World Journal of Surgery

© 2006 by the Société Internationale de Chirurgie
Published Online: 13 February 2006

World J Surg (2006) 30: 255–266
DOI: 10.1007/s00268-005-0156-0

SURGICAL OUTCOMES

Surgical Outcomes Research Based on Administrative Data: Inferior or Complementary to Prospective Randomized Clinical Trials?

Ulrich Guller, MD, MHS

Department of Surgery, Divisions of General Surgery and Surgical Research, University Hospital Basel, CH-4031 Basel, Switzerland

Abstract

The importance of surgical research has gained new prominence over the past decades as the relevance of well designed and well conducted studies has become increasingly evident. There are two basic but diametrically different methods of conducting research: the prospective randomized clinical trial and the retrospective surgical outcomes study based on administrative data. Administrative databases contain data that were initially collected for purposes other than scientific research. Whereas the prospective randomized clinical trial is familiar to most surgeons, surgical outcomes research based on administrative data constitutes a genre of investigation that is often unfamiliar to and even disparaged by the surgical community. In the present article, the strengths and weaknesses of both prospective randomized clinical trials and retrospective surgical outcomes research are discussed. Specifically, the advantages and limitations of investigations based on large administrative databases are outlined. Because both study designs play an important role in surgical research, carefully designed and implemented surgical outcomes research based on administrative data should be viewed as being complementary and not inferior to prospective randomized clinical trials.

The subject here—outcomes research based on administrative databases—begs the question: goldmine or fool's gold? Although "outcomes research" has become increasingly visible in the surgical literature over the past few years, a clear definition is still lacking, and descriptions of "outcomes research" are numerous[1] and often confusing. Outcomes research includes a variety of study types, including traditional clinical research (prospective randomized clinical trials, cohort studies, case-control studies, case-series)[1] as well as volume-outcomes research, small-area analyses, trends analyses, access to health care investigations, cost-effectiveness studies, and quality of life research.[2] The overall objective of surgical outcomes research is to assess the effectiveness, appropriateness, and costs of surgical care.[1,3]

Outcomes research based on secondary or administrative data represents a specific subset of clinical research. Administrative data have been defined as "large, computerized data files generally compiled in billing for health care services such as hospitalizations."[4] Therefore, administrative data contain information that were primarily collected for purposes other than scientific research (e.g., billing).[5,6] Herein I use the terminology "secondary database" as a synonym for "administrative" database.

Surgical outcomes research based on administrative data has a number of important advantages compared with randomized clinical trials. First, it is generally less

Correspondence to: Ulrich Guller, MD, MHS, Department of Surgery, Divisions of General Surgery and Surgical Research, University Hospital Basel, CH-4031 Basel, Switzerland, e-mail: uguller@yahoo.com

costly and time-consuming, as the data are readily available.[5–8] Second, exclusion criteria are usually chosen parsimoniously, and thus the generalizability of the findings of outcomes research studies may exceed that of tightly controlled randomized trials.[9] Hence, the issue of selection bias—owing to the population-based collection of administrative data—is less problematic,[2,9] and the effectiveness, the actual benefit in the real world, of an intervention can be assessed.[10] This is in contrast to prospective randomized clinical trials that evaluate a procedure's efficacy in highly selected populations under ideal and somewhat artificial circumstances. Third, as administrative databases contain often thousands or even millions of patients, lack of power does not represent a threat to the analyses. Even the evaluations of important outcomes in subsets (e.g., elderly patients, women, children, patients with a specific tumor stage) or assessments of rare diseases or infrequent endpoints can usually be done with sufficient statistical power.[9] Fourth, administrative databases allow the performance of descriptive analyses.[10] For instance, such databases enable the approximate determination of patients with a certain disease as well as the age, gender, and race distribution of these patients. Administrative databases also allow comparison of mortality, morbidity, or reoperation rates among hospitals and regions.[8] Finally, the most important advantage of outcomes research is that it enables researchers to answer relevant questions that cannot be answered through a randomized clinical trial because the latter would require prohibitively complex, costly, or even ethically unacceptable practices.[1]

Herein I briefly describe some of the most important subtypes of surgical outcomes research, including small-area variations, volume-outcomes research, and access to health care investigations. As shown below, administrative data can be well suited to perform such investigations.

SMALL-AREA VARIATIONS OF SURGICAL PROCEDURES

The primary objective of small-area analyses is to assess differences in the use of surgical procedures among various geographic regions. If relevant differences are found, it is likely that certain areas are underserved whereas others may be experiencing overutilization of surgical procedures.

More than 30 years ago, Wennberg and colleagues performed pioneering work in small-area analyses. Their ground-breaking study assessed whether differences existed in the use of a variety of surgical procedures in different regions of Vermont.[11] They found striking discrepancies in the age-adjusted rates for nine frequently performed surgical procedures. The most important difference was observed for tonsillectomy, which ranged from 13 cases to 151 cases per 10,000 persons. Building on Wennberg's work, numerous investigations have shown large discrepancies in the use of surgical procedures for breast cancer,[12,13] back pain,[14] colorectal cancer,[15] knee arthroplasty,[16] tonsillectomy, hemorrhoidectomy, hysterectomy, and prostatectomy during the past few decades.[17]

Administrative data are useful in the performance of small-area variation studies. In one investigation, Nattinger and associates[18] used Medicare administrative data to assess variations in the use of breast-conserving surgery among 36,982 women with breast cancer in various states of the United States. The authors found considerable differences (ranging from 3.5% to 21.2%) in the use of breast-conserving therapy among different U.S. states. Nattinger *et al*. concluded that this variation in the use of breast-conserving treatment could not be explained by differences in hospital characteristics.

Area variations analyses of surgical procedures have significant potential in surgical research, as they allow the identification of large differences in surgical practice. These differences are partially attributable to a lack of consensus among surgeons and to uncertainty regarding the effectiveness or appropriateness of a given procedure.[11,17,19] These analyses have proven useful in stimulating surgeons to reflect critically on the reasons for the existing differences and may help establishing a consensus regarding the indication for a surgical intervention. This not only may result in decreased health care costs but, more importantly, may lead to tremendous patient benefit.

