How the science of practice will improve evidence-based care

Robert E. Harbaugh, MD

Department of Neurosurgery, Penn State University, Milton S. Hershey Medical Center, Hershey, Pennsylvania

This review article analyzes the present evidence-based medicine (EBM) algorithm, compares it to the science of practice (SOP) algorithm, and demonstrates how the SOP can evolve from a quality assurance and quality improvement tool into a clinical research tool. Using appropriately constructed prospective observational databases (PODs), the SOP algorithm can be used to draw causal inferences from nonrandomized data, perform innovative comparative effectiveness research, and generate reliable information that can be used to guide treatment decisions.

<https://thejns.org/doi/abs/10.3171/2020.2.FOCUS202>

KEYWORDS comparative effectiveness research; evidence-based medicine; propensity score matching; randomized controlled trials; science of practice

IN 1938 Harvey Cushing and Louise Eisenhardt published *Meningiomas, Their Classification, Regional Behavior, Life History, and Surgical End Results*.² It demonstrated the systematic collection and analysis of patient characteristics, processes of care, and clinical outcomes to improve the quality of neurosurgical care. These are the basic tenets of the science of practice (SOP) algorithm. In 2020, we collect data in digital databases and analyze it with sophisticated biostatistics, but the essential aspects of the SOP are the same.

This SOP approach allowed our field to advance dramatically over the last 100 years, demonstrating the limited role of randomized controlled trials (RCTs) in establishing the effectiveness of procedural interventions.¹⁶ However, in our present evidence-based medicine (EBM) algorithm, non-RCT studies, including registries, are viewed disparagingly. The EBM algorithm created a hierarchy of clinical data, with data from RCTs sitting atop the hierarchy.³ However, despite decades of investment in EBM and the proliferation of evidence-based guidelines, neurosurgeons have not been able to agree on the best treatment for many patients. We need additional tools to help guide our treatment decisions.

Analyzing the EBM Algorithm for Neurosurgical Clinical Research

In the present system we use clinical research published

in peer-reviewed journals as the source of evidence, accept a standard hierarchy of strength of evidence based on trial design, and develop practice parameters from an analysis of the evidence.³ The acceptance of surgical RCTs as a gold standard requires analysis.

Problems With Surgical RCTs: Bias, Confounding, and Chance

The sources of error in all clinical studies are bias, confounding, and chance. Bias is a systematic error, introduced consciously or unconsciously by observers or patients, that leads to incorrect estimates of treatment effects. Confounders are variables related to treatment and outcome that suggest a false causal relationship. Chance errors occur because of sampling problems, which decrease as sample size increases. The RCT has been accepted as a gold standard in clinical research because the ideal RCT addresses errors from bias, confounding, and chance.

There are 3 essential components to an ideal RCT: 1) concurrent comparisons to eliminate temporal bias, 2) masked adjudication of clear and clinically meaningful endpoints to eliminate physician and patient bias, and 3) randomization of an adequately sized, representative population to equally distribute confounding variables and reduce the likelihood of chance errors (Table 1). The ideal RCT is a true gold standard. Unfortunately, surgical RCTs rarely approximate this ideal.

ABBREVIATIONS EBM = evidence-based medicine; ISUIA = International Study of Unruptured Intracranial Aneurysms; POD = prospective observational database; RCT = randomized controlled trial; SOP = science of practice; UIA = unruptured intracranial aneurysm.

SUBMITTED January 1, 2020. **ACCEPTED** February 6, 2020.

INCLUDE WHEN CITING DOI: 10.3171/2020.2.FOCUS202.

Problems With Surgical RCTs: Null Hypothesis Testing, p Values, and Power Calculations

An RCT, like all clinical studies that employ null hypothesis testing, is performed to determine the presence or absence of a treatment effect. Before beginning the trial, the null hypothesis—no statistically significant difference between treatments—is accepted. In a positive study, the null hypothesis is rejected. If the null hypothesis cannot be rejected, the study concludes that no difference between treatments exists. For positive trials, the chance that a difference was seen, even though the null hypothesis is true, is represented by the p value. A p value of less than 0.05 tells us there is less than a 5% chance that results as different as those observed in the study occurred by chance alone. Although commonly misused as a substitute for clinical benefit, the p value tells you nothing about the magnitude of the treatment effect or whether or not there is any clinical benefit. A study may be reported as a “positive” study whenever a statistically significant difference is observed—even if the observed difference is less than the prospectively identified value judged to be clinically meaningful for the study’s power calculation.

