must be directed toward maximizing "returns" most in line with the strategic and tactical goals of the program.

Table 31.1 Sample research project management worksheet

Project name	Principal investigator	Support category	Funded effort (%)	Past week actual effort (%)	1 week ago (%)	Past 4 week average (%)	Action items to complete
PE prevention	Jones	Externally funded	20	18	22	21	Data analysis for aim 1
OR simulation	Jones	Externally funded	20	22	12	19	Data analysis for aim 2
Elderly vascular dz	Wong	Externally funded	15	10	19	16	Complete data tables
DoD free flap	Sargeant	Externally funded	30	27	30	28	Methods for manuscript
Lung resection	Hernandez	Department-sponsored	n/a	5	6	4	Analytic plan for Grant
Thyroid cancer	Abdul	Department-sponsored	n/a	10	3	5	Analytic plan for Grant

31.4 Challenges in Faculty Development

Faculty development is a critical part of any surgical HSR program, but is often difficult to execute. Surgeons seeking to become HSR experts face several challenges; including time constraints, monetary disincentives, limited training opportunities (i.e. fellowships), and a paucity of senior surgical faculty mentors with expertise in HSR.

Compared to other specialties in medicine, surgery has historically demanded greater time commitment to clinical activity, making it difficult for surgeons to find time for research. This challenge is compounded by the common misperception among many academic surgeons that HSR requires little time to perform. Surgical HSR program leaders must convince surgeons who want to pursue HSR (and their clinical leaders) to make the significant time investment required to build a successful surgical HSR portfolio.

Financial disincentives to focus on HSR are also particularly difficult for surgeons to overcome. The gap between research funding for salary support and what a surgeon typically earns with clinical activity is larger than for less generously remunerated specialties: the maximum NIH salary cap at the time of this writing is US$189,600 compared to a typical academic surgeon salary of approximately US$320,000. When 0.20 FTE salary support is awarded in an NIH grant, this would typically cover only about 12% of a typical surgeon salary. This salary gap widens even more for junior faculty on K-awards where they must commit 75% full-time professional effort to their research project and career development activities. In the current tight market for research funding, doing clinical work is by far the easiest way for a surgeon to achieve targets for income generation, whether determined institutionally or personally.

There are a limited number of HSR fellowship training opportunities offered to surgeons, and the ones that do exist are highly competitive. Furthermore, the training paradigm in surgery is challenging for those pursuing a research career. Trainees in the medical specialties can enter research fellowships directly following residency training, which provides a smooth transition to an academic career. In contrast, surgeons typically complete HSR fellowships during "lab years" in the middle of residency, followed by 2–6 more years of clinical training. When surgeons finally start their first faculty job, they must balance the demands of developing a research portfolio with the steep clinical learning curve needed to build an academic surgical practice. This is a common barrier for most young surgeons and can slow down or stagnate their research and academic progress.

Surgical HSR is a relatively young field, but is growing rapidly. Many academic surgery departments aim to recruit surgeon scientists who have the skill set to analyze predictors of surgical quality and outcomes. Faculty members with HSR expertise can significantly contribute to the dialogue related to optimizing healthcare delivery at their own institution as well as on a national level. These factors have placed HSR experts in high demand. But while this environment provides many opportunities for faculty to assume positions of leadership and responsibility early in their careers, it also points to a paucity of senior mentors for the increasing number of trainees and junior faculty who want to pursue surgical HSR.

31.5 Creating a "Faculty Pipeline"

Given the above challenges, creating an effective "pipeline" for faculty is of utmost importance to a surgical HSR program. The essential components of a faculty development pipeline include (1) securing protected research time; (2) identifying training programs and mentorship; and (3) providing access to active intellectual forums and grants administration support.

As with any field of surgical research, a young surgeon entering an academic faculty position needs protected research time. Since very few new faculty have mentored career development training grants at the very onset of their career, this protection typically comes from academic departments in the form of salary guarantees and reduced clinical volume targets. Leaders of surgical HSR programs must work closely with their department chairs and clinical division leaders to ensure that newly recruited research faculty obtain adequate support.

