Neurosurg Focus 34 (1):E3, 2013
©AANS, 2013

Measuring clinical practice: science and methods

Peter D. Angevine, M.D., M.P.H., and Paul C. McCormick, M.D., M.P.H.

Department of Neurological Surgery, Columbia University College of Physicians and Surgeons, New York, New York

The routine practice of neurosurgery generates a large amount of clinical data. The structured, systematic capture of this information using clinical registries or other rigorously designed observational studies can yield useful evidence to help improve the care of patients. Registries in particular can be designed to measure outcomes in real-world clinical settings and to study differences in outcomes between subgroups. This information can help clinicians to advise patients regarding their treatment options.

To provide valid, generalizable evidence, however, registries and other observational studies must be designed and conducted with a rigor similar to that of randomized clinical trials. Neurosurgeons with a basic understanding of the potential advantages and pitfalls of nonrandomized trials and the methods of statistical analysis will be able to assess the quality of clinical data and to incorporate the findings appropriately into their patients' care.
(http://thejns.org/doi/abs/10.3171/2012.10.FOCUS12299)

Key Words • **observational study** • **study design** • **clinical registry**

Despite decades of laboratory and clinical research and the development of a multitude of evidence-based guidelines, neurosurgeons still face patients every day with uncertainty regarding the best treatment for them. Additionally, although they strive to provide the best care for each patient, neurosurgeons also face a lack of ways to measure the quality of the care they provide objectively and in real time. Practitioners need tools to help them provide their patients with accurate, up-to-date information to use in shared decision making, and to allow them to assess the quality of their care compared with benchmark data and within their own practice over time. A recent Institute of Medicine report emphasized the need to "produce and deliver practical evidence that clinicians and patients can apply to clinical questions."[7]

The clinical practice of medicine generates an enormous amount of data. These data, recorded and analyzed properly, can help practitioners treat individual patients, medical practices optimize their systems of care, and national organizations and policy makers improve overall health by developing tools to assist in shared medical decision making. Measuring clinical activities should occur in real time, providing the target audiences with up-to-date, clinically relevant information, and should do so without interfering with the physician-patient relationship or requiring an excessive commitment of resources.

Achieving these goals requires a clear delineation of the purpose of the project, an effective design, and the use of appropriate statistical analyses and interpretation. It also requires a fundamental understanding of study design principles and the potential pitfalls of observational studies.

Registries are a specific type of observational study. They are formal databases of a specific group of patients and are potentially powerful tools for examining many aspects of health, including the treatment of rare diseases, safety of medical devices, and outcomes of interventions in situations in which RCTs are not appropriate or feasible. A well-planned and -executed registry can help to assess quality of care and outcomes, and to determine positive and negative risk factors for response to treatment. This information can often help the practitioner answer the clinically salient question "[i]n whom does this treatment work?" In contrast, RCTs are best designed to address the question "[w]hich intervention provides the greater average benefit when delivered to a select group of patients?" The difference is between an individual clinical decision, as faced by patients and their physicians daily, and a public health or policy recommendation.

There are many publications that explain the technical aspects of the design and execution of observational studies including registries. The goal of this paper is not to provide a how-to guide, but rather it is to discuss briefly some of the key conceptual issues in the design, conduct, and interpretation of the study of the practice of clini-

Abbreviations used in this paper: H_0 = null hypothesis; RCT = randomized clinical trial.

cal neurosurgery. Data accrual done haphazardly or with poor methodology is unlikely to provide high-quality evidence or help refine patient care. With a basic understanding of study methodology, neurosurgeons will be informed contributors and consumers of data generated by their clinical activities.

Defining the Purpose

The first step in designing a study to measure clinical practice data is to determine the purpose for doing so. The proposed questions will, in large part, determine the appropriate design of the project. Direct comparisons of medical interventions and some surgical techniques may best be performed with a clinical trial. The causes of rare diseases or complications are often investigated with case-control studies. The natural history, outcomes, or complications of relatively common conditions and their management, however, are often best studied with well-designed prospective cohort studies or registries. According to the Institute of Medicine's report, "[t]he issue is not determining which research method is best for a particular condition, but which method provides the information most appropriate to a particular clinical need."[7]

The purpose of the study and its intended audience must be considered in determining what data are to be collected for each patient. To maximize complete data collection, information unnecessary for the study's purpose should not be solicited. Prudent selection of the database components can also improve the validity of the study. Minimizing bias requires, in part, as complete enrollment as possible of eligible patients, and obtaining all data on those who are enrolled. As the burden of form completion or data entry increases, the likelihood of patient or investigator compliance decreases. It is therefore important to avoid the temptation to try to capture an excessively large amount of data, either by including many types of information or assessment scales, or by including frequent assessment time points. A clear delineation of the purpose of the study will help to determine what information should be obtained.