For negative studies, the likelihood of detecting a positive result if there is a real therapeutic difference between treatments—the power of the study—is important. Stated simplistically, a power of 0.80 means that there is an 80% chance of finding a difference of a magnitude deemed to be of clinical importance, if such a difference exists. The power of a study is dependent on sample size, the magnitude of the treatment effect, and the statistical tests employed. Without knowing the power of the study, negative studies add very little to our knowledge.

Problems With Surgical RCTs: Crossovers

Surgical trials differ from the ideal RCT example in several important ways. In many surgical trials, patients may elect to cross over from one treatment arm to another. To preserve the benefits of randomization, it is necessary to analyze patients in their assigned groups, even if they cross over to another treatment arm—an intention-to-treat analysis. For example, in a trial comparing surgical to nonsurgical treatment of back pain, the patient randomized to medical treatment may be unsuccessfully treated medically, choose to have surgery, and then do well. In an intention-to-treat analysis, this good outcome following surgery would be assigned to the medical treatment arm (the patient’s initial treatment assignment). This is necessary to preserve the value of randomization. Accepting the as-treated analysis voids the protections of randomization. There are good examples where the as-treated analysis differs substantially from the intention-to-treat analysis.¹⁰ We know the as-treated analysis represents what really occurred during the study but, because it is not a randomized study, it is given far less weight in the EBM hierarchy than the intention-to-treat analysis.

Problems With Surgical RCTs: Unmasked Trials

Most surgical trials are unmasked trials, and endpoints

are often subjective. This greatly increases the risk of bias affecting the results of the trial. Patients may experience a substantial placebo effect with surgery, and unmasked investigators may harbor a surgical or nonsurgical bias. Having someone other than the operating surgeon evaluate patients postoperatively does not solve this problem; rather, it substitutes another unmasked observer’s bias for the surgeon’s bias. One reason that the ideal RCT is viewed as a gold standard is that a double-masked trial eliminates patient and investigator bias. However, an unmasked trial is a biased trial and thus most surgical RCTs are not ideal.

Problems With Surgical RCTs: Equipoise

It is often difficult in surgical trials to choose a representative patient population because of problems with equipoise. Surgeons have an implicit contract with their patients to offer the best care available. If surgeons do not believe that the treatment arms of an RCT (e.g., medical management and surgical management) are equally efficacious, they will offer surgical treatment, outside the trial, to patients they believe are most likely to benefit.

Consider the following proposed RCT comparing observation, clipping, and coiling of unruptured intracranial aneurysms (UIAs) 3–12 mm in greatest dimension in patients 18–75 years of age. It is assumed that the 3 treatment options are equal in regard to one or more predetermined endpoints, such as aneurysm rupture, death, or functional health status. We now begin enrolling patients in the trial and will consider the following two candidates. Patient 1 is a 25-year-old woman, a cigarette smoker with a history of hypertension and a family history of aneurysm rupture, who harbors an 11-mm, irregular, basilar apex aneurysm. Patient 2 is a 64-year-old man, a nonsmoker, without hypertension, who harbors a 4-mm, smooth, ophthalmic artery aneurysm and has no family history of aneurysm rupture. Although both meet the inclusion and exclusion criteria for the study, it is much more likely that invasive treatment, outside the trial, will be offered to patient 1 because the treating physician believes that the risk of aneurysm rupture is higher.

The RCT does not address the problem of confounding factors; it moves them to the patient-selection phase of the study. Treating physicians will consider these factors, patient preference, and other confounding variables before making a recommendation for management. Those patients believed to be at the greatest risk of having their aneurysm rupture may be encouraged to undergo treatment outside the trial. If equipoise does not exist in the treating physicians and they are correct in their judgments, a patient population that is at particularly low risk of rupture will be selected for the study.

These patients are now randomized to observation, surgery, or endovascular treatment. At follow-up, assessment of function will be influenced by patient and observer bias because the study cannot be masked. If, at a predetermined point in time, we find that patients in the observation group are faring significantly better, then the study will report observation as the preferred treatment for UIAs. Now the 25-year-old woman may no longer have the option of aneurysm treatment because we have RCT-based evidence that observation is the superior option.

