Journal of Gastrointestinal Surgery (2020) 24:1127–1137
https://doi.org/10.1007/s11605-020-04536-3

Big Data vs. Clinical Trials in HPB Surgery

Susanna W.L. de Geus[1] · Teviah E. Sachs[1] · Jennifer F. Tseng[1]

Received: 21 June 2019 / Accepted: 29 January 2020 / Published online: 19 February 2020

Abstract
Randomized controlled clinical trials (RCTs) are at the heart of "evidence-based" medicine. However, in surgical practice, RCTs remain uncommon. Conducting well-designed RCTs for surgical procedures is often challenged by inadequate recruitment accrual, blinding, or standardization of the surgical procedure, as well as lack of funding and evolution of the treatment strategy during the many years over which such trials are conducted. In addition, most clinical trials are performed in academic high-volume centers in highly selected patients, which may not necessarily reflect a "real-world" practice setting. Over the past decades, surgical outcomes research using nationwide administrative and registry databases has become increasingly common. Large databases provide easy and inexpensive access to data on a large and diverse patient population at a variety of treatment centers. Furthermore, large database studies provide the opportunity to answer questions that would be impossible or very arduous to answer using RCTs, including questions regarding health policy efficacy, trends in surgical practice, access to health care, impact of hospital volume, and adherence to practice guidelines, as well as research questions regarding rare disease, infrequent surgical outcomes, and specific subpopulation. Prospective data registries may also allow for quality benchmarking and auditing. This review outlines the role, advantages, and limitations of RCTs and large database studies in answering important research questions in surgery.

Keyword Randomized clinical trials · large database studies

Introduction

The term evidence-based medicine (EBM) was first coined in 1992 by Gordan Guyatt in an article in JAMA, and quickly became the sine qua non of medical practice.[1,2] However, the struggle to balance the uncontrolled experience of physicians with observations obtained by rigorous empirical evaluation of the effect of health interventions goes back to the time of Hippocrates.[3] The core epistemology of EBM is that scientific evidence should guide clinical decision-making, and the extent to which we believe and implement that evidence is determined by a credible process.[3] EBM views randomized controlled trials (RCTs) and meta-analyses of those trials as the "gold standard" for evidence-based practice.[3]

Surgical interventions remain less likely to be investigated using full-scale RCTs than medical therapies, and the number of surgical RCTs has decreased over the past decades, especially in the USA.[4,5] Furthermore, the quality of many surgical RCTs that are published is lacking.[6,7] Conducting RCTs in surgery differs substantially from medicine, as surgical procedures are more difficult to standardize than pharmacologic or radiotherapeutics interventions. Every surgery is unique, since it is highly dependent of the skill and learning curve of the surgeon, as well as the patient's anatomic variation and habitus.[8–10] In addition, surgical RCTs are often challenged by inadequate patient accrual, blinding, and funding.[8]

Although RCTs have long been presumed to be the ideal source for evidence regarding treatment effects, there is growing interest in other methods of obtaining evidence for decisive action, giving rise to new research methods, such as the use of large database studies, to leverage the strengths and overcome the limitations of RCTs.[11] The purpose of this review is to reflect on the role, advantages, and limitations of RCT and large database studies in surgical science.

Advantages of Surgical Trials

The main reason why RCTs are considered more rigorous than other methods is that randomization of study subjects not only controls for measurable confounders between two treatment

Communication: Presentation Surgery of the Alimentary Tract (SSAT) debate, "Big Data are Better," San Diego, California, May 2019.

✉ Jennifer F. Tseng
Jennifer.Tseng@BMC.org

[1] Department of Surgery, Boston Medical Center, Boston University School of Medicine, Boston, MA, USA

groups, but also balances unmeasured confounders.[12,13] In addition, although often challenging in surgical research, RCTs allow for blinding of participants, physicians, outcomes assessors, and data analysts to reduce detection and performance bias.[14] Therefore, well-designed and conducted RCTs are able to identify causal relationships and establish definitively which treatment methods is superior.[11] Previously, many procedures were introduced into surgical practice solely based on observational data until a randomized trial disproved their efficacy. Surgical procedures such as the extra cranial intracranial bypass operation for stroke prevention were common practice for at least 15 years before and an RCT demonstrated that it actually increased the likelihood of developing a stroke.[15,16] In similar fashion, the unmerited use of internal mammary ligation for ischemic heart disease was discredited.[15,17]

Limitation of Surgical Trials

Poor Accrual

There are several barriers to conducting surgical RCTs. It is often difficult to accrue a sufficient number of patients for RCTs in a timely and cost-effective manner. Unless this procedure or disease process is a relatively common one, it can take months to years to collect sufficient patients in a single-institutions to power a study adequately. While multi-institutional studies can recruit patients more quickly, such studies are resource intensive and demand complicated coordination to assure consistent protocol application.[18] In addition, the lack of clinical equipoise can be another obstacle to randomization and trial inclusion. Surgeons are often convinced that what they do is the best for their patients, when other ways to achieve comparable or even superior results might actually exist. In particular, surgeons pioneering a new technique are often "true-believers," and less likely to participate in a RCT. Furthermore, surgeons are frequently not reimbursed, or only partially reimbursed, for performing additional therapeutic or diagnostic interventions, which makes participation in a clinical trial less attractive.[8] Another problem applies to the competitive culture in which surgeons' work. Many surgeons may not enroll patients in trials where the patients are assigned to a non-operative study arm, because of competition among surgeons to recruit patients and out of fear to losing a source of referrals.[19]

As a result of poor accrual, a third of RCTs remains unpublished, and 20% of trials are discontinued at 5 years.[8,7,20] In addition, a third of published surgical trials are underpowered to demonstrate clinically significant differences. Unfortunately, negative findings in underpowered trials are often interpreted as showing the equivalence of the treatment arms with no discussion of the issue of being underpowered, resulting in a type II error.[7,21] This may lead clinicians to accept new treatments that have not been validated.[21]

Generalizability

RCTs often have strong internal validity, but sometimes lack external validity; generalizations of findings outside the study population may be invalid.[11] Although RCTs exist on a continuum, with a progression from efficacy to effectiveness studies. The primary goal of efficacy trials is to determine whether an intervention produced the expected result under ideal circumstances, whereas effectiveness trials (also known as pragmatic trials) are designed to inform general guidelines, clinical or policy decisions by measuring the degree of real-world effectiveness.[22] Nonetheless, in reality, the majority of RCTs are optimized to determine efficacy and may not necessary adequately inform practice.[23] The treatment setting and the patients included in most RCTs do not reflect "real" world population. Previous studies have shown that only approximately 2 to 3% of all patients with cancer ever enroll in a trial.[24,25] Traveling to and receiving care at tertiary medical centers where a trial involving complex operations are often conducted requires considerable financial resources and often takes patients away from their family and support system. Consequently, racial and ethnic minorities, elderly, and women are less likely to enroll in RCTs. In addition, patients enrolled in trials are often healthier, more compliant and of a higher socioeconomic status.[24,26,19] Furthermore, these types of studies are often performed in highly controlled conditions, with strict in- and exclusion criteria.[18] Limiting the generalizability of the findings in RCTs to "real" world practice.

