Chapter 14
Evidence-Based Surgery

Toby A. Gordon

With the national focus on patient safety, the importance of understanding the principles of evidence-based surgical practice and outcomes research has never been greater. The surgery literature manifests the continuous striving of the surgical profession to improve the delivery of healthcare and develop and incorporate new technologies into practice. Advances such as imaging and minimally invasive interventions have blurred boundaries among some specialties. Patient safety has become an area of national focus and priority. These trends, coupled with the rising demand for surgical services, the close scrutiny of health-care costs by government and managed care, consumers' increasing independence in selecting their own care, and, especially of late, the malpractice crisis in the United States, require all surgeons to understand the development and application of evidence-based surgery and surgical outcomes research.

Evidence-based surgery draws from the fields of medicine and public health and centers on the acquisition, evaluation, and application of evidence for the care of the individual patient. More broadly defined beyond the care of the single patient, evidence-based surgery also encompasses population-based outcomes research focused on clinical, economic, and patient-reported data. This chapter was written to assist surgeons in producing and interpreting “evidence” related to surgical practice, for use in educational and patient care forums alike.

SURGICAL EVIDENCE: A HISTORICAL PERSPECTIVE
Throughout the history of surgery, surgical evidence has progressed from anecdotes and observations, case studies and reports of patient series, to sophisticated laboratory, clinical, and epidemiologic research. The modern
surgical era began in the mid-19th century, when major obstacles to the progress of surgery—pain, infection, and hemorrhage—were eliminated.

The proceedings of the American Surgical Association (ASA), founded in 1880, serve as a record of the development of surgery in America and allow a historical perspective on the types of surgical “evidence” in use. Major advances in surgery reported from 1880 to 1930 reflect the shifting focus of the surgical research. Technologic developments such as radiation, direct examination of the urinary tract by cystoscopy, blood transfusion, and electric cautery and improvements in operations such as thyroidectomy, gastric resection, and craniotomy resulted in decreased mortality. Early research reports were descriptive and anecdotal but reflected interest in outcomes of surgical treatment. From 1880 to 1942, new operative approaches often were developed on the basis of intuition and insight and evaluated primarily on the basis of trial and error.

The turning point in the evolution of evidence-based surgery can be attributed to Ernest Amory Codman who, in 1910, began efforts to reform clinical medicine and surgery. With his surgical mentor, F. B. Harrington, Codman created a case monitoring system in 1900 to record outcomes. Codman's “end-results” system proposed individual patient cards for data collection, on which a determination regarding outcome would be recorded. Cases in which “perfection” was not obtained would be classified according to his “nosology of errors.” By using his system of comparison, Codman reasoned, surgeons could specialize in operations they did best. His approach exemplified the application of scientific management principles of industrial efficiency to the practice of medicine.

Major influences on evidence-based surgery since World War I include the public health movement and the evolution of public policy related to health-care delivery, military medicine's contributions from World War II, and clinical and health services research in practice variation. The introduction of employer- and government-sponsored health insurance had an important impact on evidence-based medicine. Following World War II, the government freeze on wages brought about the provision of health insurance as a form of additional compensation. By 1966, much of the nonworking population had been given the basic right to health benefits under the federally sponsored Medicare coverage for the elderly and the joint federal and state Medicaid program for the poor. The Medicare program brought about peer review, which gave way to performance improvement activities under the mandate of the Joint Commission on Accreditation of Healthcare Organizations, further stimulating interest in outcomes research. The Medicare prospective payment system, implemented in 1983, led to the widespread use of diagnosis-related groups; the creation of comprehensive claims databases
now used for economic and clinical outcomes research; and the institution of “centers of excellence,” designated cost-effective providers for coronary artery bypass surgery and organ transplantation, based on clinical and outcomes measures. In the 1980s, the Centers for Medicare and Medicaid Services (then the Health Care Financing Administration) publicly disseminated mortality data developed from Medicare claims files, a practice ultimately discontinued following debate over risk adjustment of data.

In the 1920s and 1930s, tonsillectomies came under scrutiny, due to the widespread use of this procedure and the anesthesia-related mortality risks. The next notable study of variation in practice was Wennberg's 1973 research comparing surgical procedure rates in 13 hospital service areas in Vermont. He found that physicians' preferences were the greatest influence on rates of tonsillectomies, appendectomies, hysterectomies, mastectomies, hemorrhoidectomies, and surgeries for other common conditions, with considerable variation in rates of surgery across service areas.3

Since the 1960s, increased federal and private-sector funding has supported further clinical research. Surgical outcomes research in the 1980s and 1990s tested the efficacy and effectiveness of new surgical procedures; examined clinical outcomes and cost-effectiveness of new procedures versus existing ones; compared surgical treatment with medical treatment; analyzed the relationship between volume and outcome; and, increasingly, focused on quality-of-life considerations. Refinements in clinical and quality-of-life outcomes measurements have contributed to these areas of study. Other burgeoning areas of research include the use and evaluation of critical pathways and practice guidelines. The managed care movement provided a powerful stimulus for quality and outcomes measurement.

The most significant recent influence on evidence-based surgery comes from the Institute of Medicine's (IOM's) 1996 health-care quality initiative and its 1998 report documenting “the serious and pervasive nature of the nation's overall quality problem.”4 The IOM has identified the overuse, misuse, and underuse of health-care services and laid out a vision for the transformation of the health-care system and health policy to achieve quality care.