Problems With Surgical RCTs: Defining Equipoise

There are several definitions of equipoise that have different implications for clinical research.^{5,9} Trialists often consider equipoise to be a lack of consensus in the literature about choice of treatment. However, in regard to the performance of RCTs, it is the equipoise of those making treatment decisions—providers, referring physicians, patients, and family members—that is important. These versions of equipoise differ, and each has implications for the reliability of clinical research findings.

Any clinical study comparing treatments requires the existence of sufficient equipoise to allow physicians to participate or refer patients.^{5,9} A representative and generalizable RCT can only be carried out if provider and patient equipoise exist. It should also be noted that the timing of an RCT has a powerful influence on the feasibility and reliability of the RCT.¹⁶ If an intervention has not been adequately refined, early trials may not show benefit, whereas trials with a more fully developed technology could yield very different results.¹⁶ Conversely, if trials are delayed and the procedure is allowed to disseminate widely, patients and surgeons may view it as a standard of care, making an RCT unfeasible.

In nonrandomized studies such as a prospective observational database (POD), individual clinician equipoise is not an issue because all patients are included in the study. If there is already a consensus in the neurosurgical community, a POD will demonstrate this. The advantage of the POD in this case is that the lack of equipoise is obvious, and the factors that determine this can be analyzed. The lack of equipoise in treating physicians and patients in an RCT may be hidden because of prerandomization patient selection.

Problems With Surgical RCTs: Surgeon Selection

Surgical RCT results may also be unrepresentative because of surgeon selection. A study showing a benefit from surgery with a highly selected group of surgeons will not be applicable to the general population if the outcomes of most surgeons fail to match the outcomes of those in the study. In surgical RCTs, surgical skill and judgment are constants. In the real world, and in the SOP algorithm, we know that surgical expertise is a variable that affects patient outcomes.

Other Problems With Surgical RCTs: A Tarnished Gold Standard

Perhaps the biggest problem is the deference we give to the RCT even if it is flawed for surgical trials. Unmasked surgical RCTs do not eliminate either patient or observer bias. As for confounding variables, unless equipoise exists in regard to management assignment for the surgeons involved in the trial, the confounders will not be randomly distributed. The surgeons and patients involved in the trial determine who enrolls. Without equipoise in those who determine enrollment, we are left with data that do not represent the population at large.

Problems With Surgical RCTs: Why We Cannot Reach Consensus

Many factors affect the design of clinical trials created to determine the efficacy and effectiveness of care.^{4,14,15} Traditionally, efficacy studies such as RCTs determine the effect of treatment under carefully controlled circumstances, while effectiveness trials such as those using PODs are meant to determine the benefit of treatment in a real-world setting. The RCT, with highly selected patients and surgeons and prescribed interventions, will be internally rigorous but may not be generalizable to the target population.¹⁵ Conversely, accepting all patients, providers, and treatment interventions in a POD makes the results more generalizable but also makes analysis more difficult. Although there are clear differences between efficacy and effectiveness studies, it should be noted that PODs and RCTs evaluating the same clinical question usually concur on treatment benefit.^{1,7,8}

We perform clinical research to change the behavior of physicians and patients in ways that improve the quality of our care. In order to change behavior, physicians and patients need to have confidence in the data. Reliable data must be both precise and accurate. Precision is based on how close a repeat measurement will be to a previous value. We use frequentist statistics such as p values and confidence intervals to determine the precision of our data. Accuracy, however, is determined by how closely the data in a study reflect the real-world value, which is a challenging task when the real-world value is not known. An important variable in regard to accuracy is how confident are physicians and patients that a clinical study represents a real-world situation. If the constraints of an RCT do not reflect the real-world environment, there will be little confidence in the results of the RCT.

Problems With Surgical RCTs: Summary

For all the reasons noted above, many surgical RCTs are not a gold standard. They are often unmasked studies, subject to observer and patient bias, involving unrepresentative patients, treated by unrepresentative surgeons, who may not have equipoise in regard to the treatments offered. This may yield studies showing statistically significant differences that have no clinical benefit, meaningful associations that are missed because of underpowered studies, and studies that do not accurately represent the patients and physicians in the community.