Resource Intensive

RCTs are resource intensive with regard to costs and time.[11] The lion's share of funding to carry out RCTs comes from two main sources: industry and the federal government. While financial support is often readily available for pharmaceutical trials, there are fewer industry sponsors of surgical research.[19] The majority of industry research payments towards surgeons are related to novel pharmaceuticals, with the most funding being procured from Novartis, Amgen, and Merck.[27] Previous studies have also shown that surgical grant proposals are less likely to be funded by the National Institute of Health (NIH) and carry significantly smaller awards compared to nonsurgical proposals.[28] The latter has been partly explained by the low percentage of surgeon-scientist participation in the reviewing of NIH grant proposals. High costs may sometimes result in RCT designs with inadequate sample size.

In addition, RCTs often take years to plan, implement, and analyze reducing the ability of RCTs to keep pace with clinical innovations; new products and standards of care are often developed before earlier studies' complete evaluation.[11] This

makes trial outcomes at risk of becoming obsolete before they get published. Consequently, some interventions have been widely adopted without rigorous evaluation.[29] The increasingly high costs and time constraints of RCTs can also lead to reliance on surrogate markers that may not correlate well with the outcome of interest and create additional bias.[11]

Bias

Many surgeons believe that every RCT is bias-free. This belief is not, in fact, reflected in reality, as poorly designed and conducted RCTs provide distorted, confounded results that are not beneficial for improving current surgical practice.[19] A common problem with conducting surgical RCT is adequate blinding, due to the frequent lack of placebo controls (surgical placebo, sham surgery). A surgical placebo represents a simulated operation in which the skin incisions are done without actually performing the operation. In most cases, both the patient and surgeon are able to determine which procedure was done, potentially leading to post-randomization bias.[18] Another frequent criticism of surgical RCTs, particularly when evaluating a new innovation against a standard intervention, is that the comparison may be inherently "unfair" due to an imbalance in expertise.[29]

Advantages of Big Data

Widely Accessible

The first large database study in the USA is the Framingham Heart Study, which was initiated in 1948 to study risk factors for cardiovascular disease.[30] Data from this large longitudinal database have resulted in 3819 publications to data.[31] In recent decades, the use of big data for research studies has increased significantly (Fig. 1), and currently nine of ten research papers published in clinical specialty journal describe observational research.[32,33] This development has been fueled in large part by the Health Information Technology for Economic and Health Act of 2009 which helped fund the adoption of electronic health records which in turn facilitated the creation of

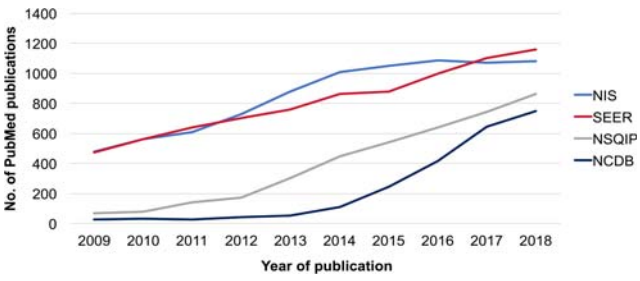

Fig. 1 Trends in the number of PubMed publication of manuscripts using data obtained from the Nationwide Inpatient Sample (NIS), Surveillance, Epidemiology, and End Results (SEER), National Surgical Quality Improvement Program (NSQIP), and National Cancer Database (NCDB)

large clinical databases.[34] The first surgery-specific database was the Veterans Affairs National Quality Improvement Program (VA NSQIP), which was created in response to concerns regarding high mortality rates in the VA system and ultimately gave rise to the American College of Surgeons (ACS) NSQIP in 2004.[35] Other frequently used databases include the National Cancer Database (NCDB); the Surveillance, Epidemiology, and End Results (SEER) database; Healthcare Costs and Utilization Project (HCUP) National Inpatient Samples; and Medicare Claims Data.[36–39]

Outcomes research is in general defined as any study of the end results of health services, including mortality, physiological functional measures, definable clinical events, and patient satisfaction.[40] For the purpose of this study, we use a more narrow definition of outcomes research, only referring to outcomes research that is observational in nature and performed using multi-institutional pro- or retrospectively collected datasets. The use of large database studies for surgical research has several advantages. First, large databases are easy and cheap (sometimes free) to obtain, and most of the time the data can be analyzed using ubiquitous statistical programs.[12] Second, in contrast to RCTs, large database studies often provide sufficient power to detect a significant difference. This may be of particular importance, considering the relative rarity of certain diseases treated by hepato-pancreato-biliary (HPB) surgeons, such as hilar cholangiocarcinoma and pancreatic neuroendocrine tumors. In addition, the large sample sizes of nationwide databases also provides the opportunity to answer a wide variety of research questions with sufficient statistical power, as well as study rare diseases, infrequent postoperative outcomes, and subsets of patients that benefit the most from a specific procedure.[24,26,19] Third, RCTs often adhere to strict inclusion and exclusion criteria, and are often performed at high-volume academic centers in highly selected patients, which limits the generalizability of its findings. Large database studies include a wide variety of patients and treatment centers, and allow for the investigation of "real" world practice patterns, treatment efficacy, and outcomes.[12]

Finally, outcomes research 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.[19,41] Large database studies have shown to be ideal to investigate the impact of hospital volume, access to health care, geographic variations in care, risk stratification protocols, trends in practice patterns, adherence to practice guideline, effectiveness of novel treatment strategy, and also to evaluate health care policy effectiveness.[34,42] Furthermore, large data registries are a useful tool for quality benchmarking. A major strength of the ACS NSQIP is not only that it provides granular risk-adjusted and case-mix-adjusted surgical outcomes data, but also that it allows participating hospital to benchmark their performance to an estimate average of all

hospitals providing data to NSQIP, which has resulted in significantly reduced morbidity and mortality in these centers.[43,44] This is similar to large nationwide audit programs common in Europe.[45–47] In addition, some of the most important published surgical research is based on retrospective studies reported in high impact-factor journal, often because randomizing patients between two very diverse treatment arms is impossible, although this limitation is not always explicitly stated as the reason for not undertaking a prospective RCT.[48–51] Much can be learned from the use of prospective registries for the introduction of innovative procedures in other aspects of surgery, such as the introduction of associating liver partition and portal vein ligation (the ALPSS procedure), the modified 2-stage hepatectomy procedure for liver tumors—for which the registry demonstrated increased perioperative morbidity in older patients compared to conventional liver resection procedures.[52]

