What Is Medical Evidence?

Fred G. Barker, II, M.D.

What is medical evidence? How can we best find it, interpret it, and contribute to it? The knowledge base that underpins clinical practice has become almost unimaginably broad and deep. With increasing numbers of medical articles published yearly, it has long been impossible for any individual to know all there is to know about human health and disease. Even within one's own narrowly defined subspecialty, trying to keep up with the rapid pace of the medical literature is a very difficult task. We need to have some sort of structure within which we can frame both new and old medical knowledge in terms of its relative value—its current importance, its applicability to our actual practice, its reliability, and its likely degree of permanence. However, when we consider the broad range of possible sources of medical knowledge—our own personal experience, the shared wisdom of teachers and colleagues, information seen or heard at medical meetings, published peer-reviewed and non–peer-reviewed articles, internet sources—we quickly see that any such hierarchy of knowledge can be only a rough guide. Nonetheless, our daily responsibility to patients requires us to exercise judgment about best estimates of prognosis, best procedures for diagnosis, and, as surgeons, especially the *best measures of active treatment*. This article attempts to outline the relatively widely accepted aspects of the so-called hierarchy of evidence (*Table 4.1*) as it applies to treatment comparisons. The hierarchy classifies evidence based on prospective, randomized comparisons as level 1; those based on prospective, nonrandomized comparisons as level 2; retrospective comparative studies as level 3; retrospective studies without clearly defined controls as level 4; and all other evidence as level 5. It is clear not only that broad aspects of medical knowledge applying to diagnosis, prognosis, resource use, health services research, basic laboratory research, and so on are often not well described by this hierarchy, but also that different versions of the table of "levels of evidence" differ in minor details of the ranking. But one has to start somewhere.

0148-703/09/5601-0024

EXPERT OPINION, CASE REPORTS, AND UNCONTROLLED CASE SERIES

One often sees the hierarchy of evidence presented from the top down, that is, starting with the randomized, controlled trial (RCT) and progressing to less and less reliable forms of evidence (*Table 4.1*). This can lead to a dismissive view of the unimportance of all levels of evidence below the RCT. Although there is nearly universal agreement that the RCT is a highly reliable source of medical knowledge, neurosurgeons have the unique problem that almost nothing that we do can be supported by reference to a randomized trial. For craniotomy, for example, we have good evidence for administering prophylactic antibiotics before making the incision, but then, if RCTs were the only source of knowledge, we really would not know what to do next in treating even such a familiar entity as symptomatic epidural hematoma, which every junior resident must know immediately how to handle.

Every neurosurgeon uses daily the body of knowledge known as expert opinion, usually in the form of what they learned during residency but sometimes in a more immediate form. For example, a surgeon confronted with difficulty closing an incision might ask a senior colleague or plastic surgeon to step down to the operating room for advice, and scoring the galea or making a relaxing incision in the lumbar fascia might be suggested. Although no RCTs exist to support these suggestions, we know that they do in fact solve many wound closure problems effectively. This class of problem—immediate technical difficulty—is most often resolved through expert advice, as are many problems of preoperative judgment and postoperative management that depend largely on successful pattern recognition by a more experienced surgeon. It is characteristic of such clinical problems that they are highly dependent on individual circumstances or conditions so that RCTs may be difficult or impossible to apply to them.

Published parallels to expert or anecdotal opinion are case reports or small case series.[35,60,84] Case reports have been an important teaching and learning device since at least the time of Hippocrates.[101] During the early development of neurosurgery, case reports and small series were the only means of dissemination of diagnostic and therapeutic advances.[11,27,87] More than half of the current neurosurgical literature consists of case reports and small, uncontrolled case series.[44,90]

TABLE 4.1. The hierarchy of medical evidence

Meta-analysis of randomized, controlled trials without significant heterogeneity	Level 1
Randomized, controlled trial	Level 1
Prospective cohort study with concurrently treated controls	Level 2
Case-control study	Level 3
Retrospective cohort study with concurrent or historical controls	Level 3
Case series (with historical or without controls)	Level 4
Expert opinion, "first principles" argument, laboratory evidence	Level 5

The qualities of novelty and rarity continue to be characteristics that suggest anecdotal experience (case reports or small case series) may be the only source of evidence for a clinical question. Vandenbroucke[121] identified potential roles of case reports and case series as including recognition and description of new diseases[21,103] or uncommon manifestations of common diseases[36]; demonstration of new surgical techniques[55]; detection of uncommon adverse drug reactions[79,123] or drug benefits in rare diseases or unique groups, especially for dire clinical situations in which there is no standard treatment[61]; studying mechanisms of disease (or disturbances of normal physiology)[4,33,69,94]; and medical education.[48] Recent neurosurgical examples include the identification of superior semicircular canal syndrome,[86] current interest in anti-mTOR (mammalian target of rapamycin) treatment for chordoma based on reports of occurrence in patients with tuberous sclerosis with loss of heterozygosity for the TSC1 and TSC2 loci,[19,75,105] and response of some neurofibromatosis-related vestibular schwannomas to bevacizumab treatment.[97]