Why Observational Studies?

Although the methodological strengths of the RCT design are indisputable, an RCT may be unnecessary, impractical, or unethical in certain circumstances. Some of the situations in which an observational design, such as a registry, may be the best option include studies to assess outcomes and quality of care as delivered, studies that require a large number of patients, or investigations of conditions in which patient preferences largely determine management and compliance with random treatment assignment is likely to be suboptimal.[1]

High-quality observational studies can be conducted with minimal disruption of normal clinical practice. The patient-doctor interactions remain intact, and only the information gathered or the format in which it is collected may change. Clinical recommendations and decision making occur as usual, and therefore the study provides information about real-world practice outcomes, complications, and quality. Furthermore, the inclusion criteria may be set to include a wide spectrum of patients with other comorbidities and other factors that may influence outcome. This, as discussed below, allows the investigator to study how outcomes interact with a variety of patient factors. Finally, because observational studies are based on actual treatment, they help clinicians assess current management, and may allow insight into the effects of changing treatment strategy.

Registries are a subtype of observational study that are generally used to obtain data about all patients or a representative sample of patients with a specific diagnosis or who were exposed to particular conditions, including natural phenomena or specific medical treatments or devices. A structured data collection system is developed and maintained for the lifetime of the registry. Registries can provide information regarding long-term consequences of exposures to medicine or environmental factors, durability data for medical devices, treatment outcomes for rare disorders, and quality data for patient care, among others. Because they are purpose-built and observational, they are ideal for capturing relevant patient and treatment factors that may be associated with outcome. This can help investigators model and understand observed differences in treatment effect, and it can help clinicians provide patients with recommendations based on their individual characteristics.

Strengths of Observational Studies

Approach to Heterogeneity. Differences between patients may be viewed in different ways. In RCTs, selection criteria are often developed to enroll a fairly homogeneous study population and minimize confounding and bias. This may affect the generalizability of these studies, as discussed above. Furthermore, the randomization process itself is designed to distribute measured and unmeasured characteristics between the experimental and control cohorts in a way that results in a balance of positive and negative factors. Variables that are known to affect outcome may be accounted for in the randomization process to ensure an equal distribution through stratification.

Patient heterogeneity is, however, a daily reality in clinical practice. When faced with an individual patient, his or her individual characteristics are not factors to be balanced or controlled for, but they are features that may influence the success or failure of an intervention. Whereas RCTs are not generally designed to explore more than a few subgroups of patients for differences in outcome, large-scale registries are ideally suited to do just that, with the accrual of a large number of patients and clinically relevant variables. Statistical models can be constructed to test the interactions of patient characteristics with outcome, both singly and in combinations. This information can help the practitioner understand which patients are the best candidates for a given treatment, and it can help patients understand their individual risks and potential benefits.

Contemporaneity. Of the several limitations of traditional studies and trials, one of the least tractable is that large-scale investigations take years to design, conduct, analyze, and report. Frequently, by the time the results of RCTs are available, clinical practice has evolved and the techniques or treatments subject to investigation no longer represent the state of the art. Retrospective cohort studies are by design backward looking, and therefore also do not provide practitioners with real-time feedback about patient outcomes and quality of care.

In addition to the practical issues of RCTs and the design of retrospective studies, the statistical methods used in most health care investigations also require completion of a study prior to performing the analysis. Null hypothesis statistical testing, based primarily on the work of Ronald A. Fisher, is a logically rigorous deductive statistical analysis.[3]

Pitfalls of Observational Studies

Potential for Bias. Observational or nonexperimental studies are often dismissed as providing low-quality evidence of the effectiveness of a treatment or intervention. Much of this criticism is valid, and is based on the significant opportunity for bias to affect the results of the study. Bias is the presence of a systematic error in a study that can lead to faulty conclusions. No study design is completely immune to bias, but several of the design elements common to most high-quality RCTs can reduce some of the most common forms of the problem. Careful planning and execution of observational studies can reduce the potential sources of bias.

Selection bias is present when patients are included or excluded from a study based on factors that may also be associated with their outcomes.[6] A rigorous study, whether observational or experimental, can minimize the possibility of selection bias by prespecifying inclusion and exclusion criteria and carefully tracking all patients from assessment of eligibility through enrollment or refusal to participate.