While not absolutely mandatory, formal training in research methods and clinical investigation is extraordinarily helpful to young surgeon investigators. Presently, many junior faculty with an interest in HSR will have completed a masters degree or research fellowship during residency training. If not, there are an increasingly large number of university settings that offer participation in part-time or short-term graduate programs that teach the fundamentals of HSR. Such programs can help fast-track the development of junior surgical faculty members.

Mentoring of junior faculty is perhaps the most critical part of growing and developing a successful surgical HSR program. While junior faculty may have strong analytic skills and bring valuable perspectives and ideas to their work, they often do not know how to leverage these assets to advance their research careers. Faculty mentors not only help junior faculty develop hypotheses and design research, but also help them set realistic career goals, such as funding milestones and academic rank advancement. In addition, mentors should be able to help junior faculty develop important networks for academic engagement and collaboration. Mentoring along the typical research trajectory for a junior faculty member includes helping them identify seed funding for pilot studies early on, and mentored career development funding when possible (e.g. NIH K-level or VA CDA funding). These early mentored efforts ideally provide the basis for applications for higher levels of external grant funding sufficient to support time devoted to independent research (e.g. NIH R-level funding or grants from major foundations).

Faculty development is also facilitated by infrastructure to support grant writing. Competitive grant funders such as the NIH typically provide complicated and daunting grant application processes that can be very time-consuming. The complexity of grant applications is a nearly prohibitive technical and psychological barrier to a surgeon if unaided. To the extent that surgeon investigators can focus exclusively on developing and writing the science of a grant application, they will be more eager and able to pursue them. From the perspective of HSR program building, support of grant writing to make the process as easy as possible for faculty should be seen as an investment in the program. Indeed, if more grant applications are submitted, there will be a higher likelihood of obtaining sufficient funding to support the overall research program.

In addition to individual mentoring and grant-writing support, effective faculty development also requires access to intellectual forums where research ideas, methods, and interpretation of results are exposed to colleagues' constructive critique, and fertilized with new ideas, study designs, and analytic approaches. These forums can take a variety of forms, such as regular research meetings, "work-in-progress" seminars, and interdisciplinary conferences. To build strong intellectual forums for a surgical HSR program, one cannot underestimate the importance of creating effective working space, preferably with offices clustered together (and actually used by the program members), with formal and informal gathering areas. Not only does dedicated space create important opportunities for interaction between researchers, it also provides an escape from competing obligations (e.g. clinical work). While constructing a "virtual" research center is attractive conceptually in a space-constrained academic center, this model is typically unsuccessful in achieving all the goals of a surgical HSR program.

Finally, research fellowship training programs for surgical residents should also be viewed as an important part of the faculty pipeline. Resident research fellows not only bring energy to the research environment and extend the capacity of faculty researchers, but also become the “farm team” for future faculty recruitment efforts.

31.6 Creating a “Funding Pipeline”

Early on, surgical HSR programs typically depend on infrastructure investment from their hosting institutions (e.g. hospital, department, or occasionally other sources of public or private grant funding). Eventually, however, programs are expected to stand on their own financially, or at least incur only minimal ongoing local funding to support the services the program provides its host institution. To gain relative independence financially requires considerable focus on the grant production pipeline.

A funding pipeline starts of course with motivated investigators who have well-articulated and meaningful research plans and teams well-suited to carrying them out. However, the transition to a complete, competitive grant application requires a lot more. At a minimum, administrative support for the “pre-award” process is a critical component of a HSR program. This function requires key personnel with fastidious attention to detail, good communication with internal regulatory bodies (e.g. human subjects review committees, human resources departments), and familiarity with the unique requirements set by a diverse group of funding organizations.

As competition for limited federal funding increases, investigators are increasingly looking toward alternative sources of funding. Surgical HSR programs need to help investigators identify other sources of funding, including private foundations and philanthropy, as well as industry partnerships. Finally, partnerships with healthcare payers and large employers have also emerged as a source of research funding for surgical health services researchers with the goals of understanding and improving surgical value and quality of care for covered populations [1].