Limitation of Big Data

Statistical Versus Clinical Significant

The large sample size available in administrative data sets have the potential to reveal statistical significance even when very small absolute differences exist. Although the conventional threshold for statistical significance of $P < 0.05$ is widely used, one should keep in mind that this threshold is arbitrary.[53] Previous studies have shown that when the total sample size of two groups combined exceeds 250,000, the p value will meet traditional significance levels (i.e., a of 0.05) without substantial differences in outcomes.[54] Disproportionate focus on a P value of less than 0.05 can exaggerate the importance of statistically significant, but clinically meaningless results. Similarly, this approach can cast aside potentially meaningful information obtained simply because the P value exceeds an arbitrary threshold. This practice carries a higher risk of type I error, concluding that a treatment is effective or a difference exists between two groups when in reality the treatment is not effective or no difference exists.[12]

In particular, any large database analysis that does not begin with an a priori hypothesis is susceptible to "data mining" or "data dredging"—a non-hypothesis-driven quest for a statistically significant result.[53] In addition, the difference in the effect estimate should be reported as a patient-centered, clinically meaningful, and interpretable difference in addition to the statistical result.[53,55] The use of confidence intervals (CIs) should help distinguish between clinical significance and statistical significance. CIs are in general more informative when comparing two treatment groups because they are generated around the absolute or relative difference between those populations.[54] When reporting the results of observational studies, authors should also consider following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.

Coding Errors

Although databases have the ability to investigate a wide range of surgical hypotheses, the information included in different national databases is highly variable and, as such, the questions that can be answered and the conclusions that can be reached are restricted by the extent of the available data (Table 1).[42] There are two types of databases: administrative and registry databases (Fig. 2). Administrative databases are generally assembled from billing information and were not created for clinical research. These databases obtain their information typically from two sources: requests to insurers for healthcare payments and claims for clinical services. In general, reimbursed procedures are more often coded accurately, while the coding of comorbidities and complications (other than death) may be less dependable. Over-coding has been reported as a potential source of distortion of administrative data. For instance, if hospitals are reimbursed based on the complexity of the patient's disease, there may be a propensity to over-code primary and secondary diagnoses, a phenomenon called diagnosis-related group (DRG) creep.[56] Miscoding constitutes an innate limitation that must be carefully considered when interpreting the results of studies based on administrative data (Table 2).[19]

Clinical or registry databases, on the other hand, are composed of a given patient population with a priori defined patient information. These databases were created to record and track information, allowing for the investigation of specific clinical questions.[57] Compared with administrative data, registry data are widely considered to have a greater degree of accuracy, mainly due to (I) a greater level of clinical training within the staff abstracting data and (II) a rigorous set of clinical criteria used in interpreting clinical phenomena. When registry data are considered a gold standard, these comparisons find administrative data to have rates of false positives ranging from 48 to 84% and false negatives ranging from < 1 to 5%.[58]

A common limitation of both administrative and registry databases is that they are dependent on International Statistical Classification of Disease, ninth edition (ICD-9) or ICD-10, and Current Procedural Terminology (CPT) coding to isolate comorbidities, diagnoses, procedures, and complications. These codes were not originally created for research purposes and their use may only be valid for certain diagnoses, procedures, or complications.[59,60] Variations in coding may be caused by clerical errors or different interpretations of discharge summaries, operative records, or other healthcare documentation. Diagnostic accuracy may also differ between types of treatment facilities; challenging diagnoses that require extensive workup by an experienced physician may be more accurate coming from a tertiary referral center than a community hospital.[54,42]

Table 1 Overview characteristics of the Medicare Claims Data, Healthcare Cost, and Utilization Project National Inpatient Sample (NIS), National Surgical Quality Improvement Program (NSQIP), National Cancer Database (NCDB), and the Surveillance, Epidemiology, and End Results (SEER) database

Database	Population	Advantages	Limitations
Medicare Claims Data36	- 70% of adults aged 65 years and older - People who qualify for Social Security Administration disability benefits.	- The data sets available from the Centers for Medicare and Medicaid Services (CMS) are suitable for linkage to several existing data sets (e.g., other CMS data, SEER, Medicaid). - Data can be tracked longitudinally across episodes of care, making this a uniquely positioned dataset to study long-term outcomes in surgical patients.	- Only includes diagnosis documented via international Classification of Disease, Ninth version (ICD-9) or ICD-10 codes. - No physiological or biochemical patient information, such as vital signs, laboratory test results, and pathology results. - Lack of data on uncovered outpatient services and managed care information.
NCDB38	- 70% of all newly diagnosed cancer cases in the USA. - Not population based, only patients treated at commission on cancer approved centers.	- Strengths of the NCDB are in examining treatment patterns and trends over time across the USA.	- Only reports treatment that was used in the 6 months after diagnosis. - Surgical procedures reported will only include the most definitive intervention. - The readmission variable only captures readmission to the same hospital within 30 days of discharge (reporting bias).
NSQIP44	- Random sampling of one out of eight cases performed at the ± 700 hospitals participating in NSQIP. - Does not represent a valid nationally representative sample. - Excludes trauma and transplant cases.	- Provides data on a broad range of 30-day outcomes, including mortality, readmission, and length of stay, as well as timing of postoperative discharge complications. - Provides the ability to account for preoperative comorbidity, as well as complications that occur in the perioperative period. - Targeted NSQIP for hepatectomy and pancreatectomy.	- Does not contain hospital or clinical identifiers. - No data on type of insurance, type of treatment facility, and surgeon or hospital volume. - Follow-up limited to 30 days.
NIS39	- 20% representative sampling of all inpatient hospital encounters in the USA. - Designed to be representative for health care use overall.	- Ideal for researching national prevalence/incidence, changes over time, and associations between diagnosis, procedures, and outcomes.	- Lack of longitudinal data. - Only diagnosis identified by ICD-9 ICD-10 codes. - Systematic undercoding of certain low-cost diagnostic procedures can lead to inaccurate estimations of procedure us. - Redesign of NIS in 2012.
SEER37	- Cancer cases only. - Population-based. - Captures 28% of the US population.	- Includes a high proportion of racial/ethnic minorities, foreign-born individuals, and those with income below the poverty line. - Longitudinal trends in cancer incidence, prevalence, treatment, and survival can be analyzed starting from 1974 to the present. - Longitudinal studies on specific subpopulations and rare or indolent cancer types.	- Comparative effectiveness analyses are limited by lack of information on comorbidities, surgical approach (minimally invasive vs. open), systemic treatment (chemotherapy, hormonal therapy, or immunotherapy), radiation dose, and recurrence. - HER2 status is coded inconsistently and should not be used in analysis.[84]