An Alternative to RCTs: The Science of Practice

Research methodologies apart from traditional RCTs, including randomized registry trials, platform trials, and adaptive designs, represent an alternative.¹⁶ The validity of these nonrandomized studies may be enhanced by designing them to mimic a corresponding RCT.^{12,13,17} An SOP approach may complement the present EBM algorithm as a means of generating reliable clinical data. PODs that collect a standard data set of patient characteristics, treatment processes, and outcomes have allowed us to do important quality improvement work. The SOP may be superior to

TABLE 1. The ideal RCT

Essential Component	Rationale
Concurrent comparison	Eliminates temporal bias (intention-to-treat analysis)
Masked adjudication of unambiguous and clinically meaningful endpoints	Eliminates patient and investigator bias (the “double-blind” study) and ensures that we are assessing clinically meaningful data
Randomization within an adequately sized and representative sample	Equally distributes confounding variables and reduces the risk of sampling error in patients who are representative of the target population

EBM for clinical research and quality improvement in neurosurgery because the clinical research and quality improvement are more generalizable to the target population. The routine collection, analysis, and distribution of clinical data, inseparable from practice, is the basis for the SOP. However, refinements to the SOP algorithm may allow us to perform comparative effectiveness research of high reliability without randomization.

The Evolution of the SOP: SOP and Big Data Studies

Whether the EBM or the SOP algorithm is used, reliable clinical studies must be methodologically sound and statistically rigorous if we hope to obtain data that can guide our treatment. The SOP should not be confused with “big data” studies. It is now simple to manipulate huge data sets and perform millions of comparisons. This creates a great risk of selective reporting of correlations without causal relationships.¹¹ A vital component of scientific rigor is the need to explicitly disclose prespecified endpoints of a study. For the SOP to evolve to the point where we can draw causal inferences from nonrandomized data, we need to create a POD that is the equivalent to the most rigorous clinical trials.

The Evolution of the SOP: Advantages of PODs

PODs have the benefit of being able to enroll large numbers of patients from real-world clinical practice.¹⁴ PODs are designed without a designated stopping point for data collection, while RCTs are intermittent with a definite start and stop date. This is important when rapidly changing technology may have an influence on treatment outcomes. RCTs cannot address the effectiveness of procedures done by surgeons of varying skill in diverse practice settings that may not have been represented in the RCT. A comparison of the EBM model and the SOP model is presented in Table 2.

The Evolution of the SOP: Drawing Causal Inferences From PODS

The well-designed POD differs from the well-designed RCT only in regard to randomization. The purpose of randomization is to equally distribute confounders among treatment groups. If we can address confounding variables in a POD, we can use POD data to draw causal inferences

without a randomized study. Although other options exist for drawing causal inferences from nonrandomized data,⁶ I will focus the discussion on propensity score matching.^{12,13}

To draw causal inferences from nonrandomized POD data we need to construct our PODs with attention to factors used to determine treatment assignment. If this is done, propensity score matching can be accomplished, and we can compare patients with the same distribution of covariates who differ only in regard to treatment assignment. This equal distribution of confounders allows propensity score–matched PODs to mimic RCTs.^{12,13}

The Evolution of the SOP: The Rubin Causal Model

A POD designed for propensity score matching relies on the framework of the Rubin Causal Model.^{12,13} RCTs randomize treatment assignment to distribute confounders equally among groups. This generates cohorts of patients who differ only in regard to treatment. In this situation, outcomes following various treatments can be directly compared. To derive a statistically meaningful causal inference from PODs, we must mimic the RCT and find a way other than randomization to generate cohorts of patients who differ only in regard to treatment. The propensity score is an estimate of the probability of treatment assignment given multiple variables used by decision-makers to assign treatment. Generating a valid propensity score requires expert insight into the covariates that influence treatment assignment.^{12,13}

Propensity score matching is a way to determine treatment effects in nonrandomized trials by controlling for the existence of known confounding factors prior to outcome analysis. This is much different from analyzing patient cohorts created due to the personal treatment biases of patients and physicians or applying multivariable analysis of the data after the outcomes are known. Comparing the outcomes of patients who are as similar as possible on a wide range of potential confounders allows us to perform a pseudorandomized study and draw causal inferences using a POD. The extent to which propensity score matching can eliminate bias from confounders depends on the completeness and quality of the variables on which the propensity score is based.

Propensity score matching will not eliminate bias introduced by unknown confounding variables.¹⁷ For example, assume the patients’ astrological signs have a profound

TABLE 2. A comparison of EBM and SOP approaches to clinical research

EBM Approach	SOP Approach
RCTs, cohort studies, case-control studies	PODs
Discontinuous or punctuated process	Iterative process
Surgical skill is assumed to be uniform	Surgical skill may be a variable
Prescriptive and proscriptive	Encourages innovation
All or none; there is or there is not rejection of the null hypothesis	Nuanced: treatment effectiveness varies depending on patient-specific variables

influence on outcome. Because this is not expected, this will not be one of the data elements used in creating the propensity score. This unknown confounder could distort or invalidate the causal inferences we make.