Patients should then be accounted for at all prespecified follow-up time points to avoid follow-up bias. This can occur when patients drop out and their data are not recorded. Patients who are lost to follow-up are very likely to differ from those who remain in the study, and a high proportion of dropouts can undermine the validity of a study. Flowcharts are often used in publications of studies to succinctly demonstrate the recruitment and follow-up of patients and to allow readers to assess the possibility of selection bias or differential loss to follow-up.

A variation of selection bias can affect the results of comparative observational studies. Patients may be selected for one treatment over another based on criteria that may make them more or less likely to have a good outcome. This selection process may or may not be conscious. If the goal of an observational study is to compare outcomes between treatments, special methods such as propensity score matching or a case-control study design may be necessary to address this concern.

In registries designed primarily to assess quality of care, either all eligible patients or a sufficient representative sample should be included. Clearly defined inclusion criteria and enrollment processes can minimize the chance of selection bias, intentional or not. Audits of participating centers may be used to confirm the enrollment of all eligible patients. Studies such as those published by the Spinal Deformity Study Group, for example, provided important information but were seriously limited because no information was provided about any patients who were eligible for enrollment but did not participate. In other words, it is not possible to determine, without this information, if a sample is representative of a population, or the nature of that underlying population.

Generalizability. Most studies are undertaken, at least in part, to develop knowledge that can benefit patients outside of the study. The applicability of a study's results to a broader population is its generalizability. To assess the generalizability of a study, one must be able to determine the population from which the study sample was drawn. Some of the same techniques used to minimize selection bias can also improve the generalizability of a study. A study that includes all patients at several centers with a particular diagnosis is likely to be more generalizable than one that enrolls haphazardly based on convenience or cooperation at a single center.

It still can be difficult, however, to determine if a study is generalizable to an entirely new setting. For example, urban academic medical centers are likely to draw from a different population than a suburban private practice. The racial, socioeconomic, or educational makeup of the study population may also raise concerns about its applicability to other groups. In general, the broader the variety of settings and inclusion criteria, the more broadly generalizable the results are likely to be.

Inferences and Analysis

Inferences are the conclusions that one can reach based on the evidence provided by a study. The design, execution, and analysis of a study determine, in large part, the inferences that may be supported by it. Inferences regarding outcome of treatment, complications of procedures, or natural history of disease may all be of interest to practitioners, and supported, with varying degrees of strength, by observational studies.

Physicians are often most concerned with the singular predictive inference: an estimate of the outcome of a treatment for an individual patient. Although some aspects of patient outcome may appear to be due to random chance, the scientific practice of medicine is predicated on the assumption that particular factors, some within the control of physicians and some not, affect a patient's outcome. The identification, quantification, and incorporation of these factors into medical decision making is the goal of much medical research, and observational studies can provide important evidence. An understanding of the statistical methods used to analyze studies is necessary to understand the appropriate inferences that may be made from a study. The statistical design of an observational study or registry has important consequences for the type of inferences that can be made from it and for the way new information is incorporated with previous knowledge.

Frequentist Analysis. The most common method of statistical analysis used in medical studies, involving H_0 testing, is based on the work of Fisher.[5] This is a deductive method, assessing the particular based on the general, and is therefore logically rigorous. Conceptually, it involves the construction of a H_0 that posits, for example, that a treatment has no effect on quality of life. A study is conducted and the results are analyzed. Statistical analysis is used to determine the probability (called the p value) of obtaining the observed results, or results more extreme, under the H_0 if the experiment were repeated. A p value < 0.05, therefore, is the probability of obtaining results as different as or more different than the null, assuming that H_0 is true.

Two points warrant emphasis. First, the underlying assumption is that H_0 is true. It is rare that mean outcomes between time points, treatments, or cohorts will differ by any prespecified value, including zero. Second, this is a backward-looking analysis. The probability space must be well defined, and so the study size and parameters must be prespecified and the analysis performed only after all the data have been accrued. This is because the probability in question is the probability of obtaining the actual results out of all the other possible results of conducting the study.[4]

Although this method has the strength of logic behind it (is never proven—it is only rejected or not rejected), it neither provides results that are intuitive nor applies particularly well to the way medicine is actually practiced. Physicians and patients generally understand direct probabilities; the explanation of p values above demonstrates that frequentist analysis, as this method is often called, does not provide this information. Confidence intervals are even more counterintuitive and difficult to understand; they represent not a range of values that contains the parameter, but a probability that similarly calculated intervals from multiple repetitions of an experiment will include the parameter of interest.