31.7 Creating a “Surgical Research Analytic Core”

Once surgical HSR programs become more firmly established in the Department and gain some degree of financial independence, it is usually worth the investment to develop an internal surgical research analytic core. A surgical core comprises both key personnel and data sources to streamline the entire process of conducting HSR for surgical investigators. Key personnel would include staff members that assist with IRB submission, programmers to manage various surgical datasets (including preparation for analysis), and biostatisticians (masters and Phd. level) to work closely with surgical investigators and other core members with the design and execution of statistical analyses. Moreover, data sources maintained by the surgical analytic core may include administrative datasets purchased by the HSR program (e.g. HCUP datasets), surgical registries, institutional datasets, or large working datasets for investigator initiated research projects.

Table 31.2 Step-by-step process for using the surgical population analysis research core (SPARC)

Steps	SPARC workflow process
Step 1	Surgical principal investigator (PI) PI submits a SPARC request form on the Department of Surgery Research website (https://medicine.utah.edu/surgery/research/sparc/).
Step 2	SPARC project manager will arrange a meeting between the surgical PI and core staff including programmer and statistician to discuss research aims, data sources needed, and develop a plan of action. Statistician works with the PI to write the analysis plan.
Step 3	Surgical PI submits final set of research aims to SPARC staff.
Step 4	SPARC provides a cost estimate to the surgical PI to complete the research study.
Step 5	SPARC assists with preparation of IRB application and any data use agreements (DUA).
Step 6	After IRB approval, a research dataset is pulled and cleaned for the surgical PI by the SPARC programmer.
Step 7	SPARC programmer hands off research dataset to surgical PI and statistician for analysis.

Once established, a research analytic core can provide a more seamless pipeline for surgical investigators to carry out their quantitative HSR projects. An example of how this internal process works is provided by workflow map used by the Surgical Population Analysis Research Core (SPARC) in the University of Utah's Department of Surgery (see Table 31.2). This starts by having surgical investigators submit an on-line SPARC request for each individual research proposal, followed by a project meeting with core staff to discuss specific research aims and data sources needed. If the project is deemed feasible and data is available, SPARC assists in the IRB submission and helps prepare a detailed analytic plan. This also includes a cost estimate for all aspects of the study including data usage, programmer time, and statistical analysis. When possible, the same statistician works with the surgical investigator on writing the analytic plan and completing the final analysis.

31.8 Research Platforms

Successful surgical health services research requires an appropriate platform in order to undertake meaningful studies. A useful research platform can take any of a number of forms, including electronic datasets, clinical settings, research networks or collaboratives, learning health systems, or specialized analytic tools. Building a surgical HSR program requires identifying a set of research platforms that are appropriate to the kinds of questions investigators want to answer, and that are within reach given the resources available.

31.8.1 Electronic Datasets

Electronic datasets are a common, often easily accessible research platform, and have traditionally been the backbone of surgical health services research. Because of electronic data accessibility and ease of use, many successful health services researchers have started their careers using large datasets to establish a research niche, and have then graduated to other research platforms as their work has gained momentum and funding has been garnered. The least expensive electronic data is administrative data, which is typically drawn from sources not originally intended for research purposes, such as hospital discharge abstracts or data created for billing. Many such electronic datasets are now organized and augmented to facilitate their use as research tools, such as the National Inpatient Sample (NIS) and other various Healthcare Cost and Utilization Project (HCUP) datasets made available through the federal Agency for Healthcare Research and Quality [2]. In addition to administrative data, clinical registries have become another important source of electronic data. Over the past several decades, several clinical registries have become prominent data sources for surgical HSR including the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) [3], the Society of Thoracic Surgeons (STS) National Database [4], and the Society of Vascular Surgery Vascular Quality Initiative (SVS-VQI) [5]. Other specialized sources of electronic data that are now frequently used in health services research include the US Census, physician workforce data from the American Medical Association, geographical data, and publically reported hospital quality measures.