The Evolution of SOP: Designing a POD With Propensity Score Matching

To design a POD that maximizes the effect of propensity score matching, we need to ask and answer the following questions: 1) What randomized experiment do we want to model? 2) Who are the decision-makers for treatment assignment? 3) What are the key covariates used to assign treatment? 4) Can we measure the key covariates well? 5) What clinically meaningful outcomes will we measure? 6) What sample sizes will be needed? In addition, as for a well-designed RCT, a well-designed POD must prospectively specify outcomes and analytical methods to be used without resorting to the actual outcome data.

Improved accuracy of causal inferences using propensity score matching has been demonstrated using the International Study of Unruptured Intracranial Aneurysms (ISUIA).¹⁴ In ISUIA, the risk of rupture of a previously unruptured aneurysm depends on many factors, including aneurysm location, size, and other geometrical morphological factors. Patient outcomes were correlated with patient age, aneurysm location, and aneurysm size. The ISUIA was not an RCT. Patients were selected for clipping, coiling, or medical management with surveillance based on the providers' recommendations. Therefore, to compare outcomes of various management options, propensity scores of the aggregate data were used. With propensity score matching, invasive therapy was statistically significantly superior to medical management with surveillance for both hemorrhage and overall outcome at 5 and 10 years.¹⁴

The Evolution of the SOP: A POD for the Management of UIAs

We previously considered an RCT comparing observation, clipping, and coiling of UIAs. We will now consider a POD designed for propensity score matching. We want to mimic an RCT evaluating clipping, coiling, and observation for patients aged 18–75 years with UIAs 3–12 mm in size. The decision-makers for treatment assignment are the patient, the patient's family members, and the treating physician. Our database must capture key covariates used to assign treatment. These include patient-specific vari-

ables like age and medical comorbidities, aneurysm-specific variables like size, shape, and location, and provider-specific variables like specialty training and practice site. We must collect data on a large number of such variables in order to refine our propensity score match.

We also need to prospectively determine what clinically meaningful outcomes we will collect and at what time points these will be collected. We must also calculate what sample size will be needed for an adequately powered study. If we can successfully complete all of these steps, we should be able to construct a POD that will allow us to draw causal inferences about the comparative effectiveness of our management options.

We now begin enrolling patients. We no longer have to worry about whether the treating physician or surgeon will be equally likely to recommend enrollment in a trial because all patients are entered into the POD. We may find that across numerous practice sites endovascular treatment is appropriate for young people with larger basilar apex aneurysms and that observation with imaging surveillance is the preferred management option for older patients with small ophthalmic artery aneurysms. If so, we will not be able to generate matched propensity scores for cohorts of patients who differ only in regard to treatment choice as equipoise does not exist for the various treatment options for these patients. However, we will now know which factors determine management assignment across numerous practices and will not make the mistake of applying the outcomes of a subset of patients to all patients as could occur in an RCT.

It is likely that we will be able to obtain good propensity score–matched cohorts for a wide range of patients. Analysis of the POD data would allow us to draw causal inferences regarding the comparative effectiveness of the various management options for these patients. The POD will allow increasingly refined analysis over time as more patients are entered and followed for a longer period of time. A comparison of the strengths and weaknesses of an ideal RCT, surgical RCT, and a POD designed for propensity score matching is given in Table 3.

Summary: How the SOP Will Improve Evidence-Based Care

While RCTs have high internal validity, there are concerns about the generalizability of such studies. Neurosurgeons are skeptical of the applicability of RCTs to heterogeneous patient populations and practice settings. The SOP algorithm provides an alternative for performing

TABLE 3. Comparing an ideal RCT, surgical RCT, and POD with propensity score matching

Sources of Error	Ideal RCT	Surgical RCT	POD*
Observer bias (masked study)	Yes	No	No
Temporal bias (prospective data)	Yes	Yes	Yes
Known confounders	Yes	Yes	Yes
Unknown confounders	Yes	Yes	No
Representative of the target population	Yes	No†	Yes
Adequately powered	Yes	Maybe‡	Yes‡

* POD designed for propensity score matching.