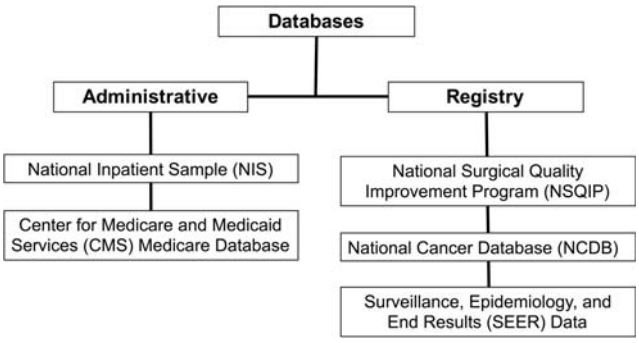

Fig. 2 Administrative and registry databases commonly used for surgical outcomes research in hepato-pancreato-biliary surgery

Missing Data

Missing data are common in clinical research, particularly for variables requiring complex, time-sensitive, resource-intensive, or longitudinal data collection methods.[61] Some variables are missing at random, which does not necessarily presume patients with missing values are similar to those with complete data, but instead presumes that observed values can be used to "explain" which values are missing and assist predicting what the missing values would be. There are various methods of dealing with missing variables. Often patients with missing variables are omitted from an analysis, which is known as complete case analysis and is the default methods

Table 2 Checklist for good conduct large database research

- □ Identify a hypothesis driven research question
- □ Choose the appropriate data set to address the question of interest (see Table 1).[38]
- □ A flow diagram should be included that shows the number of patients included and excluded, along with reasons for exclusion, documenting a stepwise derivation of the final sample.[85–87]
- □ All predictor and outcomes variables should be defined a priori.[87]
- □ A justification should be provided regarding categorizing continuous variables.[87]
- □ Check or variables of interest have changed over time.3[8]
- □ Shifts in cancer stage classifications over time should be accounted for[37]
- □ Define study population according to Current Procedural Terminology (CPT) codes and then validate this population using the International Classification of Disease Ninth Revision (ICD-9) and ICD-10 diagnostic codes.[44]
- □ Account for differences between the International Classification of Disease, Ninth Edition (ICD-9) and ICD-10 systems.[39]
- □ When the study spans many years, determine whether the period qualitatively changes the study results, if present, consider a stratified analysis.[38]
- □ Variables with less than 50% of data available for analysis should be discarded.[38]
- □ In case, over 30% of patients has missing variables, multiple imputation should be used to control for missing variables. In addition, the cause of 'missingness' should be investigated and described.[61,85]
- □ In case of health policy evaluation, difference-in-difference assessments should be performed to controlled for any unrelated changed over time.[36]
- □ For multivariable analyses, variable selection should be based on prior evidence and biological/clinical plausibility, not necessary any variable that is statistically significant.[86,87]
- □ In case multivariable models include variables based on statistical significant criteria, model performance statistics and whether multicollinearity and effect modification were assessed should be specified.[87]
- □ For logistic multivariable regression analysis, coefficients should be interpreted using odds ratios, while linear and Poisson models should incorporate effect size.[86]
- □ In case of rare events of interest (less than 10–15 events per variable in the model), the propensity score method should be used instead of a multivariable analysis.[53]
- □ In case the covariates of two groups under investigation are not sufficiently overlapping, the propensity score method should be used instead of a multivariable analysis.[53]
- □ Check for immortal time bias, especially in studies investigating the efficacy of (neo)adjuvant therapy or transplantation, perform a landmark analysis or extended Cox model.[80]
- □ Perform extensive sensitivity analyses to evaluate and address confounding and selection bias[38]
- □ Emphasize practical clinical findings instead of incidental statistically significant results.[39,86]

  A power calculation should be included when dealing with small subgroups or rare disease; the findings for a subgroup or rare disease may be susceptible to bias if the sample size is small.[86]

  In case, the database included facility identifiers, hierarchical analyses should be used including the identifier as the random effect in the model to account for the correlated patient outcomes, as patients are nested within facilities.[87,38,88]
- □ Avoid use of language implying causal inference in reporting results from observational studies; Instead, these studies are best suited for hypothesis generation.[39]
- □ Ensure that your article has a clear take-home message that addresses how your research advances current knowledge and has important policy or clinical implications.[85]

used by most statistical software. The primary limitation of complete case analysis is reduced sample size, resulting in reduced study power. In addition, unless variables are missing completely at random (very unlikely), estimates using observed case analysis will be biased and the direction of the bias unpredictable. In general, multiple imputation is the best approach for modeling the effects of missing data in studies.[61] Multiple imputation uses the available data to predict plausible values for missing data through the use of regression models. Missing data are then replaced with predicted, or imputed, values. By using multiple imputed data sets, the subsequent analyses appropriately consider both the uncertainty of the observed values and the uncertainty of the imputed values, thereby resulting in more valid inferences.[61]

Variables that are missing not at random are the most troublesome and occur when missing values are dependent on unobserved or unknown factors. When variables are missing not at random, statistical adjustment for missing information is effectively impossible. Because an investigator usually cannot establish the actual mechanism for why the date is missing, statistical

analyses usually continue assuming the data are missing at random.[61]

Studies using administrative data should report the extent of missing data, use proper methods to account for missing data in analyses, and describe their potential impact on inferences and conclusions. The proportion of missing data for the variables and outcomes of interest should be clearly discussed in the resulting manuscript. When there is a large proportion of missing data (> 30%), the author should investigate and describe the pattern of "missingness" in the data, and there should be consideration for using techniques such as multiple imputation.[61]