Evidence-based surgery in its current form manifests the influences of historic public health concerns, the IOM quality initiative, ever more sophisticated medical treatments and interventions, the increasing application of information technologies in medical care, the economic and political forces that shape health insurance, managed care and health-
care reform, and the growing interest of consumers in self-directing their personal care. Ultimately, evidence-based surgical research should enable stakeholders in the health-care delivery process—patients, physicians, payers, and policymakers—to provide the highest quality, most appropriate, and most cost-effective care possible.

FRAMEWORK FOREVIDENCE-BASED SURGERY

Quality Defined
The IOM defines quality as the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge. Surgical outcomes research studies the quality of surgical care, from the selected treatment approaches and clinical interventions to systems of care delivery, with results of interest to physicians, patients, and policymakers alike.

Measures of the Performance of Health-Care Delivery
Traditional measures of the performance of health-care delivery systems focus on access to care, quality, and cost. Donabedian, the foremost researcher in medical quality assessment, defined a conceptual model in which the outcomes of healthcare are measured by examining the structures and processes of care as well as patient factors and risks. Structural features comprise patient, provider, and payer characteristics; process-of-care measures describe what was done to and for the patient; and outcomes are classified as clinical and physiologic, patient-reported, and economic. Key considerations in evidence-based surgery include the measurement and interpretation of outcomes and the design of studies to obtain them.

Outcomes Measurement
Guice described successful outcomes measurement as documenting changes in clinical condition as a result of medical intervention; collecting data in a common format; maintaining data collected from multiple clinical sites in a single site to facilitate comparison of outcomes; incorporating standardized and validated methods of accounting for a health-care organization’s effect on health and quality; enabling physicians to assess and select medical treatments and accurately predict the resources needed for care on the basis of the actual results and cost of a treatment; providing data to establish standards or guidelines for treatment; and providing patients with specific facts to help them make medical decisions.
decisions, including facts concerning treatments and their cost, efficacy, and impact on quality of life.

Comprehensive measurement of outcomes includes clinical, economic, and patient-reported data.

**CLINICAL OUTCOMES MEASUREMENT**

Much of the evidence-based surgery literature assesses clinical and physiologic measures, which may include symptoms, anatomic descriptors and physical signs, physiologic or functional data as measured or observed by clinicians at a specific point in time or within a specified time period, and the end results of clinical care as measured by morbidity and mortality. Outcomes measures can be taken at the patient, study group, or population level. The most commonly used clinical measures of the consequences of surgical therapy are rates of mortality and morbidity, gains in life expectancy, relative risk, relative and absolute risk reduction, and potential impact in absolute numbers.

**Mortality**

Mortality is the most reliably measured clinical outcome. It is most meaningfully expressed as the proportion of deaths from a particular cause over a defined time interval and most reliably measured from death certificates. *Postoperative mortality* generally refers to death within 30 days after a procedure. However, because patients today commonly have hospital stays that are far shorter than a month, care must be taken to differentiate between *in-hospital* and *postoperative* mortality. In *mortality rates*, the denominator represents the entire population at risk of dying from the disease, that is, those with a disease or condition or at risk for the condition. *Case fatality rate* includes only those with the disease in the denominator. *Gain in life expectancy* is generally discerned from life-table analysis, but its interpretation can be problematic without carefully designed clinical trials that control for competing explanations for the gains.

The *relative risk* is the ratio of probabilities of adverse outcomes in two treatments being compared. We interpret relative risk and the odds ratio the same way, and under certain conditions the two rates approximate each other. Alternatively, retrospective study designs commonly measure odds ratios rather than relative risk. *Relative risk reduction* is a treatment’s reduction of adverse clinical outcomes due to the progression of disease. It is expressed in the difference in event rates between the control and treatment groups, divided by the event rate in the control group. Relative risk reduction does not reflect the magnitude of the risk without therapy, and thus it will overestimate or underestimate the effect.
of therapy when adverse events in untreated patients are very rare or very common, respectively. The absolute risk reduction (also known as the attributable risk reduction) is the difference in event rates between the control and treatment groups. As an expression of the consequences of giving no treatment, it provides an additional measure of clinical effect. Last, the number needed to treat is the number of patients who must be treated to prevent one adverse event. The shortcomings of the foregoing measures of clinical benefit result from the properties of the measures themselves, as well as from the data used. Any measure of the benefit of treatment may vary considerably in different trials of the same or similar therapy because of different patient populations, study design, or chance; thus, the applicability of results of a study must be evaluated carefully. Because trials are of finite duration, the effects of continuing therapy beyond the period of the trial are not known. Some treatments may not be effective until long after they have been started, and there needs to be an adequate duration of follow-up.  

**Morbidity**

Mortality has less meaning in the study of surgical procedures in which death is an extremely rare event; therefore, it is more meaningful to report morbidity, often referred to in the surgical literature as complications. Morbidity measures the presence of illness and the degree of dysfunction, which can be assessed as days of work missed or bed-disability days. Because surgical complications can range in severity from simple wound infections to life-threatening conditions, they should be reported separately, considering the underlying procedure or treatment of study. Generic surgical complications—such as wound infection, pneumonia, urinary tract infection, and bleeding requiring blood transfusion—should be examined in addition to disease- or procedure-specific complications. Complications are coded according to the International Classification of Diseases (ICD), now in its tenth revision. Because coding categories and regulations change from one revision to another, any study of time trends in morbidity that spans more than one revision must examine the possibility that observed changes could be due entirely or in part to changes in the ICD. Changes in disease definition can also have a significant effect on the number of cases of the disease that are reported or that are reported and subsequently classified as meeting the diagnostic criteria for the disease.