For instance, investigations have shown that a between-area variation in the use of breast-conserving therapy has stimulated a decrease of the rate of mastectomy performances in breast cancer patients, although differences persist.[20] Indeed, the diffusion of certain guidelines appears to be faster in teaching hospitals, large hospitals, and urban areas.[20,21]

VOLUME OUTCOMES ANALYSES

Volume outcomes analyses assess whether surgeons or hospitals with high case loads have better outcomes than do low-volume providers. Although this hypothesis seems intuitive, the association between higher volumes and better outcomes for certain procedures has been a matter of great debate for many years.

One of the most important and most widely referenced volume outcomes investigations was performed by Birkmeyer and colleagues.[22] Based on Medicare data that includes approximately 2.5 million procedures, they assessed whether high hospital volume was associated with decreased mortality for six cardiovascular procedures and eight types of cancer resection. Differences in sociodemographics and risk factors between the hospital volume categories were adjusted for in multivariable analyses. The authors found diminished mortality with increasing hospital volume for all 14 surgical procedures. Based on these findings, Birkmeyer *et al*. concluded that patients undergoing cardiovascular or cancer procedures can significantly decrease their mortality by selecting a high-volume hospital. A variety of other analyses based on secondary data have shown that higher volume was inversely related to lower mortality for pancreatectomy, colorectal surgery, esophagectomy, liver resection, prostatectomy, lung or bronchial tumor resection, and pelvic exenteration.[23–29]

In a recent investigation based on the Nationwide Inpatient Sample 1997,[30] we assessed whether patients with rectal cancer are more likely to undergo sphincter-sparing procedures versus abdominoperineal resection if operated on by high-volume versus low-volume surgeons. We found a risk-adjusted odds ratio that exceeded 5 (patients operated on by high-volume surgeons were more than five times more likely to undergo sphincter-sparing procedures than were patients who were operated on by low-volume surgeons). As we were unable to risk-adjust for tumor size, tumor stage, and grading because these parameters could not be ascertained from the Nationwide Inpatient Sample, patient selection could in part explain the differences in performing sphincter-sparing procedures. Nonetheless, we believe that the important difference (risk-adjusted odds ratio exceeding 5) of undergoing sphincter-sparing procedures between high- and low-volume surgeons cannot be explained solely by residual confounding. However, because investigations based on administrative data are best suited for generating hypotheses, a similar investigation with "real" data that contain potential confounders would be warranted. This would help confirm the hypothesis that patients may decrease the risk of undergoing abdominoperineal resection associated with definitive colostomy if they are operated on by high-volume surgeons.

It is noteworthy that the relation between higher volume and improved outcomes should not be assumed automatically. Although this relation is now generally accepted for high-risk procedures, it remains unclear whether it applies equally to low-risk surgery. High volume represents a surrogate marker for outcomes and thus cannot necessarily be taken as an indicator of quality. Furthermore, although regionalization is justified for high risk procedures, it is obvious that, for logistical reasons, not all procedures can be performed in highly specialized centers.

ACCESS TO HEALTH CARE

Equality in health care utilization has become an increasingly important issue for health services research over the past decade. Racial[31–33] and socioeconomic[32] differences have been identified as independent factors for inequality of access to health care. In cancer patients, several investigations reported that African American patients receive less intensive treatment or have poorer outcomes for breast,[34] prostate,[35] colon and rectum,[36,37] bladder, and lung[32,38] cancer. It is clear that well designed studies that reveal potential socioeconomic or racial discrepancies in access to health care are of greatest importance to the medical community, policymakers, and the general public.

For obvious reasons, however, access to health care cannot be assessed in a randomized clinical trial. We then see that surgical outcomes research complements the randomized clinical trial in this regard, as outlined in two examples.

First, Cooper and colleagues identified predictors associated with patients undergoing potentially curable surgical therapy for resectable colorectal cancer, basing their research on a large administrative database that included 81,579 Medicare beneficiaries.[37] The authors found that African Americans were significantly less likely to undergo potentially curative surgery than were white patients (78% vs. 68%, $P < 0.001$). This difference remained statistically significant even after controlling for age, co-morbidity, and location and extent of the tumor. Also, African Americans had a significantly higher mortality rate than white patients, even in multivariable and subset (teaching versus nonteaching, private versus public) analyses.

Second, in a recent study, our group[39] investigated whether private insurance status and race represented independent predictors for undergoing laparoscopic appendectomy in patients with appendicitis. Patients (n = 145,456) with primary ICD-9 procedure codes for laparoscopic and open appendectomy were selected from the 1998, 1999, and 2000 Nationwide Inpatient

Samples. Even after adjusting for potential confounders such as age, gender, the patient's co-morbidity and median zip code income, hospital location and teaching status, and presence of appendiceal abscess or perforation, privately insured patients and white patients were significantly more likely to undergo laparoscopic surgery than were African Americans and Medicaid patients.

TREND ANALYSES

Administrative data may be well suited for trend analyses. Often, nationwide or statewide administrative data have been collected for several years or even decades, allowing evaluation of a change over time.

A good example of the use of administrative data for trend analyses was provided by Flum and colleagues.[40] They used the Washington state hospital discharge database and the U.S. Census Bureau data for 1987–1998 to evaluate whether misdiagnosis of appendicitis has declined with increasing use of diagnostic tools such as computed tomography (CT) scans and ultrasonography. The analysis included 63,707 nonincidental appendectomy patients. Among them, 84.5% had appendicitis and 15.5% had no diagnosis of appendicitis. Interestingly, the percentage of misdiagnosed appendicitis did not change over time, implying that the correct diagnosis of appendicitis has not significantly improved with more frequent use of diagnostic imaging techniques.