To provide surgical health services researchers some guidance with using these datasets and others, a comprehensive series of articles focused on surgical datasets was published in JAMA Surgery between June and October of 2018. This included a checklist that can be used to ensure that key conceptual and methodological issues are addressed when using administrative databases and registries for secondary data analysis [6]. In addition, review articles were published as part of this series specific for the ten most common data sources used by surgical health service researchers, each serving as a practical guide to understand the limitations and strengths of each dataset.

31.8.2 “Local Labs”

Surgical health services researchers also find research platforms within their own clinical settings, using local communities, hospital clinics, and operating rooms to examine, measure, and intervene in surgical care delivery. Local clinical settings as a research platform have particular utility in patient-centered outcomes research, wherein measures and outcomes often require new information derived from direct patient contact. The same can be said for clinical decision science, preference assessment, surveys, implementation science, and qualitative research. Local clinical settings are also useful to measure the effect of system design innovations on surgical outcomes and quality. Research using a hospital as a local lab for research has the potential advantage of providing direct benefit to the hospital, which may be willing to provide funds to support it. Simulation centers are also a kind of local lab that can be used as a controlled setting to examine the provider behavior component of surgical care delivery.

31.8.3 Provider Networks and Quality Collaboratives

Provider networks and collaboratives created to measure and improve quality and value of care are increasingly used as platforms for health services research. Examples include the Northern New England Cardiovascular Disease Study Group [7], the Michigan Surgical Quality Collaborative [1], and the Surgical Care and Outcomes Assessment Program (SCOAP) [8] in Washington State. These types of regional quality collaboratives provide a platform for research and have resulted in important discoveries related to surgical care delivery and improving patient outcomes. An illustration of this synergy is the Michigan Opioid Prescribing Engagement Network (OPEN), that used the statewide collaborative to develop and implement guidelines for post-operative opioid prescribing [9].

A national network for research collaboration, known as PCORnet, was recently established with funding from the Patient-Centered Outcomes Research Institute (PCORI). PCORNet is a network of healthcare institutions across the US focused on collecting and analyzing patient-centered and patient-reported outcome (PRO) data [10]. PCORnet is organized into Collaborative Research Groups (CRGs) that are focused on collecting data and studying specific health-related conditions within a multi-hospital research network. Each participating site transforms its data into a Common Data Model format in order to facilitate the collection of data for comparative effectiveness research.

31.8.4 Learning Health Systems

A learning health system (LHS) is another platform for surgical HSR that builds upon the data networks established by local labs and/or research collaboratives [11]. The Institute of Medicine (IOM) conceived of a LHS as a health care system in which "science, informatics, incentives and culture are all aligned for continuous quality improvement and innovation, with best practices seamlessly embedded in the delivery process and new knowledge captured as an integral by-product of the delivery experience." In a LHS, the data generated from clinical practice create new knowledge, and that knowledge is directly fed back into the system to change and improves how we practice. Surgical HSR can be undertaken within this perpetual LHS cycle, which is designed to systematically gather and analyze outcomes on every patient. Each time care processes are refined in response to patient outcomes, this feedback cycle allows an opportunity for outcome assessment and research analysis.

31.8.5 Analytic Tools

A large body of health services research has been built on the use of analytic tools that synthesize information drawn largely from medical literature, including most notably meta-analysis, decision analysis, and cost-effectiveness analysis. Expertise in these methods can serve as a platform for health services research to refine research questions and motivate further work using other research platforms.

Surgical health services researchers also benefit from collaborating with experts in bioinformatics and computer science, and adopting innovative health information technology (IT) tools. This includes developing dashboards, apps, and other IT-based tools that allow data to be captured as well as displayed for research and clinical purposes. Finally, there is growing interest within the surgical HSR community to utilize natural language processing technology and other forms of artificial intelligence to extract discrete information from written text in clinical notes in order to develop more granular sources of data [12].

31.9 Collaborations

Building a health services research program is greatly facilitated by the ability to create collaborations across research disciplines and expertise. While a single investigator with a little statistical knowledge, an electronic dataset, and a personal computer can write and publish plenty of papers, the most meaningful health services research typically draws from a range of collaborations with other investigators who bring a variety of skills and knowledge to bear on the targeted research questions. Furthermore, multidisciplinary collaboration is almost always necessary to pursue extramural funding in health services research.