Single observations are always at risk of resulting from simple chance, and no statistical tests can distinguish truth from coincidence. Some characteristics that have been suggested as indicating true associations include immediacy and strength of the effect, stability of the baseline condition, consistency across settings, biological plausibility, and specificity (treatment causes the effect and little else).[43] As time passes and rare observations begin to accumulate in the literature, a "meta-analysis" based on case reports can be produced. Such reports can sometimes define prognostic factors in rare diseases, such as trilateral retinoblastoma[66] or malignant glioma of the cerebellum.[32] Such compilations of case series do not ordinarily provide strong evidence of treatment effects because of the usual confounding factors that affect nonrandomized treatment assessments. For example, a meta-analysis of case reports and small series of osteosarcoma of the head and neck showed that survival time was shorter when postoperative adjuvant radiation and chemotherapy were used.[63] The authors explained the effect as due to "confounding by indication,"[84] that is, the survival difference was more directly related to the indication for treatment than to the effect of the treatment itself. The authors of this study concluded that adjuvant therapies were probably beneficial despite the results of their review.

CASE SERIES WITH INTERNAL OR EXTERNAL CONTROLS

Many questions that are closely related to those best handled by expert advice should, however, be resolved by more reliable evidence. With reference to wound closure, for example, the relative merits of skin suture versus staples would not usually be settled reliably by appealing to a senior surgeon's personal experience. The relatively low incidence of wound problems with either type of closure and the lack of the immediately verifiable difference between the two methods mean that careful data collection from a large number of cases would be necessary to gain even a rough impression of the superiority of one method over the other.

A relatively common research design for testing treatments is the observational (i.e., nonrandomized) case series with controls. The cases may be studied prospectively or retrospectively, and the controls may be historical or concurrently treated. Sometimes historical controls are drawn from previously published literature (external controls) and sometimes they are directly studied by the investigators (internal controls). In the oncology literature, case series with an implied or explicit external control group are called phase II trials and are used to explore new antitumor drugs for evidence of activity. The purpose of phase II trials is to select promising drugs for definitive proof of efficacy in randomized phase III trials. In neurosurgery, a case series with comparison with some type of controls is often the best existing evidence for a neurosurgical treatment because neurosurgical RCTs are so rare, approximately 1% to 2% of articles in neurosurgical journals.[44,90] In drug treatment of stroke, Parkinson's disease, and head trauma, increasing emphasis is being placed on phase II trials because the track record of phase III trials in these fields, often based purely on preclinical animal model data, has been very poor.[20,93,120] A well-designed and well-conducted phase II trial can often avoid the much greater expense and investment of other resources in a phase III trial of an ineffective agent.

When judging the quality of an observational study, it is helpful to examine the methods used to select and describe the cases separately from those used to define the control population. Strong methods for generating a case series worthy of further study are to define the cases prospectively using prespecified criteria and to ensure that the case series is consecutive. Case series that produce results that generalize well to actual practice tend to have few exclusion criteria and to be population based. Patients who seek care at specialized

centers tend to be healthier than patients in unselected practice as well as often being enriched for social factors that are associated with better prognoses, such as high levels of education, social support, and wealth.[72] Tight restrictions on comorbidity, age, presenting symptoms and signs, performance status, and so on obviously have a similar effect. Retrospectively studied case series can have incomplete case ascertainment and missing data for important prognostic factors. As well, investigators may have chosen to report their series because the results have been particularly favorable, a form of publication bias. For example, in-hospital mortality after pancreatic resections was 2.4 times higher in a population-based sample than in reported case series in one study.[116] A similar bias has been reported for published single-center unruptured aneurysm clipping series.[127]

The selection of the control population is important in defining the reliability of an observational study. Historical controls often are more convenient and inexpensive compared with concurrently treated controls, and when the trial design is "before and after" the introduction of some new treatment or technology, there may be no other option. However, experience has shown that the prognosis for historical controls is often worse than concurrently treated controls even in the absence of any perceptible shift in the field.[46] For example, in-hospital mortality for craniotomy for tumor decreased substantially during the period from 1988 to 2000, despite any dramatic single innovation in care during this time.[8] Furthermore, the average length of stay decreased substantially during this period (as for other types of craniotomy[6]), largely because of pressure on hospital reimbursement by payers for care. Study designs comparing length of stay (and the closely dependent variable hospital cost) before and after the introduction of new technology always show a (spurious) improvement with the new method unless it is actually harmful. In addition, historical controls often lack data on important prognostic factors that may hinder the selection of a truly comparable control group.