COMPARISONS OF SURGICAL PROCEDURES BASED ON ADMINISTRATIVE DATA

Whereas surgical outcome studies usually assess questions regarding the distribution and effects of health care provided to average persons in typical clinical practice, randomized clinical trials measure the relative efficacy of a treatment in highly selected patient samples under ideal and somewhat artificial circumstances. Therefore, the objectives of the randomized clinical trial and surgical outcomes research based on administrative data are often different, and there is usually not a choice of whether to use one design or the other. Rather, the research question determines which study type should be used. However, an overlapping area between randomized clinical trials and surgical outcomes research is the comparison of two surgical procedures. In an investigation from our group, laparoscopic versus open appendectomy were compared using the Nationwide Inpatient Sample 1997, an administrative database with patient discharges from various states across the United States.[41] A total of 43,757 patients were included in the investigation; and outcomes such as in-hospital morbidity, in-hospital mortality, length of hospital stay, and the rate of routine patient discharge were assessed. We found the laparoscopic procedure to be advantageous over the open procedure for these outcomes.

Interestingly, Benson and Hartz compared infection rates after laparoscopic and open appendectomy between published prospective randomized clinical trials and observational studies. The authors found similar results for the two study types.[42]

CAVEATS OF USING ADMINISTRATIVE DATABASES

Limited Clinical Data Availability

It is clear that administrative databases have several inherent limitations and drawbacks. Administrative databases are usually established to serve billing purposes but not to answer specific research questions. Therefore, the amount of clinically relevant data in administrative databases may be limited.[7,8,43,44] For instance, information regarding disease severity, tumor size, lymph node status, and grading may be missing. It is thus critically important to consider whether between-group differences of those parameters affect the study findings and conclusions. For instance, in the above-mentioned volume-outcomes study,[22] a variety of putative confounding factors were not adjusted for in the multivariable analysis, as they could not be ascertained from the administrative database used. Nonetheless, Birkmeyer and colleagues[22] concluded that the mortality differences observed between low- and high-volume providers could not be explained by unmeasured confounding alone.

Similar to putative confounders, administrative databases do not contain certain important endpoints, such as postoperative quality of life and functional status.[8] Nonetheless, length of hospital stay, postoperative morbidity, postoperative mortality, and rate of reoperation can be ascertained from many administrative databases. These are relevant outcomes that allow important research questions to be addressed and have the potential to affect surgical practice.

Miscoding, Undercoding, and DRG Creep

The data of administrative databases may be inaccurate, first, because of miscoding or undercoding.[7,15,43,45–50] There are various reasons that may lead to miscoding, as the medical record is nowadays increasingly used for cost containment, legal, or administrative purposes.[6] Thus the medical record may in fact not reflect the actual health status of the patient.

Second, overcoding has been described as a potential source of distortion of administrative data.[51] For instance, if hospitals are reimbursed based on the complexity of the patient's disease, there may be a tendency for overcoding primary and secondary diagnoses, a phenomenon called diagnosis-related group (DRG) creep.[51] Miscoding represents an inherent limitation that must be carefully considered when interpreting the findings of studies based on administrative data.

Data Mining

Research based on administrative data should be hypothesis-driven. It is critically important that, similar to a randomized clinical trial, an a priori hypothesis is stated. Then, one should ascertain whether an administrative database is well suited to test this a priori hypothesis. When interpreting an investigation based on administrative data, it is essential to make the distinction between hypotheses that were created prior to performing the study (a priori hypothesis) and hypotheses that were stated after the study was conducted (a posteriori hypothesis). A priori hypotheses do not carry the risk that the investigator was influenced by the readily available data and thus are less prone to generating erroneous conclusions. If hypotheses are stated a posteriori, it is possible that the investigator looked at various patient subsets until he or she found significant results. This phenomenon is often referred to as "data mining," "data dredging," or a "fishing expedition"; and it has an inherent increased risk of a type I error (obtaining a false-positive finding).[52] Investigations that formulate their hypotheses after the study has been conducted should be viewed as hypothesis-generating rather than hypothesis-testing, and even more so if they examine patient subsets and perform multiple comparisons.[52,53] If an investigator performs multiple comparisons, the threshold of statistical significance (usually set at 0.05) should be adjusted for using the Bonferroni or other statistical method to decrease the risk of a type I error.[52] For instance, if three independent hypotheses are tested, the threshold of statistical significance should be lowered to 0.05:3.00 (= 0.017).

Statistical Significance versus Clinical Relevance

As administrative databases can contain up to several million patients, with even extracted patient samples potentially being large, it is essential to differentiate between statistical significance and clinical relevance. If the sample size is large, even tiny differences between study groups become statistically significant.[52] The question, however, is whether these small differences are clinically relevant. A clinically relevant difference is associated with a change in health care that represents a meaningful improvement to the patient. It is thus critically important to consider the absolute results of an analysis based on large administrative databases, as they may be clinically irrelevant despite being statistically significant.

Confounding

Prior to defining a confounding variable it is important to understand the meaning of, and the association between, a predictor variable and an outcome. Commonly, studies are designed to show a link between a predictor variable (independent variable) and an outcome (dependent variable). Predictor variables can be either a diagnostic or therapeutic intervention (e.g., new surgical therapy, new diagnostic procedure) or a risk/prognostic factor such as age, patient co-morbidities, tumor size, or lymph node status.[52] Frequently assessed outcomes in the surgical literature are disease-free survival, overall survival, response to a treatment, and postoperative morbidity. A confounding variable (also known as a confounding factor or confounder) is an extrinsic factor that is linked to the predictor variable and also affects the outcome. The perceived association between the predictor and the outcome variable is distorted because of the confounder.[52] Also a confounder cannot be an intermediate in the causal pathway between exposure and outcome.[54] Because of the nonrandomized study design of retrospective outcomes research, the results must be adjusted for potential confounding factors using multivariable analyses (or other statistical techniques, such as propensity score analyses or an instrumental variable method) to minimize bias.[55] It is clear that bias cannot be perfectly adjusted for, as some known confounders may not be in the database. Moreover, although it is possible to risk-adjust for known confounders if available in the database, researchers cannot control for unknown confounding. Nonetheless, as pointed out by Birkmeyer and colleagues,[22] if the differences are large even after adjusting for putative confounding factors, it can be assumed that they cannot be explained solely by residual or hidden confounding.