Clinical signs and symptoms are important but subjective measures of morbidity. Care must be taken to standardize definitions. When reporting a test value as a clinical outcome, it is essential to define the normal range for the test, because definitions might vary among laboratories or
institutions. In addition, it is important to define clearly the clinical context in which the test was ordered, so that interpretation of the result is appropriate. Trends in test results are as important as their absolute values, because changes over time may represent return of disease and poor clinical outcome.

**ECONOMIC OUTCOMES MEASUREMENT**

Concerns about the costs of healthcare have dramatically increased the demand for economic outcome measures, which quantify the costs and/or benefits of medical and surgical care. In the evidence-based surgery literature, clinical measures of outcome such as mortality or treatment complications are reported much more frequently than are economic measures, either alone or in combination with clinical measures. This scarcity of information may be attributed not only to the interests of investigators but also the difficulties in measuring economic outcomes.

Key parameters for economic analysis include the time period for analysis, the breadth of services provided, and the perspective from which costs are defined, that is, the patient's, the provider's, or the payer's. Ideally, an economic analysis would include long-term comparisons of the costs of all services from the perspectives of all stakeholders (patient, provider, payer, and society). In actuality, the scope of analysis is much narrower; for example, an examination of the perioperative mortality for high-risk surgical procedures may limit the cost analysis to inpatient length of stay, as a proxy for cost. Because the benefits and ultimate costs of a surgical intervention may not be realized until years later, studies taking only hospitalization into account do not reflect all societal costs. Thus, a longer time frame of analysis is desirable, such as an episode of illness, or 1-year or longer periods. The ideal economic measure of costs from the societal perspective is the social opportunity cost of the inputs to the health-care process, that is, the highest value the inputs could earn if used for other purposes. Analyses that compare the utility of additional spending for health-care services with the utility of other societal needs are infrequently reported. Such economic research is generally performed for purposes of making policy. An example is the effort of the Oregon Medicaid program in 1994 to rationalize service delivery by prioritizing all services provided to beneficiaries, based on cost and utility analysis.

Accounting for the costs of health-care services is a complex process. Reimbursement methodologies provide incentives to classify and allocate costs differently than would be desirable for outcomes research. Additionally, the data reflect what providers are paid for services rather than the true economic costs of services. Economic measures for clinical outcomes studies may be developed in several ways. Cost analysis can be
completed prospectively as part of a clinical trial. Retrospective cost data can be analyzed typically based on secondary or administrative databases. In other instances, standardized or estimated costs may be applied to models of clinical outcomes. Typically reported costs include hospital and physician charges, obtained from billing data. Administrative databases are also used for standard cost data from Medicare claims and cost reports.

**PATIENT-REPORTED OUTCOMES MEASUREMENTS**

Surgical procedures aim at improving the quality of life as well as prolonging life, and surgical studies have increasingly examined patient-reported outcome measures such as patient-reported health status or health-related quality of life, including functioning and well-being, and patient satisfaction with healthcare. These data are usually collected using standardized questionnaires or surveys. Most surgical studies before 5 years ago neglected to collect standardized data about patient-reported health status and quality of life.

Health status, functional status, quality of life, and health-related quality of life are terms used almost interchangeably to refer to the concept of patient reports of their own health. In 1948, the World Health Organization defined *health* as “a state of complete physical, mental, and social well-being, and not merely the absence of disease and infirmity.” This definition reflects the multidimensional nature of health, and that it has both positive and negative aspects. Bergner identified five dimensions of *health status* (Table. 14-1).

**TABLE 14.1 BERGNER’S FIVE DIMENSIONS OF HEALTH STATUS**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic and inherited characteristics</td>
<td>Genetic and inherited characteristics</td>
</tr>
<tr>
<td>Biochemical, physiologic, and anatomic condition, including</td>
<td>Biochemical, physiologic, and anatomic condition, including</td>
</tr>
<tr>
<td>impairment of these systems, disease, signs, and symptoms</td>
<td>impairment of these systems, disease, signs, and symptoms</td>
</tr>
<tr>
<td>Functional status, including performance of the usual activities of daily</td>
<td>Functional status, including performance of the usual activities of daily</td>
</tr>
<tr>
<td>living, such as self-care, physical activities, cognition, and work</td>
<td>living, such as self-care, physical activities, cognition, and work</td>
</tr>
<tr>
<td>Mental condition, which includes positive and negative emotions</td>
<td>Mental condition, which includes positive and negative emotions</td>
</tr>
<tr>
<td>Health potential, including prognosis for longevity and future functioning</td>
<td>Health potential, including prognosis for longevity and future functioning</td>
</tr>
</tbody>
</table>
Quality of life is a broad concept that encompasses a person's experience and assessment of aspects of life. Health-related quality of life encompasses several dimensions of health status that are directly experienced by the person, including physical functioning, psychological well-being, cognitive functioning, social and role functioning, and general health perceptions. The patient's symptoms are often also included under this definition.