Bias and Confounding

Large database studies are a valuable tool for surgical outcomes research. However, a constant challenge in observational designs is to rule out bias.[62] Bias is the systematic deviation of study results or inferences from the truth. Because bias can lead to erroneous conclusions, its minimization is pivotal to all good research.[63] The two most common types of bias are selection bias and confounding. Selection bias arises when certain types of patients are more or less likely to receive treatment owing to possible confounding by indication.[64] Interestingly, a common and often intractable form of confounding results from good medical practice: overall healthier patients tend to undergo more aggressive treatments, which improves survival seen with these more aggressive treatments, but may actually be more of a sign of the patients' overall health at diagnosis rather than the treatment itself.[65] Another common, but often unrecognized, type of selection bias is immortal time bias, also known as guarantee-time bias or survivorship bias, which occurs when a time-dependent exposure (such as initiation of a medical treatment) is not included appropriately in an analysis of a survival outcome. It is termed immortal time bias, because patients must survive sufficiently long enough to receive treatment; hence, they are immortal by definition before exposure.[66] The latter places a disproportionate number of the early deaths in the control group, lowers its survival rate, and artificially makes the treatment group seem better in comparison.[67] In a systematic review, over 40% of studies with a survival end point and time-varying treatment were susceptible to immortal time bias.[63]

Confounding stems from measured or unmeasured factors that affect the outcome of interest and are unevenly distributed among study arms. A variable may introduce confounding only if it manifests three characteristics. First, it must be a risk factor for the outcome of interest. Second, it must be associated with the exposure of primary interest. Finally, it must not be affected by the exposure or the outcome of interest.[12] For example, when extended lymphadenectomy is more commonly performed at high-volume treatment centers, studies may demonstrate that patients undergoing extended lymphadenectomy have better long-term outcomes compared to patients who did not undergo extended lymphadenectomy. However, the improved survival is actually caused by receiving better care at a high-volume center, which is in this case the confounder.

Statistical Consideration

There are several approaches to dealing with potential selection bias and confounding. However, multivariate regression is the most often used technique to adjust for the presence of confounding variables. When using a multivariable model, the theoretical rationale of the model should be reported. The type of model (e.g., logistic, linear, Poisson) and the assumptions on which it is based should be clearly stated (e.g., the model assumed linearity or normality of the distribution of the data). The authors should demonstrate that model's assumptions were not violated (e.g., the hazard are proportional), thereby confirming the validity of the model. In addition, it should be clearly stated why certain predictor variables and which variables were chosen for the model. Ideally, a model and its predictors will not be selected based on statistical significance. Rather, the predictor variables should be chosen based on background literature and/or biological and clinical plausibility. If selection is performed purely based on statistical significance, the model should be presented as hypothesis-generating, rather than conclusive.[68,53] Furthermore, for every covariate included in the model, there should be at least 10 to 15 participants with the outcome of interest.[53]

Propensity score methodology can be especially useful when a treatment is common but the outcome of interest is rare, a situation in which multivariate regression analysis is particularly troublesome.[69] In addition, propensity score methods should be preferred over multivariable regression strategies when the distribution of the covariates of the two treatment groups under investigation do not overlap sufficiently.[70] With propensity scores, patient and provider characteristics are used to calculate the probability that a patient will receive the intervention of interest.[71,72] There are 4 general ways these propensity scores can be further used. The most common is propensity score matching, which involves assembling 2 groups of study participants, one group that received the treatment of interest and the other that did not, while matching individuals with similar or identical propensity scores.[73,74] Other methods include stratification on the propensity score, covariate adjustment using the propensity score, and inverse probability of treatment weighting using the propensity score.[74,75] In general, propensity score matching minimizes bias to a greater extent than propensity score stratification.[74] Previous studies have demonstrated that propensity score methods eliminate approximately 90% of the bias.[76–78] In addition, in a review of treatment effects of

published surgical studies, results by RCT and non-RCT studies were found to be very similar when non-RCT data were analyzed after matching by use of propensity analysis.[79]

Immortal time bias cannot be controlled for using multivariable models or propensity score methods. The common techniques to control or remove immortal time bias are conditional landmark analysis, time-dependent Cox regression model, and inverse probability weighting.[80] Condition landmark analyses are most frequently used: a fixed time point after the initiation of follow-up is chosen as a landmark for conducting the analysis.[81] Treatment status (exposure) is determined at the landmark, with patients having the event of interest or censored before the landmark excluded from the analysis. Patients who initiate treatment after the landmark are included in the no-exposure group.

It is critical to remember, however, that propensity score techniques, or any of the other statistical methods, can only reliably account for measured determinants of treatment selection, but not for unknown confounding. Nonetheless, as pointed out by Birkmeyer and colleagues, if the differences are large even after adjusting for putative confounding factors, it can be presumed that they cannot be explained solely by residual confounding.[82,19] In addition, even more important than post hoc adjustment—because it is never perfect—is the thorough investigation and description of the potential impact of both overt and hidden biases in the manuscript of a study.[12] Furthermore, extensive sensitivity analyses should be performed to test the robustness of outcomes for confounders and missing data.[12]

Conclusions

RCTs in surgery rightfully constitute the most reliable scientific approach to comparative effectiveness studies and should be conducted when feasible. However, considering the rarity of certain surgical conditions, lack of funding, and time constrains, not every research question can be answered by a RCT. Furthermore, RCTS are limited by strict inclusion criteria and exclusion criteria, as well as their highly selected patient population. This raised concerns regarding the translation of results obtain from RCTs into everyday practice. Large database studies are able to provide a "real" world perspective on surgical practice, and could aid in conducting well-designed RCTs by provided pretrial data to enable power calculations, and to clarify the definition and indication of the intervention, as well as to develop quality measures. Furthermore, they could provide external validation of RCT results after completion. There even have been initiatives to integrated the strengths of large database studies and RCTs by performing for example registry-based pragmatic RCTs.[83] In addition, large database studies are able to reflect on research questions regarding the efficacy of health policy, access to health care, and trends and geographic variation in practice patterns, as well as the treatment of rare disease or patients subgroups, which would be impossible or very strenuous by using RCTs. Moreover, large nationwide datasets, such as NSQIP, provide the tremendous opportunity to benchmark surgical outcomes and subsequently improve quality of care. On the other hand, it has been well established that large database studies are prone to bias. Therefore, comprehensive understanding of the limitation of these studies, well-thought study designs, and rigorous statistical analyses are pivotal to conducting worthwhile large database studies.