† Lack of equipoise and highly selected patients and surgeons in many surgical trials.

‡ PODs include all patients, so it is easier to ensure patient cohorts large enough for an adequately powered study.

clinical research that addresses many of the shortcomings of surgical RCTs. The SOP, a POD-based approach to clinical research, will allow us to compare the effectiveness of neurosurgical interventions.

Continued, exclusive reliance on RCTs in surgical specialties may be harmful, as we will have few reliable data points to guide care.¹⁶ While there are methods to improve the feasibility of conducting neurosurgical RCTs, pragmatism is required to fairly assess the relative strengths and weaknesses of RCTs versus a POD approach. Given the difficulty of conducting RCTs that yield reliable data in surgical disciplines, it is likely that there will continue to be limited RCT data to analyze many procedural interventions.¹⁶

Improving information technology will result in clinical research based on PODs becoming a routine part of most daily neurosurgical practices. Large databases with detailed data on patient characteristics, processes of care, and clinically meaningful outcomes will be a very powerful research tool. This routine and robust data collection inseparable from clinical practice will guide neurosurgical practice. Traditionally, a small number of physicians have produced new medical knowledge. This reliance on a scientific elite for the generation of new knowledge should be replaced by a system in which many physicians participate in clinical research through the acquisition of practice data.

References

- Ahn JM, Kang SJ, Yoon SH, Park HW, Kang SM, Lee JY, et al: Meta-analysis of outcomes after intravascular ultrasound-guided versus angiography-guided drug-eluting stent implantation in 26,503 patients enrolled in three randomized trials and 14 observational studies. *Am J Cardiol* **113**:1338–1347, 2014
- Cushing H, Eisenhardt L: **Meningiomas, Their Classification, Regional Behavior, Life History, and Surgical End Results**. Springfield, IL: Charles C Thomas, 1938
- Esene IN, Baesa SS, Ammar A: Evidence-based neurosurgery. Basic concepts for the appraisal and application of scientific information to patient care (Part II). *Neurosciences (Riyadh)* **21**:197–206, 2016
- Flay BR: Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Prev Med* **15**:451–474, 1986
- Freedman B: Equipoise and the ethics of clinical research. *N Engl J Med* **317**:141–145, 1987
- Hogan JW, Lancaster T: Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Stat Methods Med Res* **13**:17–48, 2004
- Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al: Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* **286**:821–830, 2001
- Li Q, Zhang Z, Yin RX: Drug-eluting stents or coronary artery bypass grafting for unprotected left main coronary artery disease: a meta-analysis of four randomized trials and seventeen observational studies. *Trials* **14**:133, 2013
- Lilford RJ: Ethics of clinical trials from a bayesian and decision analytic perspective: whose equipoise is it anyway? *BMJ* **326**:980–981, 2003
- Lurie JD, Tosteson TD, Tosteson A, Abdu WA, Zhao W, Morgan TS, et al: Long-term outcomes of lumbar spinal stenosis: eight-year results of the Spine Patient Outcomes Research Trial (SPORT). *Spine (Phila Pa 1976)* **40**:63–76, 2015
- Psaty BM, Breckenridge AM: Mini-Sentinel and regulatory science—big data rendered fit and functional. *N Engl J Med* **370**:2165–2167, 2014
- Rubin DB: For objective causal inference, design trumps analysis. *Ann Appl Stat* **2**:808–840, 2008
- Rubin D: Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. *Handbook of Statistics* **27**:28–63, 2008
- Torner JC, Zhang J, Piegras D, Huston J, Meissner I, Brown R, et al: Comparative effectiveness of long-term outcomes of treatment of unruptured intracranial aneurysms. *Stroke* **43**:A3906, 2012 (Abstract)
- Treweek S, Zwarenstein M: Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials* **10**:37, 2009
- Wallis CJD, Detsky AS, Fan E: Establishing the effectiveness of procedural interventions: the limited role of randomized trials. *JAMA* **320**:2421–2422, 2018
- Yang JY, Webster-Clark M, Lund JL, Sandler RS, Dellon ES, Stürmer T: Propensity score methods to control for confounding in observational cohort studies: a statistical primer and application to endoscopy research. *Gastrointest Endosc* **90**:360–369, 2019

Disclosures

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

Correspondence

Robert E. Harbaugh: Penn State University, Milton S. Hershey Medical Center, Hershey, PA. reh21@psu.edu.