There are two basic approaches to quality-of-life assessment: generic and disease-specific. Generic instruments are designed for use across different diseases, treatments, settings, and patient groups. The major advantage is that they can be used in any population and allow comparisons of the relative impact of various health interventions. However, they may be unresponsive to changes in specific conditions and may be too general to guide clinical decision making. Disease-specific measures focus on dimensions of health related to a particular disease, population, symptom, or problem and may be more responsive to a change in the patient's condition than a generic instrument.

Health profiles attempt to measure multiple important dimensions of health-related quality of life. For example, the Sickness Impact Profile assesses a physical dimension (including ambulation, mobility, body care, and movement); a psychosocial dimension (including social interaction, alertness behavior, communication, and emotional behavior); and domains such as eating, work, home management, sleep and rest, and recreation and pastimes. The SF-36 Health Survey (Ware, 1992) is a brief (36-item), widely used questionnaire that assesses general health perceptions, physical functioning, role limitations due to physical health, role

limitations due to mental health, social functioning, pain, mental health, and energy.\textsuperscript{16} The \textit{Quality of Well-being Scale}\textsuperscript{17} is a widely used instrument that combines questions about various dimensions of functional status to generate a score. The Duke Health Profile is an abbreviated version of the 63-item Duke-UNC Health Profile, developed in 1990, which generically measures patient health status in primary care settings.\textsuperscript{18} Intended for use in evaluative studies and policy research, the European Quality of Life (EuroQol) measure, developed in 1990 and revised in 1993, shows health status as a single index score.\textsuperscript{18} The European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ) of 1993 is a standardized system for evaluating the quality of life of cancer patients as research subjects in international clinical trials.\textsuperscript{18}

\textit{Descriptive or psychometric measures} are based on the patient's report or rating of their health state on a continuum. \textit{Utility measures}, derived from economic and decision theory,\textsuperscript{19} refers to the value placed by the individual on a particular health state. Utility is summarized as a score ranging from 0.0, representing death, to 1.0, representing perfect health. In economic analyses, utilities are used to justify devoting resources to a treatment. Because they weight the duration of life according to its quality, they can be used to generate quality-adjusted life-years. However, because they are expressed as a single score, they do not provide detail about how specific aspects of patients' lives are affected.\textsuperscript{12}

\textit{Patient satisfaction} refers to patients' subjective evaluations of their healthcare.\textsuperscript{20} Patient ratings of care reflect what they think is important about the quality of care, including the doctor-patient relationship and their perception of the adequacy of diagnosis and therapy. They predict patients' subsequent behavior, including how well they comply with medications prescribed, whether they return or go elsewhere, and whether they recommend a physician to others.\textsuperscript{21} The Patient Satisfaction Questionnaire (PSQ)\textsuperscript{22} and the Medical Outcomes Study Nine-Item Visit Rating Form\textsuperscript{23} are examples of instruments that assess general medical care and specific physician visits. The Consumer Assessment of Health Plans (CAHPS) surveys are intended to assess health plans and services and help consumers select among them.\textsuperscript{24} A new federal initiative, the Hospital CAHPS, or HCAHPS, aims to assess patient satisfaction with inpatient care.\textsuperscript{25} The methodology and results of the 25-item survey instrument, developed by the Centers for Medicare and Medicaid Services in partnership with the Agency for Healthcare Research and Quality, will be made available to the public. Although there are no surgery-specific questions, pain control is an area of focus. There are few, if any, established measures of patient satisfaction with surgical care.
Quality-of-life assessments can have relevance to surgical research and practice for defining the indication for surgery, for monitoring the patient, and for evaluating the impact of treatment. Surgical quality-of-life assessments should be made when different treatment alternatives might affect the patient's quality of life differently, when new interventions are implemented, when there is scarcity of resources, when the timing of an operative intervention must be determined, and when improving quality of life is the goal of intervention. When surgery is clearly life-saving or is the only treatment alternative, quality-of-life assessment may be less important. Because early clinical trials that examined the impact of surgical treatment on quality of life, studies have helped to identify treatments that are preferable based on both decreased morbidity and cost-effectiveness. At the present time, quality-of-life assessment has evolved into a crucial component of clinical trials of new and existing treatments, as well as of cohort studies. Selecting an appropriate quality-of-life measure for a specific surgical problem requires a clear formulation of the question to be answered, consideration of the factors that must be assessed, review of available instruments, review of the evidence for usefulness of instruments in a comparable population, and examination of practical considerations.

**STUDY DESIGN**

The quality of the evidence and the strength of the study design are critical factors in conducting and evaluating outcomes research. Determining causal links between two variables (to establish which procedures are most effective) requires randomized clinical trials or matched-pair experimental studies with blinding. The number of participants needed in a randomized controlled trial is often greater than can be recruited at any one center. The multicenter randomized clinical trial was developed as the means to deal with such circumstances. In this study design, the same randomized clinical trial is conducted simultaneously at several different clinical centers. The clinical centers recruit the participants, collect the data, and administer the treatments, but the randomization of assignment and the data analysis from the trial are performed by a coordinating center. Randomization takes place within each clinical center, and the treatments are distributed at each clinical center.