While collaboration in surgical health services research once meant finding other surgeon investigators to join in a project, many successful health services research programs have discovered significant benefit in working side-by-side with non-surgeon health services researchers who provide different perspectives, suggest alternate study designs and analytic methods, and provide important opportunities for junior investigators to find research mentorship. As surgical health services research has become more sophisticated, surgical investigators have benefited from finding collaborators across a very broad range of expertise, including economists, psychologists, anthropologists, sociologists, systems engineers, and experts in informatics, biostatistics, qualitative methods, management, and health policy.

31.10 Conclusion

In summary, building a surgical HSR program requires careful planning, resources and strategic relationships at your institution. This usually starts by surveying the local academic landscape to identify what resources are available in terms of potential funding, research platforms, and expertise, then creating a clear vision with goals that are well aligned with the inventory of assets available. Surgeons with HSR training and expertise often need to be recruited to the program, but existing faculty and residents with the right talents should also be identified and cultivated. Successful surgical HSR programs build bridges to collaborators and mentors with valuable expertise, invest available funding and resources judiciously to maximize their return, and identify and pursue potential internal and external funding sources through a carefully constructed and meaningful research agenda. Ultimately, success in the endeavor comes to those who work hard, collaborate well, genuinely care about the questions their research aims to answer, and are prepared to benefit from good providence when it comes their way.

References

1. Birkmeyer NO, Share D, Campbell DA, et al. Partnering with payers to improve surgical quality: the Michigan plan. Surgery. 2005;138:815–20.
2. Stulberg JJ, Haut ER. Practical guide to surgical data sets: healthcare cost and utilization project national inpatient sample (NIS). JAMA Surg. 2018;153(6):586–7.
3. Raval MV, Pawlik TM. Practical guide to surgical data sets: national surgical quality improvement program (NSQIP) and pediatric NSQIP. JAMA Surg. 2018;153(8):764–5.
4. Farjah F, Kaji AH, Chu D. Practical guide to surgical data sets: society of thoracic surgeons (STS) national database. JAMA Surg. 2018;153(10):955–6.
5. Desai SS, Kaji AH, Upchurch G Jr. Practical guide to surgical data sets: society for vascular surgery vascular quality initiative (SVS-VQI). JAMA Surg. 2018;153(10):957–8.
6. Haider AH, Bilimoria KY, Kibbe MR. A Checklist to Elevate the Science of Surgical Database Research. JAMA Surg. 2018;153(6):505–7.
7. O'Conner GT, Plume SK, Olmstead EM, et al. A regional prospective study of in-hospital mortality associated with coronary artery bypass grafting. JAMA. 1991;266:803–9.
8. Flum DR, Fisher N, Thompson J, et al. Washington State's approach to variability in surgical processes/outcomes: surgical clinical outcomes assessment program. Surgery. 2005;138:821–8.
9. Howard R, Waljee J, Brummett C, Englesbe M, Lee J. Reduction in opioid prescribing through evidence-based prescribing guidelines. JAMA Surg. 2018;153(3):285–7.
10. Selby JV, Grossman C, Zirkle M, Barbash S. Multistakeholder engagement in PCORnet, the National Patient-Centered Clinical Research Network. Med Care. 2018;56(suppl 1):S4–5.
11. Guise J-M, Savitz LA, Friedman CP. Mind the gap: putting evidence into practice in the era of learning health systems. J Gen Intern Med. 2018;33(12):2237–9.
12. Chapman AB, Mowery DL, Swords DS, Chapman WW, Bucher BT. Detecting evidence of intra-abdominal surgical site infections from radiology reports using natural language processing. AMIA Annu Symp Proc. 2018;16:515–24.

Index

J. B. Dimick, C. C. Lubitz (eds.), *Health Services Research*, Success
in Academic Surgery, https://doi.org/10.1007/978-3-030-28357-5