Because of these inherent limitations and drawbacks, it is important to interpret and scrutinize critically the surgical outcomes research based on administrative data prior to incorporating the studies‘ recommendations into clinical practice. Also, efforts must be undertaken to improve the accuracy of administrative databases even further, which makes them an even more valuable tool for assessing outcomes and quality of care. However, most of the above-mentioned investigations have been performed with greatest attention to scientific rigor and are of clear relevance to the medical community despite being based on administrative data. It must be concluded that—if well designed and well conducted—administrative databases are goldmines for surgical research rather than fool's gold.

CHALLENGES WHEN PERFORMING CLINICAL TRIALS IN SURGERY

It is generally agreed that randomized clinical trials, when designed and conducted properly, provide the highest standards of scientific evidence and are considered the gold standard for evaluating the efficacy of therapies.[56–59] By randomly assigning patients to either the experimental arm or the control arm, the investigator can control for extraneous factors (confounders). In contrast to nonrandomized studies, random allocation allows controlling for both known and unknown confounders. Theoretically, the only difference between the two groups is the intervention (A versus B). Thus, the investigator is better able to demonstrate the causal link between the intervention and the endpoints under investigation.[60] Despite the obvious advantages and strengths of randomized clinical trials in surgery, they are complex, costly, and time-consuming undertakings.

Many surgeons believe that every prospective randomized study is bias-free. This belief is not in fact reflected in reality, as poorly designed and conducted randomized trials provide distorted, confounded results that are not useful for improving current surgical practice. I have herein summarized the particular challenges that pertain to the performance of prospective randomized clinical trials in surgery.

Patient Accrual and Clinical Equipoise

Unless a randomized clinical trial is performed for common diseases, recruiting a sufficient number of patients in a timely manner is often difficult. For instance, performing a single-institution trial for adrenal or rare thyroid cancer might not be feasible, as the number of patients with the disease under investigation is prohibitively small. Equally important, patients often do not want to be randomized.[61] They may be reluctant to have random chance decide into which arm they go. Similarly, surgeons often do not want to randomize their patients, as they frequently believe that one therapeutic option is better than the other. This phenomenon is referred to as lack of clinical equipoise. Equipoise represents a state of uncertainty regarding the benefits of alternative treatments.[62] The lack of clinical equipoise is prevalent in the surgical community and represents a challenging factor when performing surgical clinical trials. Thus it is critically important that the investigators emphasize to patients and physicians that the premise of a randomized clinical trial is based on the absence of current scientific evidence that the experimental arm is superior to the control arm. Furthermore, to facilitate patient accrual and increase the feasibility of the study, one should consider the option of performing a multicenter trial for rare diseases.

Selection Bias, Generalizability of Results, and "Pragmatic Trials"

Most randomized controlled trials have clearly defined inclusion and exclusion criteria, are based on a relatively homogeneous patient population,[63] and are performed under somewhat artificial and controlled conditions.[1,64] Moreover, only a small percentage of potentially eligible patients agree to participate in surgical randomized clinical trials.[64] This phenomenon is accentuated in randomized controlled trials in surgical oncology, for which it is estimated that less than 3% of cancer patients participate.[65] However, it is well known and extensively documented in the literature that patients who agree to participate in randomized clinical trials are systematically different from patients who do not participate.[58,64] Patients in clinical trials are, on average, healthier, more compliant, and enjoy higher socioeconomic status,[58,64] resulting in a selection bias of unknown magnitude. Thus even if an intervention works in the somewhat artificial setting of a randomized clinical trial, it is unclear whether it will have the same benefit in the "real world,"[1,2,58,63] as numerous examples in the medical literature have demonstrated. Also, it is clear that the findings of a randomized clinical trial cannot be extrapolated to patient populations that were excluded from the study. For instance, if a surgical intervention was shown to have significant overall survival advantage in male Caucasians with stage I/II disease, aged 40 to 55 years, the results cannot be generalized to women, patients with advanced stage disease, African Americans, Hispanics, Asians, or the elderly. Therefore exclusion criteria must be discussed carefully during the planning phase of a trial and should be chosen parsimoniously.[64,66] Clearly, the more stringent the exclusion criteria, the greater is the selection bias and the less generalizable are the results.

The performance of "pragmatic trials" may help diminish selection bias. Pragmatic trials aim to reflect the real-world situation as much as possible and often evaluate a range of outcomes, including cost-effectiveness and quality of life aspects in addition to the clinical endpoints.[67] Exclusion criteria are chosen parsimoniously in the design of pragmatic trials, and patients are always analyzed in the initially assigned treatment group (intention-to-treat analysis).[68] The strength of pragmatic trials lies in providing patients and health care providers with information regarding the effectiveness of treatment options in routine clinical practice.[68,69]

Sample Size, Follow-up, and Costs

One of the most unambiguous and frequently used primary endpoints in surgical clinical trials is overall survival, as it is a wholly objective criterion. However, because the surgical treatment is often studied in early-stage disease where surgical therapy is most beneficial, the evaluation of overall survival generally requires a long follow-up[70] period, which is associated with increased logistical difficulties and higher costs. Furthermore, the longer the follow-up, the higher is the drop-out rate, which again increases the number of patients required in a prospective randomized trial. To shorten the follow-up period, the assessment of surrogate endpoints (i.e., endpoints believed to be linked to clinical endpoints, such as overall survival) has been suggested. The advantage of using surrogate endpoints is that they can be evaluated at an earlier point in time than the clinical endpoint, shortening the time required for the trial.[70–72] However, numerous investigations have shown that surrogate endpoints (e.g., tumor growth or increase in a tumor marker) are fallacious.[70,73]

Lack of Power, Type II Error, and Effect Size

Power is defined as the probability of finding a statistically significant result (of rejecting the null hypothesis) in a study if the populations are truly different.[74,75] A type II error (synonym: beta) represents the situation in which the results lead to the erroneous conclusion that there is no significant difference between the study groups when in reality a difference exists.[52,74] Beta, the false-negative rate, is complementary to the power of a study.