Bayesian Analysis. An alternative method of analysis, growing in popularity in part because of the problems with the frequentist techniques noted above, is based on inductive logic and Bayes' theorem. As detailed by Berry,[2] Bayesian analysis allows direct calculations of probability, yielding intuitive, useful results. It also avoids the backward-looking aspect of frequentist techniques that requires a study to be defined in advance so that the data can be analyzed as coming from one of many theoretical repetitions of the experiment. One cost of these benefits is that, because the logical underpinnings of induction are weaker than those of deductive reasoning, Bayesian analysis does not have the philosophical rigor of frequentist analysis. A second potential drawback, although one that seems to be fading in importance, is that Bayesian analysis requires specifying a prior probability, an estimation of the likelihood of the outcome based on previous studies or experience. These so-called priors have been criticized as being subjective, and they can be difficult to elicit. They represent the investigator's expectation of the outcome and are strongest when they are based on previous studies. In fact, the process of Bayesian analysis closely parallels in many ways the method a clinician uses to interpret a new study's results. Based on previous studies and individual experience, the clinician has a base of knowledge; this corresponds to the prior probability. Upon reading and analyzing the new study, the clinician assesses the strength of the study; the Bayes factor represents the strength of the new data. Finally, the clinician adjusts his or her knowledge in light of the new data to develop a posterior probability. If the data are compellingly strong and different from the clinician's previous knowledge, the posterior may be dissimilar to the prior. If, on the other hand, the new study produces results similar to the prior, then the strength of belief increases but its nature remains unchanged.

Once the prior probability distribution has been specified, Bayesian analysis allows the practitioner to update the parameter estimates in real time, without the restraints of frequentist methods. Furthermore, one can calculate direct probability intervals that are intuitively understood. Called "credible intervals" to avoid confusion with frequentist confidence intervals, they represent the probability area centered on the most likely value of the parameter in question. For a summary parameter estimate, for example, the 95% credible interval represents the range of values with a 95% chance of containing the true value, with the most likely value in the center of the range.

Bayesian analysis, therefore, proceeds much in the same way as clinical medicine does: previous information is used to estimate likely outcomes, additional data are obtained through clinical practice, and that evidence is used to update the original judgments. Regression models, both logistic and linear, can also be constructed and analyzed using Bayesian techniques. In this way, factors associated with good and poor outcomes can be identified and their interactions analyzed. Predictive models can be developed using patient data and then validated prospectively, giving physicians the information they need to help patients make fully informed decisions regarding their medical care.

Conclusions

The clinical practice of neurosurgery generates an enormous amount of data daily. With properly designed and conducted registries and other observational studies, these data can be evidence to help physicians, medical practices, and organizations develop the tools to improve patient care at an individual level. To maximize the validity and usefulness of observational studies, however, their potential strengths and weaknesses must be understood. Additionally, clinicians should understand the limitations of common statistical techniques and the benefits of alternative methods, such as Bayesian analysis.

Properly designed and conducted clinical registries can provide clinicians with a granularity of information unavailable from other study designs. They can do this because a large number of patients can be economically enrolled, and relevant clinical data can be obtained about each patient. With attention to the details of study methodology, the result can be high-quality, valid evidence of the quality and effectiveness of our patient care.

Disclosure

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author contributions to the study and manuscript preparation include the following. Conception and design: both authors. Drafting the article: Angevine. Critically revising the article: McCormick. Reviewed submitted version of manuscript: both authors. Approved the final version of the manuscript on behalf of both authors: Angevine.

References

1. Angevine PD, McCormick PC: Inference and validity in the SPORT herniated lumbar disc randomized clinical trial. **Spine J 7:**387–391, 2007
2. Berry DA: **Statistics: A Bayesian Perspective.** Belmont, CA: Duxbury Press, 1996
3. Fisher RA: **The Design of Experiments, ed 9.** New York: Macmillan, 1971
4. Goodman SN: Toward evidence-based medical statistics. 1: The P value fallacy. **Ann Intern Med 130:**995–1004, 1999
5. Kruschke JK: **Doing Bayesian Data Analysis: A Tutorial with R and BUGS.** Burlington, MA: Academic Press, 2011
6. Shadish WR, Cook TD, Campbell DT: **Experimental and Quasi-Experimental Designs for Generalized Causal Inference, ed 2.** Boston: Houghton Mifflin, 2002
7. Smith M, Saunders R Jr, Stuckhardt L, McGinnis JM (eds): **Best Care at Lower Cost: The Path to Continuously Learning Health Care in America.** Washington, DC: National Academy Press, 2012

Manuscript submitted September 14, 2012.

Accepted October 30, 2012.

Please include this information when citing this paper: DOI: 10.3171/2012.10.FOCUS12299.

Address correspondence to: Peter D. Angevine, M.D., Department of Neurological Surgery, Columbia University College of Physicians and Surgeons, 710 West 168th Street, Room 510, New York, New York 10032. email: pda9@columbia.edu.