Funding information This work was supported by the Perlman research scholarship (Susanna W.L. de Geus).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Smith R, Rennie D. Evidence-based medicine–an oral history. JAMA. 2014;311(4):365–7. doi:https://doi.org/10.1001/jama.2013.286182.
2. Evidence-Based Medicine Working G. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992;268(17):2420–5.
3. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415–23. doi:https://doi.org/10.1016/S0140-6736(16)31592-6.
4. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ. 2002;324(7351):1448–51. doi:https://doi.org/10.1136/bmj.324.7351.1448.
5. Farrokhyar F, Karanicolas PJ, Thoma A, Simunovic M, Bhandari M, Devereaux PJ et al. Randomized controlled trials of surgical interventions. Ann Surg. 2010;251(3):409–16. doi:https://doi.org/10.1097/SLA.0b013e3181cf863d.
6. Yu J, Chen W, Chen S, Jia P, Su G, Li Y et al. Design, Conduct, and Analysis of Surgical Randomized Controlled Trials: A Cross-sectional Survey. Ann Surg. 2018. doi:https://doi.org/10.1097/SLA.0000000000002860.
7. Ahmed Ali U, Ten Hove JR, Reiber BM, van der Sluis PC, Besselink MG. Sample size of surgical randomized controlled trials: a lack of improvement over time. J Surg Res. 2018;228:1–7. doi:https://doi.org/10.1016/j.jss.2018.02.014.
8. Evrard S, McKelvie-Sebileau P, van de Velde C, Nordlinger B, Poston G. What can we learn from oncology surgical trials? Nat Rev Clin Oncol. 2016;13(1):55–62. doi:https://doi.org/10.1038/nrclinonc.2015.176.
9. Balch CM, Nelson H, Niederhuber JE. Surgery: Limitations of prospective surgical oncology trials - a US view. Nat Rev Clin Oncol. 2016;13(1):6–8. doi:https://doi.org/10.1038/nrclinonc.2015.212.
10. Baum M. Reflections on randomised controlled trials in surgery. The Lancet. 1999;353:S6-S8. doi:https://doi.org/10.1016/s0140-6736(99)90220-9.

11. Frieden TR. Evidence for Health Decision Making - Beyond Randomized, Controlled Trials. N Engl J Med. 2017;377(5):465–75. doi:https://doi.org/10.1056/NEJMra1614394.
12. Nathan H, Pawlik TM. Limitations of claims and registry data in surgical oncology research. Ann Surg Oncol. 2008;15(2):415–23. doi:https://doi.org/10.1245/s10434-007-9658-3.
13. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med. 2016;21(4):125–7. doi:https://doi.org/10.1136/ebmed-2016-110401.
14. Probst P, Grummich K, Heger P, Zaschke S, Knebel P, Ulrich A et al. Blinding in randomized controlled trials in general and abdominal surgery: protocol for a systematic review and empirical study. Syst Rev. 2016;5:48. doi:https://doi.org/10.1186/s13643-016-0226-4.
15. Das AK. Randomised clinical trials in surgery: a look at the ethical and practical issues. Indian J Surg. 2011;73(4):245–50. doi:https://doi.org/10.1007/s12262-011-0307-5.
16. Group EIBS. Failure of extracranial-intracranial arterial bypass to reduce the risk of ischemic stroke. Results of an international randomized trial. N Engl J Med. 1985;313(19):1191–200. doi:https://doi.org/10.1056/NEJM198511073131904.
17. Cobb LA, Thomas GI, Dillard DH, Merendino KA, Bruce RA. An evaluation of internal-mammary-artery ligation by a double-blind technic. N Engl J Med. 1959;260(22):1115–8. doi:https://doi.org/10.1056/NEJM195905282602204.
18. Zhu VZ, Tuggle CT, Au AF. Promise and Limitations of Big Data Research in Plastic Surgery. Ann Plast Surg. 2016;76(4):453–8. doi:https://doi.org/10.1097/SAP.0000000000000750.
19. Guller U. Surgical outcomes research based on administrative data: inferior or complementary to prospective randomized clinical trials? World J Surg. 2006;30(3):255–66. doi:https://doi.org/10.1007/s00268-005-0156-0.
20. Chapman SJ, Shelton B, Mahmood H, Fitzgerald JE, Harrison EM, Bhangu A. Discontinuation and non-publication of surgical randomised controlled trials: observational study. BMJ. 2014;349:g6870. doi:https://doi.org/10.1136/bmj.g6870.
21. Brody BA, Ashton CM, Liu D, Xiong Y, Yao X, Wray NP. Are surgical trials with negative results being interpreted correctly? J Am Coll Surg. 2013;216(1):158–66. doi:https://doi.org/10.1016/j.jamcollsurg.2012.09.015.
22. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS. A simple and valid tool distinguished efficacy from effectiveness studies. J Clin Epidemiol. 2006;59(10):1040–8. doi:https://doi.org/10.1016/j.jclinepi.2006.01.011.
23. Ford I, Norrie J. Pragmatic Trials. N Engl J Med. 2016;375(5):454–63. doi:https://doi.org/10.1056/NEJMra1510059.
24. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. JAMA. 2004;291(22):2720–6. doi:https://doi.org/10.1001/jama.291.22.2720.
25. Stewart JH, Bertoni AG, Staten JL, Levine EA, Gross CP. Participation in surgical oncology clinical trials: gender-, race/ethnicity-, and age-based disparities. Ann Surg Oncol. 2007;14(12):3328–34. doi:https://doi.org/10.1245/s10434-007-9500-y.
26. Lamont EB, Hayreh D, Pickett KE, Dignam JJ, List MA, Stenson KM et al. Is patient travel distance associated with survival on phase II clinical trials in oncology? J Natl Cancer Inst. 2003;95(18):1370–5. doi:https://doi.org/10.1093/jnci/djg035.
27. Santamaria-Barria JA, Stern S, Khader A, Garland-Kledzik M, Scholer AJ, Fischer T et al. Changing Trends in Industry Funding for Surgical Oncologists. Ann Surg Oncol. 2019. doi:https://doi.org/10.1245/s10434-019-07380-1.
28. Weil RJ. The future of surgical research. PLoS Med. 2004;1(1):e13. doi:https://doi.org/10.1371/journal.pmed.0010013.
29. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. Trials. 2009;10:9. doi:https://doi.org/10.1186/1745-6215-10-9.
30. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. The Lancet. 2014;383(9921):999–1008. doi:https://doi.org/10.1016/s0140-6736(13)61752-3.
31. Framingham Heart Study. https://www.framinghamheartstudy.org/fhs-bibliography/. Accessed June 2019.
32. Funai EF, Rosenbush EJ, Lee MJ, Del Priore G. Distribution of study designs in four major US journals of obstetrics and gynecology. Gynecol Obstet Invest. 2001;51(1):8–11. doi:https://doi.org/10.1159/000052882.
33. Scales CD, Jr., Norris RD, Peterson BL, Preminger GM, Dahm P. Clinical research and statistical methods in the urology literature. J Urol. 2005;174(4 Pt 1):1374–9.
34. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–2. doi:https://doi.org/10.1001/jama.2013.393.
35. Fink AS, Campbell DA, Jr., Mentzer RM, Jr., Henderson WG, Daley J, Bannister J et al. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. Ann Surg. 2002;236(3):344–53; **discussion 53-4**. doi:https://doi.org/10.1097/00000658-200209000-00011.
36. Ghaferi AA, Dimick JB. Practical Guide to Surgical Data Sets: Medicare Claims Data. JAMA Surg. 2018;153(7):677–8. doi:https://doi.org/10.1001/jamasurg.2018.0489.
37. Doll KM, Rademaker A, Sosa JA. Practical Guide to Surgical Data Sets: Surveillance, Epidemiology, and End Results (SEER) Database. JAMA Surg. 2018;153(6):588–9. doi:https://doi.org/10.1001/jamasurg.2018.0501.
38. Merkow RP, Rademaker AW, Bilimoria KY. Practical Guide to Surgical Data Sets: National Cancer Database (NCDB). JAMA Surg. 2018;153(9):850–1. doi:https://doi.org/10.1001/jamasurg.2018.0492.
39. Stulberg JJ, Haut ER. Practical Guide to Surgical Data Sets: Healthcare Cost and Utilization Project National Inpatient Sample (NIS). JAMA Surg. 2018;153(6):586–7. doi:https://doi.org/10.1001/jamasurg.2018.0542.
40. Clancy CM, Eisenberg JM. Outcomes research: measuring the end results of health care. Science. 1998;282(5387):245–6. doi:https://doi.org/10.1126/science.282.5387.245.
41. Porter GA, Skibber JM. Outcomes Research in Surgical Oncology. Annals of Surgical Oncology. 2000;7(5):367–75. doi:https://doi.org/10.1007/s10434-000-0367-4.
42. Alluri RK, Leland H, Heckmann N. Surgical research using national databases. Ann Transl Med. 2016;4(20):393. doi:https://doi.org/10.21037/atm.2016.10.49.
43. Cohen ME, Liu Y, Ko CY, Hall BL. Improved Surgical Outcomes for ACS NSQIP Hospitals Over Time: Evaluation of Hospital Cohorts With up to 8 Years of Participation. Ann Surg. 2016;263(2):267–73. doi:https://doi.org/10.1097/SLA.0000000000001192.
44. Raval MV, Pawlik TM. Practical Guide to Surgical Data Sets: National Surgical Quality Improvement Program (NSQIP) and Pediatric NSQIP. JAMA Surg. 2018;153(8):764–5. doi:https://doi.org/10.1001/jamasurg.2018.0486.
45. Norstein J, Langmark F. Results of Rectal Cancer Treatment: A National Experience. 1997:17–28. doi:https://doi.org/10.1007/978-3-642-60514-7_2.
46. Dutch Snapshot Research G. Benchmarking recent national practice in rectal cancer treatment with landmark randomized controlled trials. Colorectal Dis. 2017;19(6):O219-O31. doi:https://doi.org/10.1111/codi.13644.