Randomized clinical trials are generally accepted as the definitive approach for assessing the efficacy of a new treatment. The process of randomization, when properly implemented, provides the means by which the myriad factors that may influence the results of a trial are equally distributed between the experimental treatment group and the control, or “usual care,” group. The design may provide information on the natural
history of a disease during both usual care and the experimental
treatment. It also addresses several weaknesses in the randomized
clinical trial. The costs of such studies are generally considerable. It is
impossible to subject all new therapies to a randomized clinical trial
evaluation, in part because of those costs. Such studies also require
considerable time, both for recruitment and, frequently, to obtain the
outcomes of interest. There are also instances in which undertaking a
randomized trial is simply not ethical—for example, withholding
appendectomy after a ruptured appendix to determine whether antibiotic
treatment alone was efficacious. Also, randomized clinical trials are not
based on random samples of the population of patients. The investigators
in a given trial may seek to exclude all but a very specific subset of
patients with a particular disease. It is therefore often difficult to
generalize the results of a randomized clinical trial to the population of
patients with that particular disease. These weaknesses must be carefully
weighed against the considerable strengths of this type of study.

With respect to surgery, experimental studies may be impractical because
the pace of the introduction of a new surgical technology or technique
outstrips the ability of surgical investigators to conduct a randomized
controlled trial to evaluate its effectiveness. Experimental study designs
may be limited in other ways as well: patients recruited for a clinical
study may be dissimilar to the population to which investigators wish to
generalize results, and participating centers may have profound
differences from nonparticipating centers—for example, centers studying
the efficacy of a new surgical procedure may have surgeons more skilled
in that procedure.

When clinical trials are impractical or unavailable, quasiexperimental or
observational studies can be valuable but must be interpreted with
greater caution. Even if recruited randomly, patients may fail to enroll or
respond due to factors that may be related to outcomes, such as the
effect of illness on compliance with the study protocol.

Observational studies differ from randomized clinical trials in that the
investigator merely observes the result of the exposure to the factor of
interest, as the investigator does not assign the exposure. There are two
varieties of observational studies, cohort studies and case-control studies.
In cohort studies, cohorts of individuals exposed to the factor of interest
and those not so exposed are recruited and followed by the investigator
for a period of time to see whether the outcomes of interest occur. In a
case-control study, persons with the outcomes of interest (the “cases”) are
recruited, as are persons who do not have that outcome (the “controls”).
Both groups are queried with regard to their past exposure to the factor
(s) of interest. The investigator then determines if an association exists between the exposure and subsequent outcome. The findings may reveal a relationship between exposure and outcome, but a causal link is difficult to determine in this study design, due to compounding variables the study design cannot account for. Cohort studies have been part of the evaluation of surgical procedures for much of the past century.

In the case report, or case series, a major component of clinical research in surgery, the clinician examines the response of a well-characterized disease to a new treatment. Other clinicians can then consider these data in formulating treatment plans for their own patients. Case reports present a variety of advantages, including the speed and ease with which the data may be compiled, the facility with which most clinicians can relate to the information in a given report, and the ability of all clinicians to contribute to the corpus of medical therapeutics without intensive research training. Disadvantages of case reports include the lack of a comparison group, the small sample size, and the lack of risk adjustment for comparisons to other studies and populations.

**Levels of Evidence**

Regardless of the study design employed, certain fundamentals apply to the analysis of all data. These are best considered within the context of the new evidence-based surgery movement, which has been described as “the conscientious, explicit, and judicious use of the current best evidence in making decisions about the care of individual patients; the integration of individual clinical expertise with the best available external clinical evidence from systematic research.”\(^{28}\) Such “clinical evidence” from systematic research can be of variable quality. One of the most widely used and respected classifications of levels of evidence is that of the United States Preventive Services Task Force (Table 14-2).\(^{29}\) Much of the clinical decision making performed by the average practicing surgeon falls outside the realm of evidence-based medicine. Thus, the practice of surgery represents the use of extensive clinical experience and contemporaneous research findings to determine the most appropriate treatment for an individual patient.

---

**TABLE 14.2 UNITED STATES PREVENTIVE SERVICES TASK FORCE CLASSIFICATION OF LEVELS OF EVIDENCE**

<table>
<thead>
<tr>
<th>Level</th>
<th>Quality of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
As a primer on the critical assessment of outcomes studies, the *Journal of the American Medical Association*’s comprehensive series of articles on evidence design and analysis consideration constitutes an excellent resource. Briefly, the recommended framework for critical appraisal of the quality and of a research paper and its applicability to the care of an individual patient is based on the questions in Table 14-3.

### Meta-Analysis

In contrast to the study designs discussed previously, in which investigators collect data to answer questions regarding the outcomes of specific surgical procedures or related aspects of those procedures (such as prophylaxis), meta-analysis assembles existing research findings to provide an aggregate view. In a meta-analysis, the investigator reviews the literature for all relevant studies regarding a given surgical procedure and a specific outcome. The number of subjects and the strength of the association between the procedure and the outcome are recorded for each study. Then, an aggregate strength of the association is calculated using one of the

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Evidence obtained from at least one properly conducted, randomized, controlled trial (RCT)</td>
</tr>
<tr>
<td>II-1</td>
<td>Evidence obtained from well-designed controlled trials without randomization</td>
</tr>
<tr>
<td>II-2</td>
<td>Evidence from well-designed cohort or case control analytic studies, preferably from more than one center or research group</td>
</tr>
<tr>
<td>II-3</td>
<td>Evidence obtained from several time series with or without intervention, or dramatic result in uncontrolled experiments</td>
</tr>
<tr>
<td>III</td>
<td>Opinions of respected authorities based on clinical experiences, descriptive studies and case reports, or reports of expert committees</td>
</tr>
</tbody>
</table>

As a primer on the critical assessment of outcomes studies, the *Journal of the American Medical Association*’s comprehensive series of articles on evidence design and analysis consideration constitutes an excellent resource. Briefly, the recommended framework for critical appraisal of the quality and of a research paper and its applicability to the care of an individual patient is based on the questions in Table 14-3.
statistical techniques that have been developed for this purpose. Conceptually, the estimates are weighted by the number of subjects in each study; the larger the study, the more weight that is given to that estimate in the calculation.