The choice of adequate power in a randomized clinical trial is critical, as investigators and funding agencies must be confident that an existing difference in the overall patient population can be detected using the study sample. The power of a study depends on various factors: the effect size (expected difference in the primary outcome between the study groups, see below), the chosen type I error (rate of false-positive results), and the precision (e.g., standard deviation) of the primary outcome under investigation.[74] Moreover, the power of a study is intrinsically linked to the sample size. The larger the sample size, the higher is the power. The importance of sample size consideration is clear: Even the most thoroughly planned and well executed randomized clinical trial may fail to answer the research question if the sample size is too small. Often small studies do not find statistically significant differences. It is then unclear whether there was truly no difference between the treatment options or the sample size was prohibitively small to provide sufficient evidence for a statistically significant difference.[74] Unfortunately, there is a plethora of randomized clinical trials in the surgical literature that were clearly underpowered while claiming that there was no statistically significant difference in outcomes,[76–78] an erroneous and potentially harmful conclusion.

For sample size computations, investigators start by defining a clinically meaningful difference in the primary outcome (e.g., overall survival) between treatment A and B (called effect size, or delta), which is believed to be true for the overall patient population. The effect size is often the "least important difference in outcomes" that would lead to a change in current clinical practice.[79] The smaller the expected difference in the outcome, the larger the required sample size must be.[74]

As a result of the difficulties of patient accrual, the high costs associated with prospective randomized trials in surgery, limited funding, or undertraining, the estimates on which the sample size for a randomized clinical trial is based might be too optimistic (e.g., choosing a too large effect size), and thus the resulting sample size is too small or, worse, no sample size was computed at all. Moreover, even if the sample size of a randomized clinical trial provides sufficient power to assess the primary endpoint, it is still too small to perform relevant subset analyses. For instance, it may be important to know whether an intervention has particular benefits in the elderly, in women, or for a specific disease stage.

Another challenging issue in the interpretation of randomized clinical trials that enrolled patients with early-stage disease with good prognosis is the estimation of the real treatment effect if there are few events. For instance, one might claim that a surgical procedure is safe because there were no deaths among 20 patients undergoing surgery.[80] This may, however, not be true and is difficult to gauge for the reader if 95% confidence intervals are not provided in the manuscript.

A simple aid in the interpretation of such results is "the rule of 3" for zero numerators: If an outcome (e.g., death) occurs 0 times in n patients, the upper 95% confidence limit is approximately 3/n.[80,81] In the example above, the upper 95% confidence limit would thus be about 3/20 = 0.15, or 15%. In other words, based on the sample of 20 patients, one can be 95% sure that the true mortality rate for the surgical procedure lies between 0% and 15%.

Lack of Placebo Controls and Sham Surgery

An important difficulty in the design of surgical clinical trials is the frequent lack of placebo controls (surgical placebos, sham surgery). A surgical placebo represents a simulated operation in which the skin incisions are done without actually performing the operation. This makes the blinded patient believe that he or she underwent surgery which may be associated with a placebo effect. For instance, Moseley *et al*. conducted a three-arm prospective, randomized, placebo-controlled trial in 180 patients with osteoarthritis of the knee.[82] Patients were assigned to arthroscopic débridement, arthroscopic lavage, or placebo surgery. The placebo surgery consisted of performing a skin incision only, without inserting the arthroscope. Interestingly, outcomes after arthroscopic lavage or arthroscopic débridement were not superior to those seen with the sham surgery.

It is clear that the placebo-controlled randomized clinical trial represents the most unbiased study design. However, the controversy regarding sham surgery is considerable, and therefore surgical placebos are rarely used.[79,83,84]

Postrandomization Bias

One of the most important challenges when performing clinical trials in surgery is postrandomization bias. Postrandomization bias is largely due to the impracticalities of blinding during a surgical intervention. Single blinding (blinding the patient to the arm assigned) is rarely possible,[61] and double blinding (blinding both physician and patient to the assigned arm) is even more difficult if the intervention is a surgical procedure, a chemotherapy regimen, or radiation therapy.

Postrandomization bias occurs in a multitude of forms, one of which is ascertainment bias.[66] Let us consider a Phase III trial in which disease-free survival is evaluated in esophageal cancer patients randomized to neoadjuvant radiotherapy and surgery (arm 1) versus surgery alone (arm 2). It can be hypothesized that patients assigned to arm 2 (surgery alone) believe that they should undergo more stringent follow-up diagnostic procedures as they did not receive radiation therapy. It can thus be assumed that these patients see their primary care physicians more frequently, undergo more CT scanning, upper endoscopies, and tumor marker assays among other measures and that recurrences are diagnosed earlier in this subset of patients than in the subjects who were randomized to arm 1. In this scenario a between-arm difference of disease-free survival could therefore be linked to discrepancies in follow-up diagnostic procedures even if the therapeutic options are equiefficient.

Similarly, patients who were randomized to surgery alone might seek additional postoperative therapy (e.g., immunotherapy or alternative medical treatment options), which again may affect the outcome under investigation. This phenomenon is called co-intervention.[66]

Differences in surgical expertise may also affect the outcomes under investigation.[61] Let us consider a Phase III randomized trial comparing open and laparoscopic sigmoid resection for diverticular disease. It is possible that patients who are randomized to the laparoscopic procedure are operated on by the senior surgeon, who has extensive experience in laparoscopic surgery and is highly motivated to prove that the laparoscopic approach is superior to the open procedure. Conversely, the patient randomized to open colectomy is operated on by the surgical resident, who certainly has less experience and may lack the senior surgeon's particular motivation. Although randomization equally distributes both known and unknown confounders (e.g., age, co-morbidity, gender, race) to arm 1 and arm 2, there is a postrandomization bias of unknown magnitude due to differences in surgical expertise.