47. van der Werf LR, Kok NFM, Buis CI, Grunhagen DJ, Hoogwater FJH, Swijnenburg RJ et al. Implementation and first results of a mandatory, nationwide audit on liver surgery. HPB (Oxford). 2019. doi:https://doi.org/10.1016/j.hpb.2019.02.021.
48. Adam R, Wicherts DA, de Haas RJ, Ciacio O, Levi F, Paule B et al. Patients with initially unresectable colorectal liver metastases: is there a possibility of cure? J Clin Oncol. 2009;27(11):1829–35. doi:https://doi.org/10.1200/JCO.2008.19.9273.
49. Brouquet A, Abdalla EK, Kopetz S, Garrett CR, Overman MJ, Eng C et al. High survival rate after two-stage resection of advanced colorectal liver metastases: response-based selection and complete resection define outcome. J Clin Oncol. 2011;29(8):1083–90. doi:https://doi.org/10.1200/JCO.2010.32.6132.
50. Elias D, Gilly F, Boutitie F, Quenet F, Bereder JM, Mansvelt B et al. Peritoneal colorectal carcinomatosis treated with surgery and perioperative intraperitoneal chemotherapy: retrospective analysis of 523 patients from a multicentric French study. J Clin Oncol. 2010;28(1):63–8. doi:https://doi.org/10.1200/JCO.2009.23.9285.
51. Curley SA. Radiofrequency ablation versus resection for resectable colorectal liver metastases: time for a randomized trial? Ann Surg Oncol. 2008;15(1):11–3. doi:https://doi.org/10.1245/s10434-007-9668-1.
52. Schadde E, Ardiles V, Robles-Campos R, Malago M, Machado M, Hernandez-Alejandro R et al. Early survival and safety of ALPPS: first report of the International ALPPS Registry. Ann Surg. 2014;260(5):829–36; **discussion 36-8**. doi:https://doi.org/10.1097/SLA.0000000000000947.
53. Kaji AH, Rademaker AW, Hyslop T. Tips for Analyzing Large Data Sets From the JAMA Surgery Statistical Editors. JAMA Surg. 2018;153(6):508–9. doi:https://doi.org/10.1001/jamasurg.2018.0647.
54. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. J Clin Epidemiol. 2012;65(2):126–31. doi:https://doi.org/10.1016/j.jclinepi.2011.08.002.
55. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. JAMA. 2014;312(13):1342–3. doi:https://doi.org/10.1001/jama.2014.13128.
56. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. N Engl J Med. 1988;318(6):352–5. doi:https://doi.org/10.1056/NEJM198802113180604.
57. Murphy M, Alavi K, Maykel J. Working with existing databases. Clin Colon Rectal Surg. 2013;26(1):5–11. doi:https://doi.org/10.1055/s-0033-1333627.
58. Lawson EH, Zingmond DS, Hall BL, Louie R, Brook RH, Ko CY. Comparison between clinical registry and medicare claims data on the classification of hospital quality of surgical care. Ann Surg. 2015;261(2):290–6. doi:https://doi.org/10.1097/SLA.0000000000000707.
59. Goff SL, Pekow PS, Markenson G, Knee A, Chasan-Taber L, Lindenauer PK. Validity of using ICD-9-CM codes to identify selected categories of obstetric complications, procedures and comorbidities. Paediatr Perinat Epidemiol. 2012;26(5):421–9. doi:https://doi.org/10.1111/j.1365-3016.2012.01303.x.
60. Best WR, Khuri SF, Phelan M, Hur K, Henderson WG, Demakis JG et al. Identifying patient preoperative risk factors and postoperative adverse events in administrative databases: results from the Department of Veterans Affairs National Surgical Quality Improvement Program. J Am Coll Surg. 2002;194(3):257–66.
61. Newgard CD, Lewis RJ. Missing Data: How to Best Account for What Is Not Known. JAMA. 2015;314(9):940–1. doi:https://doi.org/10.1001/jama.2015.10516.
62. Norgaard M, Ehrenstein V, Vandenbroucke JP. Confounding in observational studies based on large health care databases: problems and potential solutions - a primer for the clinician. Clin Epidemiol. 2017;9:185–93. doi:https://doi.org/10.2147/CLEP.S129879.
63. van Walraven C, Davis D, Forster AJ, Wells GA. Time-dependent bias was common in survival analyses published in leading clinical journals. J Clin Epidemiol. 2004;57(7):672–82. doi:https://doi.org/10.1016/j.jclinepi.2003.12.008.
64. Hemmila MR, Birkmeyer NJ, Arbabi S, Osborne NH, Wahl WL, Dimick JB. Introduction to propensity scores: A case study on the comparative effectiveness of laparoscopic vs open appendectomy. Arch Surg. 2010;145(10):939–45. doi:https://doi.org/10.1001/archsurg.2010.193.
65. Torgeson A, Tao R, Garrido-Laguna I, Willen B, Dursteler A, Lloyd S. Large database utilization in health outcomes research in pancreatic cancer: an update. J Gastrointest Oncol. 2018;9(6):996–1004. doi:https://doi.org/10.21037/jgo.2018.05.15.
66. Jones M, Fowler R. Immortal time bias in observational studies of time-to-event outcomes. J Crit Care. 2016;36:195–9. doi:https://doi.org/10.1016/j.jcrc.2016.07.017.
67. Kollman C. Survival Analysis and the Immortal Time Bias. JAMA Ophthalmol. 2018;136(11):1314–5. doi:https://doi.org/10.1001/jamaophthalmol.2018.3499.
68. Meurer WJ, Tolles J. Logistic Regression Diagnostics: Understanding How Well a Model Predicts Outcomes. JAMA. 2017;317(10):1068–9. doi:https://doi.org/10.1001/jama.2016.20441.
69. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med. 2002;137(8):693–5.
70. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med. 1997;127(8 Pt 2):757–63.
71. Kao LS, Dimick JB, Porter GA, Evidence-Based Reviews in Surgery G. How do administrative data compare with a clinical registry for identifying 30-day postoperative complications? J Am Coll Surg. 2014;219(6):1187–91. doi:https://doi.org/10.1016/j.jamcollsurg.2014.09.002.
72. Adamina M, Guller U, Weber WP, Oertli D. Propensity scores and the surgeon. Br J Surg. 2006;93(4):389–94. doi:https://doi.org/10.1002/bjs.5265.
73. Roze JC, Cambonie G, Marchand-Martin L, Gournay V, Durrmeyer X, Durox M et al. Association Between Early Screening for Patent Ductus Arteriosus and In-Hospital Mortality Among Extremely Preterm Infants. JAMA. 2015;313(24):2441–8. doi:https://doi.org/10.1001/jama.2015.6734.
74. Haukoos JS, Lewis RJ. The Propensity Score. JAMA. 2015;314(15):1637–8. doi:https://doi.org/10.1001/jama.2015.13480.
75. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. doi:https://doi.org/10.1093/biomet/70.1.41.
76. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011;46(3):399–424. doi:https://doi.org/10.1080/00273171.2011.568786.
77. Cochran WG. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. Biometrics. 1968;24(2):295. doi:https://doi.org/10.2307/2528036.
78. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. Journal of the American Statistical Association. 1984;79(387):516. doi:https://doi.org/10.1080/01621459.1984.10478078.
79. Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Ann Surg. 2014;259(1):18–25. doi:https://doi.org/10.1097/SLA.0000000000000256.