<table>
<thead>
<tr>
<th>TABLE 14.3 FRAMEWORK FOR CRITICAL APPRAISAL OF RESEARCH’S APPLICABILITY TO PATIENT CARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the assignment of patients to treatment really randomized?</td>
</tr>
<tr>
<td>2. Were all clinically relevant outcomes reported?</td>
</tr>
<tr>
<td>3. Were the study patients recognizably similar to your own?</td>
</tr>
<tr>
<td>4. Were both clinical and statistical significance considered?</td>
</tr>
<tr>
<td>5. Is this therapeutic maneuver feasible in your own practice?</td>
</tr>
<tr>
<td>6. Were all patients who entered the study accounted for at its conclusion?</td>
</tr>
</tbody>
</table>

Meta-analysis is based on the assumptions that the quality of the individual studies are the same, that the factors examined in the studies are the same, that the data missing from any one study will not be prejudicial for the outcomes of interest, that the populations from which the study subjects were drawn are similar, and that the definitions used among the studies are the same.60-62 There is also the assumption that all studies involving the factor and the outcome are known to the investigators.60 Frequently, this requirement means that the investigator must know about all studies conducted regarding a factor and an outcome.63 Because studies that do not attain statistical significance are not published as frequently as those that do reach it, some bias (“publication bias”) may attend the results of the meta-analysis. Even when the investigator is aware of such studies and is able to include them in the meta-analysis, it is generally difficult to be certain that all such studies have been included.64 Because all of these assumptions may not be satisfied, when interpreting the results of any meta-analysis one must consider the degree to which they were violated.

MEASUREMENT
Reliability and Validity

The quality of evidence is determined by the reliability and validity of the outcomes measurement.

Reliability refers to a measure's consistency or repeatability, that is, whether it gives the same result repeatedly when the same thing is measured. For clinical research and quality assessment, reliability is often measured by the criteria in Table 14-4.

### TABLE 14.4 CRITERIA FOR MEASURING RELIABILITY

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-retest reliability</td>
<td>Repeated use of the same measure on the same subject, yielding the same value results, when the property measured is something that should be stable over the time between the two measurements</td>
</tr>
<tr>
<td>Inter-rater reliability</td>
<td>Consistent results when several observers or judges obtain the information or make judgments</td>
</tr>
<tr>
<td>Measurement bias</td>
<td>Prejudicial influences such as investigator hypotheses and beliefs, rater tendencies, recall bias, and others&lt;sup&gt;65&lt;/sup&gt;</td>
</tr>
<tr>
<td>Random measurement error</td>
<td></td>
</tr>
</tbody>
</table>

Validity refers to whether or not a measure reflects what it is intended to measure. Validity is a function of reliability, inasmuch as unreliable measures cannot be valid, yet reliable measures may lack validity due to built-in sources of bias. For example, a scale may reliably or repeatedly yield the same weight but be inaccurate, thus invalid. Tests of the validity of measures used to collect evidence about the effectiveness or quality of surgical care are outlined in Table 14-5.

Before acting on evidence, surgical providers should be convinced that it is valid, or accurate. Several aspects of study methods can help determine
whether a study is likely to be valid, including the study design, the sampling, the completeness of the conceptual model, the measurement and reporting of confounding variables, and measurement reliability and validity.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face validity</td>
<td>Stakeholder perception that the measurement is likely to obtain accurate results</td>
</tr>
<tr>
<td>Content validity</td>
<td>Whether all the important content that is part of measuring an outcome is included in the measurements</td>
</tr>
<tr>
<td>Predictive validity</td>
<td>Whether the measurement has been demonstrated to predict future events or outcomes</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>Similar to predictive validity, but relates subsequent outcomes to a gold standard finding already proven or known</td>
</tr>
<tr>
<td>Convergent-discriminant validity</td>
<td>When there is no gold standard or future event for validation, investigators determine whether the measures agree with (or are “convergent with”) other similar measures, and disagree with (or can be “discriminated from”) measures of states that theoretically should not be related to them</td>
</tr>
</tbody>
</table>