Another form of postrandomization bias is the "differential expertise bias."[85] Differential expertise bias occurs when unequal percentages of surgeons are experienced in performing the standard versus the investigational procedure.

Let us consider a different scenario of the trial randomizing patients to open resection (arm 1) versus laparoscopic sigmoid resection (arm 2). Currently, most surgeons are better trained to perform the open procedure, and a considerable percentage has either not yet started to do laparoscopic sigmoid resection or is still climbing the learning curve. Let us assume that 90% of the patients in both study arms are operated on by surgeons with excellent expertise in performing the open resection but little experience in the laparoscopic procedure, and only 10% of patients are operated on by surgeons who are equally trained for both procedures. In this scenario, the trial is biased toward the open procedure.

A multicenter randomized trial from The Netherlands has shown that laparoscopic fundoplication produces substantially worse results than the open procedure.[86] This conclusion is somewhat counterintuitive, and it is possible that the study was confounded by differential expertise bias.

To minimize postrandomization bias in surgical clinical trials, it is critically important to standardize the surgical procedures and the diagnostic follow-up interventions as much as possible. Equally important, all surgeons participating in a clinical trial should have similar experience and technical expertise. This can be guaranteed if each participating surgeon is required to have performed a certain number of cases (ensuring that the surgeons overcome their learning curve) involving the relevant procedure prior to participating in a randomized clinical trial.

Alternatively, one could perform a surgical expertise randomized controlled trial,[85] in which patients in study arm 1 are operated on only by surgeons with expertise for the open sigmoid resection whereas patients in arm 2 are operated on only by surgeons with expertise for the laparoscopic sigmoid resection. Although this approach reduces the differences in surgical skills among surgeons and thus postrandomization bias, it does not reflect the "real world" in which most surgeons perform both the laparoscopic and open procedures and choose one or another option depending on the patient.

Surgeons

Finally, a serious challenge to the performance of clinical trials in surgery are the surgeons themselves. First, surgeons often do not have sufficient time to invest in the thorough design and performance of randomized controlled studies, which may lead to poorly designed and conducted trials that are wasteful and ethically questionable. Second, surgeons are often not reimbursed, or only partially reimbursed, for performing additional therapeutic or diagnostic interventions, which renders participation to a clinical trial less appealing. While financial support is often readily available for pharmaceutical trials because they are frequently funded by the pharmaceutical industry, there are fewer industry sponsors of surgical research. Third, a problem with randomized clinical trials in surgery relates to the competitive culture in which surgeons work. Many surgeons may not agree to enroll patients in trials where the patients are assigned to a nonoperative study arm because of competition among surgeons to attract patients and out of fear of losing a source of referrals.

Finally, it takes many years until results from a prospective randomized clinical trial are available and an article can be published. This delay in publishable data is another factor that decreases the enthusiasm of some surgeons to participate in surgical trials.

The clinical trial in surgery is a challenging undertaking and if not carefully done has many pitfalls and limitations. Clearly, randomized clinical trials in surgery are not bias-free despite persistent perceptions to the contrary. Nonetheless, it is vitally important that surgeons continue to perform clinical trials, which because of their rigorous study design often represent cornerstones in surgical research. It is critical, however, that medical centers collaborate to accrue sufficient numbers of patients in a timely manner, that exclusion criteria are chosen sparingly to increase the generalizability of the results, that postrandomization bias is minimized by standardizing surgical and diagnostic procedures, and that surgeons collaborate with clinical researchers and statisticians. Only then can results with the highest scientific value and greatest potential patient benefit be obtained.

CONCLUSIONS

Prospective clinical trials and retrospective outcomes research have their respective strengths and limitations, and both deserve a place in surgical research. Although well designed and well conducted randomized clinical trials provide the "gold standard" of scientific evidence, especially if performed in the multicenter setting, there exists a plethora of novel, relevant, and interesting research questions that cannot be addressed through randomized clinical trials because of prohibitively high costs, long follow-up, the rarity of a specific disease, or because the study would require ethically dubious practices. This gap, however, can be filled by surgical outcomes research. The use of administrative databases, if carefully planned, thoroughly performed, and cautiously interpreted, can provide invaluable data for a variety of research applications. Therefore, outcomes research based on administrative databases should be viewed as complementary and not inferior to prospective randomized clinical trials in surgery. Both study types play important roles in the critical evaluation of health care delivery and must be further explored for potential benefit to current surgical practice. It is hoped that the present article stimulates surgeons to engage more actively in surgical research using prospective randomized clinical trials as well as retrospective outcomes research. Only the active exploration of both investigational avenues can maximize the result for which we all strive: improved health care delivery.

ACKNOWLEDGMENTS

I thank Mr. Jonathan McCall for carefully reading the manuscript and making many valuable suggestions.