80. Giobbie-Hurder A, Gelber RD, Regan MM. Challenges of guarantee-time bias. J Clin Oncol. 2013;31(23):2963–9. doi:https://doi.org/10.1200/JCO.2013.49.5283.
81. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1(11):710–9. doi:https://doi.org/10.1200/JCO.1983.1.11.710.
82. Birkmeyer JD, Siewers AE, Finlayson EV, Stukel TA, Lucas FL, Batista I et al. Hospital volume and surgical mortality in the United States. N Engl J Med. 2002;346(15):1128–37. doi:https://doi.org/10.1056/NEJMsa012337.
83. Mathes T, Buehn S, Prengel P, Pieper D. Registry-based randomized controlled trials merged the strength of randomized controlled trails and observational studies and give rise to more pragmatic trials. J Clin Epidemiol. 2018;93:120–7. doi:https://doi.org/10.1016/j.jclinepi.2017.09.017.
84. Howlader N, Chen VW, Ries LA, Loch MM, Lee R, DeSantis C et al. Overview of breast cancer collaborative stage data items–their definitions, quality, usage, and clinical implications: a review of SEER data for 2004-2010. Cancer. 2014;120 Suppl 23:3771–80. doi:https://doi.org/10.1002/cncr.29059.
85. Haider AH, Bilimoria KY, Kibbe MR. A Checklist to Elevate the Science of Surgical Database Research. JAMA Surg. 2018;153(6):505–7. doi:https://doi.org/10.1001/jamasurg.2018.0628.
86. Desai SS, Kaji AH, Upchurch G, Jr. Practical Guide to Surgical Data Sets: Society for Vascular Surgery Vascular Quality Initiative (SVS VQI). JAMA Surg. 2018;153(10):957–8. doi:https://doi.org/10.1001/jamasurg.2018.0498.
87. Hashmi ZG, Kaji AH, Nathens AB. Practical Guide to Surgical Data Sets: National Trauma Data Bank (NTDB). JAMA Surg. 2018;153(9):852–3. doi:https://doi.org/10.1001/jamasurg.2018.0483.
88. Massarweh NN, Kaji AH, Itani KMF. Practical Guide to Surgical Data Sets: Veterans Affairs Surgical Quality Improvement Program (VASQIP). JAMA Surg. 2018;153(8):768–9. doi:https://doi.org/10.1001/jamasurg.2018.0504.