Statistical Analysis and Risk Adjustment

Techniques of statistical analysis are beyond the scope of this chapter,
but the univariate versus multivariate distinction merits mention here. Studies using only univariate analyses such as chi-square are limited in strength, whereas multivariate regression, necessary for risk adjustment, is more the gold standard. The risk adjustment process occurs as part of the main statistical analysis, generally multivariate regression analysis. In this method, multivariate regression accounts for the effects of risk variables of interest (the independent variables) on the outcome of interest (the dependent variable). There are many techniques of multivariate regression, a far superior method to another observed analytic approach, that is, multiple chi-square or analysis of variance (ANOVA) analyses, which are fraught with the potential to lead investigators to erroneous conclusions, primarily because with these approaches the variables are studied in isolation and not simultaneously. Simultaneous adjustment of potential confounding variables is critical for “apples to apples” comparisons. Risk adjustment is critical but not the sole component in enhancing a study’s validity. Other important study design considerations include sample size and statistical power; the distribution of data, normal or otherwise; and the effect of outliers, to name a few. Adjustment for risk is essential in comparing patient outcomes, because patient-specific risk factors can mask or confound the relationship between interventions or treatments and outcomes. Patient-specific characteristics and many aspects of patient health status, especially disease comorbidities, are causally related to the outcomes of care, as is well documented in the research literature. Therefore, risk adjustment—specifically, adjustment for disease comorbidity and severity of illness—is a way to remove the effects of confounding factors before making inferences about the outcomes of care. These adjustments are particularly relevant to surgical studies in which the effectiveness of different procedures or approaches to care is evaluated to guide evidence-based practice.

Risk defines the likelihood of a poor outcome, and the dimensions of risk are manifold. A broad set of patient risk factors can include age; sex; race and ethnicity; clinical stability; principal diagnosis; severity of principal diagnosis; extent and severity of comorbidities; physical functional status; psychological, cognitive, and psychosocial functioning; cultural and socioeconomic attributes and behaviors; health status and quality of life; and patient attitudes and preferences for outcomes.

Comorbidities, or coexisting diagnoses, are usually coded in medical records as the secondary diagnoses, diseases unrelated in etiology to the principal diagnosis. Often comorbidities are chronic conditions, such as diabetes mellitus, chronic obstructive pulmonary disease, or chronic ischemic heart disease. Patients with comorbidities often differ significantly from those without these conditions. In addition to having a
higher risk of death and complications, they are less able to tolerate
treatment and slower to respond to therapy. In the case of surgery,
operative risks often increase due to the presence of comorbidities.\textsuperscript{72,73}

Adjustment for severity of illness differs from that for comorbidities. The
definition of severity of illness is related to disease prognosis:
expectations about patients' clinical outcomes are evaluated according
to the extent and nature of diseases. For many diagnoses in which death is
not an immediate event, defining severity will involve a subjective
standard. Similarly, comparing severity among different diseases or
conditions is more of a challenge. However, differentiating by severity
among patients within a single diagnostic category is important for
describing the illness burden in general, and distinguishing patients by
the severity of their principal diagnoses is a necessary first step.
Important considerations beyond severity of the primary diagnosis include
the number and severity of comorbid diagnoses, acute physiologic
stability, functional status, and resource needs\textsuperscript{70} secondary to the illness.

Relating severity and comorbidities involves translating different stages of
clinical conditions into an overall risk score, which requires sophisticated
analysis of very large databases to obtain empirical evidence, as well as
clinical judgment and an understanding of the limits of empirical
analyses.\textsuperscript{74} To begin applying adjustments to surgical outcomes, one must
first determine which risk factors need to be accounted for in the study.
These could be patient-specific characteristics such as age, sex, race, and
medical conditions. In addition, when examining the relationship between
interventions and outcomes of care, it is important to adjust for risk
factors relating to the patient's condition, such as comorbidities, severity,
or other disease-specific conditions, or to the procedure. These factors
are generally determined by review of the literature and data analysis
examining the correlation between the outcome of interest and each risk
factor.

\textbf{Data Sources}

Data sources vary widely depending on the study design, from original
primary data collection to extraction of data from clinical databases set up
to study the condition of interest or from administrative databases such
as Medicare claims files. Types of data generally included for analysis
range from clinical signs and symptoms to laboratory results to measures
of morbidity and mortality to economic and patient-reported measures as
previously described.

Mortality, morbidity, and utilization are the most commonly used clinical
measures because they are the most accessible from medical records,
health departments, and hospital charts. Detailed clinical information can
be collected unobtrusively by retrospective review of medical records. To maximize reliability, trained reviewers with a clinical background must perform abstraction. Patient confidentiality must be assured, and institutional review board (IRB) approval must be secured prior to embarking on such a review.

Morbidity surveys on population samples, such as the National Health Survey and National Cancer Surveys, are helpful because they provide population-based descriptions of frequency of death and complications and can be used to monitor trends over time. Statistical power is obtained with relative ease, and data assembly and analysis are relatively inexpensive. Disease reporting—for communicable diseases and cancer registries—is helpful for the same reasons.

Claims data analysis uses data files, such as those maintained by the Medicare program or accumulated as a byproduct of insurance and prepaid medical care plans, to explore patterns of clinical outcomes at the population level. Several problems limit the value of claims data for assessing medical effectiveness or evaluating the quality of care, however. Because they are intended primarily for financial analysis, claims data may not contain enough detail about clinical features thought to affect prognosis, such as the stage of a disease's progress at the time of diagnosis. Chart audits should be performed to confirm the accuracy of coded information. The description of diagnoses and complications is often constrained by the ICD coding system, and clinical events out of hospital, in the ambulatory setting and at freestanding surgical centers, are frequently excluded from analysis. For confidentiality reasons, patient records tend not to be linked over time or across different settings. As a result, these are often cross-sectional, rather than longitudinal, analyses. Overall, it can be difficult, using claims data, to identify clinically relevant patient groups and to control for clinical factors likely to affect outcomes.