REFERENCES

1. Porter GA, Skibber JM. Outcomes research in surgical oncology. Ann Surg Oncol 2000;7:367–75.
2. Brenneman FD, Wright JG, Kennedy ED, *et al.* Outcomes research in surgery. World J Surg 1999;23:1220–1223.
3. Epstein AM. The outcomes movement—will it get us where we want to go? N Engl J Med 1990;323:266–270.
4. Iezzoni LI (1997) Risk Adjustment for Measuring Health Care Outcomes Health Administrative Press, Foundation of the American College of Excecutives, Chicago.
5. Best AE. Secondary data bases and their use in outcomes research: a review of the area resource file and the Healthcare Cost and Utilization Project. J Med Syst 1999;23:175–181.
6. Iezzoni LI. Using risk-adjusted outcomes to assess clinical practice: an overview of issues pertaining to risk adjustment. Ann Thorac Surg 1994;58:1822–1826.
7. Lewis NJ, Patwell JT, Briesacher BA. The role of insurance claims databases in drug therapy outcomes research. Pharmacoeconomics 1993;4:323–330.
8. Wennberg JE, Roos N, Sola L, *et al.* Use of claims data systems to evaluate health care outcomes: mortality and reoperation following prostatectomy. JAMA 1987;257: 933–936.
9. Coleman AL, Morgenstern H. Use of insurance claims databases to evaluate the outcomes of ophthalmic surgery. Surv Ophthalmol 1997;42:271–278.
10. Armstrong EP, Manuchehri F. Ambulatory care databases for managed care organizations. Am J Health Syst Pharm 1997;54:1973–2005.
11. Wennberg J, Gittelsohn. Small area variations in health care delivery. Science 1973;182:1102–1108.
12. Farrow DC, Hunt WC, Samet JM. Geographic variation in the treatment of localized breast cancer. N Engl J Med 1992;326:1097–1101.
13. Sainsbury R, Rider L, Smith A, *et al.* Does it matter where you live? Treatment variation for breast cancer in Yorkshire; The Yorkshire Breast Cancer Group. Br J Cancer 1995;71:1275–1278.
14. Nilasena DS, Vaughn RJ, Mori M, *et al.* Surgical trends in the treatment of diseases of the lumbar spine in Utah's Medicare population, 1984 to 1990. Med Care 1995;33:585–597.
15. Cooper GS, Yuan Z, Stange KC, *et al.* Use of Medicare claims data to measure county-level variations in the incidence of colorectal carcinoma. Cancer 1998;83: 673–678.
16. Skinner J, Weinstein JN, Sporer SM, *et al.* Racial, ethnic, and geographic disparities in rates of knee arthroplasty among Medicare patients. N Engl J Med 2003;349: 1350–1359.
17. McPherson K, Wennberg JE, Hovind OB, *et al.* Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway. N Engl J Med 1982;307:1310–1314.
18. Nattinger AB, Gottlieb MS, Veum J, *et al.* Geographic variation in the use of breast-conserving treatment for breast cancer. N Engl J Med 1992;326:1102–1107.
19. Wright JG, Hawker GA, Bombardier C, *et al.* Physician enthusiasm as an explanation for area variation in the utilization of knee replacement surgery. Med Care 1999;37:946–956.
20. Gilligan MA, Kneusel RT, Hoffmann RG, *et al.* Persistent differences in sociodemographic determinants of breast conserving treatment despite overall increased adoption. Med Care 2002;40:181–189.
21. Jerome-D'Emilia B, Begun JW. Diffusion of breast conserving surgery in medical communities. Soc Sci Med 2005;60:143–151.
22. Birkmeyer JD, Siewers AE, Finlayson EV, *et al.* Hospital volume and surgical mortality in the United States. N Engl J Med 2002;346:1128–1137.
23. Begg CB, Cramer LD, Hoskins WJ, *et al.* Impact of hospital volume on operative mortality for major cancer surgery. JAMA 1998;280:1747–1751.
24. Glasgow RE, Mulvihill SJ. Hospital volume influences outcome in patients undergoing pancreatic resection for cancer. West J Med 1996;165:294–300.
25. Lieberman MD, Kilburn H, Lindsey M, *et al.* Relation of perioperative deaths to hospital volume among patients undergoing pancreatic resection for malignancy. Ann Surg 1995;222:638–645.
26. Romano PS, Mark DH. Patient and hospital characteristics related to in-hospital mortality after lung cancer resection. Chest 1992;101:1332–1337.
27. Harmon JW, Tang DG, Gordon TA, *et al.* Hospital volume can serve as a surrogate for surgeon volume for achieving excellent outcomes in colorectal resection. Ann Surg 1999;230:404–413.
28. Yao SL, Lu-Yao G. Population-based study of relationships between hospital volume of prostatectomies, patient outcomes, and length of hospital stay. J Natl Cancer Inst 1999;91:1950–1956.
29. Gordon TA, Burleyson GP, Tielsch JM, *et al.* The effects of regionalization on cost and outcome for one general high-risk surgical procedure. Ann Surg 1995;221:43–49.

30. Purves H, Pietrobon R, Hervey S, *et al.* Relationship between surgeon caseload and sphincter preservation in patients with rectal cancer. Dis Colon Rectum 2005;48: 195–204.

31. Schulman KA, Berlin JA, Harless W, *et al.* The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med 1999;340:618–626.

32. Greenwald HP, Polissar NL, Borgatta EF, *et al.* Social factors, treatment, and survival in early-stage non-small cell lung cancer. Am J Public Health 1998;88:1681–1684.

33. Richards RJ, Reker DM. Racial differences in use of colonoscopy, sigmoidoscopy, and barium enema in Medicare beneficiaries. Dig Dis Sci 2002;47:2715–2719.

34. Eley JW, Hill HA, Chen VW, *et al.* Racial differences in survival from breast cancer: results of the National Cancer Institute Black/White Cancer Survival Study. JAMA 1994;272:947–954.

35. Dayal HH, Polissar L, Dahlberg S. Race, socioeconomic status, and other prognostic factors for survival from prostate cancer. J Natl Cancer Inst 1985;74:1001–1006.

36. Dayal H, Polissar L, Yang CY, *et al.* Race, socioeconomic status, and other prognostic factors for survival from colorectal cancer. J Chronic Dis 1987;40:857–864.

37. Cooper GS, Yuan Z, Landefeld CS, *et al.* Surgery for colorectal cancer: race-related differences in rates and survival among Medicare beneficiaries. Am J Public Health 1996;86:582–586.

38. Bach PB, Cramer LD, Warren JL, *et al.* Racial differences in the treatment of early-stage lung cancer. N Engl J Med 1999;341:1198–1205.

39. Guller U, Jain N, Curtis L, *et al.* Insurance status and race represent independent predictors of undergoing laparoscopic surgery for appendicitis: secondary data analysis of 145,546 patients. J Am Coll Surg 2004;199: 567–577.

40. Flum DR, Morris A, Koepsell T, *et al.* Has misdiagnosis of appendicitis decreased over time? A population-based analysis. JAMA 2001;286:1748–1753.

41. Guller U, Hervey S, Purves H, *et al.* Laparoscopic versus open appendectomy: outcomes comparison based on a large administrative database. Ann Surg 2004;239:43–52.

42. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878–1886.

43. Lloyd SS, Rissing JP. Physician and coding errors in patient records. JAMA 1985;254:1330–1336.

44. Steiner C, Elixhauser A, Schnaier J. The healthcare cost and utilization project: an overview. Eff Clin Pract 2002;5:143–151.

45. Losina E, Barrett J, Baron JA, *et al.* Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. J Clin Epidemiol 2003;56:515–519.

46. Romano PS, Roos LL, Luft HS, *et al.* A comparison of administrative versus clinical data: coronary artery bypass surgery as an example; Ischemic Heart Disease Patient Outcomes Research Team. J Clin Epidemiol 1994;47: 249–260.

47. Jollis JG, Ancukiewicz M, DeLong ER, *et al.* Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med 1993;119:844–850.

48. Cooper GS, Yuan Z, Stange KC, *et al.* The sensitivity of Medicare claims data for case ascertainment of six common cancers. Med Care 1999;37:436–444.

49. Cooper GS, Yuan Z, Stange KC, *et al.* The utility of Medicare claims data for measuring cancer stage. Med Care 1999;37:706–711.

50. Cooper GS, Yuan Z, Stange KC, *et al.* Agreement of Medicare claims and tumor registry data for assessment of cancer-related treatment. Med Care 2000;38: 411–421.

51. Hsia DC, Krushat WM, Fagan AB, *et al.* Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med 1988;318:352–355.

52. Guller U, DeLong ER. Interpreting statistics in medical literature: a vade mecum for surgeons. J Am Coll Surg 2004;198:441–458.

53. Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–349.

54. Katz MH. What are confounders and how does multivariate analysis help me to deal with them. In: Multivariable Analysis. Cambridge University Press, Cambridge, 1999.

55. Klungel OH, Martens EP, Psaty BM, *et al.* Methods to assess intended effects of drug treatment in observational studies are reviewed. J Clin Epidemiol 2004;57: 1223–1231.

56. Byar DP, Simon RM, Friedewald WT, *et al.* Randomized clinical trials: perspectives on some recent ideas. N Engl J Med 1976;295:74–80.

57. Stewart LA, Parmar MK. Bias in the analysis and reporting of randomized controlled trials. Int J Technol Assess Health Care 1996;12:264–275.

58. Bailey KR. Generalizing the results of randomized clinical trials. Control Clin Trials 1994;15:15–23.

59. Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol 1999;52:487–497.

60. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996;313:570–571.

61. McLeod RS. Issues in surgical randomized controlled trials. World J Surg 1999;23:1210–1214.

62. Young J, Harrison J, White G, *et al.* Developing measures of surgeons' equipoise to assess the feasibility of randomized controlled trials in vascular surgery. Surgery 2004;136:1070–1076.

63. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–1892.

64. Kennedy WA, Laurier C, Malo JL, *et al.* Does clinical trial subject selection restrict the ability to generalize use and

cost of health services to "real life" subjects? Int J Technol Assess Health Care 2003;19:8–16.

65. Crawford ED. On the importance of clinical trials. J Urol 1990;143:787.
66. Cummings SR, Grady D, Hulley SB. (2001) Designing an Experiment: Clinical Trials I. Designing Clinical Research. Lippincott Williams & Wilkins, Philadelphia, pp 143–155.
67. Helms PJ. 'Real world' pragmatic clinical trials: what are they and what do they tell us? Pediatr Allergy Immunol 2002;13:4–9.
68. Roland M, Torgerson DJ. What are pragmatic trials? BMJ 1998;316:285.
69. Clarke CE. A "cure" for Parkinson's disease: can neuroprotection be proven with current trial designs? Mov Disord 2004;19:491–498.
70. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996;125: 605–613.
71. Guller U, Blumenstein BA. Trends in clinical trials in surgical oncology: implications for outcomes research. Clin Ther 2003;25:(2) 684–698.
72. Ellenberg S, Hamilton JM. Surrogate endpoints in clinical trials: cancer. Stat Med 1989;8:405–413.
73. D'Agostino RB Jr. Debate: the slippery slope of surrogate outcomes. Curr Control Trials Cardiovasc Med 2000;1: 76–78.
74. Guller U, Oertli D. Sample size matters: a guide for surgeons. World J Surg 2005;29:601–605.
75. Berwick DM. Experimental power: the other side of the coin. Pediatrics 1980;65:1043–1045.
76. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122–124.
77. Vandekerckhove P, O'Donovan PA, Lilford RJ, *et al.* Infertility treatment: from cookery to science: the epidemiology of randomised controlled trials. Br J Obstet Gynaecol 1993;100:1005–1036.
78. Freiman JA, Chalmers TC, Smith H Jr, *et al.* The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 "negative" trials. N Engl J Med 1978;299:690–694.
79. Lilford R, Braunholtz D, Harris J, Gill T. Trials in surgery. Br J Surg 2004;91:6–16.
80. Montori VM, Kleinbart J, Newman TB, *et al.* Tips for learners of evidence-based medicine. 2. Measures of precision (confidence intervals). Can Med Assoc J 2004;171: 611–615.
81. Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA 1983;249:1743–1745.
82. Moseley JB, O'Malley K, Petersen NJ, *et al.* A controlled trial of arthroscopic surgery for osteoarthritis of the knee. N Engl J Med 2002;347:81–88.
83. Albin RL. Sham surgery controls: intracerebral grafting of fetal tissue for Parkinson's disease and proposed criteria for use of sham surgery controls. J Med Ethics 2002;28:322–325.
84. Macklin R. The ethical problems with sham surgery in clinical research. N Engl J Med 1999;341:992–996.
85. Devereaux PJ, Bhandari M, Clarke M, *et al.* Need for expertise based randomised controlled trials. BMJ 2005;330:88.
86. Bais JE, Bartelsman JF, Bonjer HJ, *et al.* Laparoscopic or conventional Nissen fundoplication for gastro-oesophageal reflux disease: randomised clinical trial: The Netherlands Antireflux Surgery Study Group. Lancet 2000;355:170–174.