There are three major database sources for most medical and surgical effectiveness research: administrative databases, medical records, and patient-based surveys. Administrative databases, large claims files collected for billing purposes, are very useful for outcomes study of a descriptive nature, such as exploring variations in treatment patterns.

Medical records offer a rich source of information about patients and their care. Generally, medical charts document patients' histories, chief complaints, present symptoms, physical examination results, clinical assessments and diagnoses, diagnostic laboratory results, procedures, medications, in-hospital responses to therapy, clinical courses, and discharge plans. For studies relying on medical records, investigators need to have explicit review criteria; otherwise, interobserver variation
and subjectivity could bias the results. Risk adjustment methods that rely on clinical measures obtained from medical records such as vital signs or laboratory findings are able to measure risks not measurable using administrative data systems. However, the costs of primary data collection from medical records may be prohibitive.

Patient surveys can obtain information unavailable in either administrative files or medical records. Survey instruments can be designed to capture subjective information such as the perception of quality, satisfaction, personal preferences, and utility. Currently, there are many survey scales available to measure health behavior and psychosocial characteristics. However, surveying patients for outcomes studies may be expensive. It requires much effort to develop an appropriate instrument and to validate it. In addition to the cost of conducting a survey, there are potential biases and numerous logistical concerns relating to the process of collecting data from patients. Survey-based information should be tested for its reliability and validity.

The practice of evidence-based surgery requires the ability to critically assess and interpret the surgical literature to plan the optimal care of each patient, such as the selection of surgical versus medical treatment, the most appropriate procedure, for example, palliative versus curative, or open versus minimally invasive approaches. Beyond the care of the individual patient, evidence can influence the design of systems of care by identifying structure and process attributes of ideal systems. For example, system design includes minimum procedure volume requirements for facilities and providers, the necessity of dedicated resources such as monitoring capabilities or specialty trained surgeons and staff, or availability of critical pathways and care management protocols. Surgical evidence can greatly contribute to efforts to improve quality and patient safety. In studying outcomes of care, the structural and process attributes of surgeons and facilities can be analyzed to determine those factors that influence outcome to support quality improvement efforts, for example, specialty training, board certification, level of experience, the physical environment, teamwork, and communications. Last, knowledge of the significance of these factors may influence health policy, such as provider and facility licensure requirements, reimbursement for care based on efficacy, effectiveness and quality of care, and planning for manpower and technology requirements.

The most important changes likely to come about in evidence-based surgery include the proliferation of evidence in the literature and the availability of patient data in regional medical record databases. The proliferation of evidence will stem in large part from the efforts of the National Institutes of Health (NIH) to foster more clinical research and
quickly promulgate the results and the growing interest and specialty training in surgical outcomes research. If properly designed with sufficient patient and procedural details, regional medical records databases will enable numerous studies of relevance to surgical outcomes research. Knowledge and understanding of the methods and techniques of outcomes research will be increasingly important to all surgeons.

KEY POINTS

- The most commonly used clinical measures of the consequences of surgical therapy are rates of mortality and morbidity, gains in life expectancy, relative risk, relative and absolute risk reduction, and the potential impact in absolute numbers.
- Key parameters for economic analysis include the time period for analysis, the breadth of services provided, and the perspective from which costs are defined, i.e., the patient's, the provider's or the payer's.
- Quality-of-life assessment can have relevance to surgical research and practice for defining the indications for surgery, for monitoring the patient, and for evaluating the impact of treatment.
- There are two basic approaches to quality-of-life assessment: generic and disease specific.
- Randomized clinical trials are generally accepted as the definitive approach for assessing of a new treatment and represent level I evidence.
- Observational studies differ from randomized clinical trials in that the investigator merely observes the result of the exposure to the factor of interest, as the investigator does not assign the exposure.
- A meta-analysis requires a review of the literature for the outcomes from all relevant studies regarding a given surgical procedure and a specific outcome.
- Reliability refers to a measure's consistency or repeatability, that is, whether it gives the same result repeatedly when the same thing is measured.
- Validity refers to whether or not a measure reflects what it is intended to measure.
- Multivariate regression accounts for the effects of risk variables of interest (independent variables) on the outcome of interest (dependent variable) and is the “gold standard” of
statistical analysis for risk adjustment.

References


10. Davidoff AJ, Powe NR, The role of perspective in defining economic


21. Rubin HR, Wu AW. Patient satisfaction: its importance and how to


31. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993;270:2598.

32. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B.
What were the results and will they help me in caring for my patients?


34. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? *JAMA* 1994;271:703.


38. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? *JAMA* 1995;273:1292.

39. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? *JAMA* 1995;273:1610.


43. Naylor CD, Guyatt GH; for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. X. How to use an article reporting variations in the outcomes of health services. JAMA 1996;275:554.

44. Naylor CD, Guyatt GH; for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. XI. How to use an article about a clinical utilization review. JAMA 1996;275:1435.

45. Drummond MF, Richardson WS, O'Brien BJ, et al. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? JAMA 1997; 277:1552.


47. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature. XIV. How to decide on the applicability of clinical trial results to your patients. JAMA 1998;279:545.


57. Giacomini MK, Cook DJ. Users' guides to the medical literature. XXIII. Qualitative research in health care. A. Are the results of the study valid? *JAMA* 2000;284:357.