Studies have shown that properly conducted observational studies tend to produce results that are similar to those of RCTs on the same question.[12,23,57] This is only true, however, when the methods adopted by the observational studies closely mimic those used in a typical RCT. Specifically, controls should be concurrently treated and should be equally eligible for the treatment that the patients themselves received. Failure to adhere to these requirements has frequently caused seriously misleading results in neurosurgical investigations. For example, brachytherapy, radiosurgery, and intra-arterial chemotherapy all showed apparent improvements in survival when used to treat glioblastoma in case series from specialized centers.[47,99,110] However, RCTs later showed that all of these treatments were ineffective.[1,74,108,109] Subsequent study showed that historical control patients who were selected for their eligibility for the treatment (but did not actually receive the treatment) had a survival advantage based solely on their eligibility, which explained the overoptimistic results in case series reports.[26,37,58,65,80]

Another extremely important source of bias in observational studies is the selection of patients for an investigational treatment because they are "good candidates," i.e., healthy enough to be expected to benefit from the therapy. A recent example is the widespread use of hormone replacement therapy (HRT) for postmenopausal women in the late 1990s. HRT was expected to lower mortality from coronary heart disease based on epidemiological considerations and observational treatment studies on large populations of women. When RCTs of HRT were completed, it was apparent that there was no evidence of benefit from the treatment and in fact some suggestion that HRT was harmful.[91] The apparent reason was that physicians had prescribed HRT preferentially to women who were healthier—leaner, more affluent and educated, more frequent exercisers—and with greater access to regular health care, in short, women who had longer life expectancies and who would be expected to benefit from a prophylactic treatment. The survival advantage that they enjoyed was spurious, a result of confounding by indication and not of the treatment itself.

Methods used in observational studies to generate an appropriate control group are therefore very important in reducing the chance of a misleading result. Unfortunately, many (perhaps most) neurosurgical observational studies ignore this aspect of research design. For example, almost no studies on the effect of extensive resection on survival in malignant glioma have been limited to those patients who were actually eligible for complete resection; indeed, most contain patients who underwent biopsies, presumably of unresectable tumors.[5]

Appropriate methods for selecting and adjusting for controls in observational studies include matching, multivariate analysis using individual prognostic factors, use of risk stratification scores, propensity scores, and instrumental variables.[30,68] *Matching* refers to selecting a matched control for each case based on knowledge of important prognostic factors. For example, one might choose a control for a patient with brain metastases by selecting a person with the same number of metastases, primary tumor histology, and similar age and performance score. A disadvantage of matching is that when multiple important prognostic factors exist, matching all of them for each case becomes very difficult. In this situation, using *multivariate analysis* to model the effect of all the prognostic factors on each case and (nonmatched) control is simpler, but a good knowledge of all important prognostic factors and their effects is still required.[67] In some situations, it may be possible to construct a *risk stratification score* using historical controls. An example is the Radiation Therapy Oncology Group's Recursive Partitioning Analysis classification for malignant glioma patients.[25] This widely

used scheme combines information on several important prognostic factors to generate six risk classes with distinct survival expectations. Each case is then matched to a historical control who has the same Recursive Partitioning Analysis class. A very different method of adjustment is the use of the *propensity score*.[3] This is a constructed variable that expresses the probability of each patient receiving the study treatment based on observable baseline characteristics. For example, older patients are more likely to receive coiling than clipping for ruptured aneurysms, so their propensity scores for coiling would be higher.[7] Propensity scores are usually constructed by relating baseline factors to treatment assignment using logistic regression. When each case and each potential control has been assigned a propensity score, matching can be used to construct a control cohort, or the entire group of cases and controls can be analyzed using the propensity score as a stratification variable. Finally, *instrumental variables* can be used to construct an appropriate comparison of cases and controls.[49] An instrumental variable is one that is strongly predictive of received treatment but not directly (causally) related to outcome (such as treatment assignment in a randomized trial). This analysis is commonly used in econometrics but is relatively novel in medical studies.[42] One or more of these methods should be used in all observational studies of treatment efficacy.

RCTs: PHASE III TRIALS

Although observational studies may produce results that are in accordance with results of RCTs, in other instances, results can be quite misleading, and it is never possible to be sure in the absence of well-designed and well-conducted RCTs whether observational results are reliable—the "unpredictability paradox."[71] In practice, every treatment assignment that is not formally randomized contains some degree of bias, which is never completely captured by variables that can be observed or recorded. An adequate randomization is the only known means of removing both conscious and unconscious bias from treatment assignment and providing a level playing field for a treatment comparison. This is why it is worth the undeniable extra effort and expense of mounting a good RCT for any important treatment question.

Several features of RCT design are widely believed to be especially important in ensuring freedom from bias. The first is *adequate concealment of treatment allocation*. By definition, the fundamental distinction between RCTs and nonrandomized designs is the random nature of allocation to the two treatments being compared. Any trial in which investigators can discover the treatment allocation for a patient before actually enrolling him or her allows the bias associated with nonrandom assignment to sneak back into the trial in a concealed manner. This problem is present in its strongest form when "pseudorandom" allocation is openly assigned based on information such as the patient's medical record number, birth date, or day of enrollment. In more subtle forms, investigators may have advance access to randomization lists or even be able to discover assignment by holding sealed translucent envelopes up to the light.[106] Comparisons between trials with and without adequate randomization show that trials with inadequately concealed allocation produce estimates of treatment effect that are as much as 40% larger than trials with well-concealed allocation schemes, as well as larger imbalances in prognostic factors between groups.[22,88,96,106]

A second important design feature in RCTs is *blinding*. Ideally, the treating physician, the patient, and the observer who adjudicates the trial's endpoints would all be unaware of which treatment the patient received (triple blinding). Blinding is an important protection against bias for several reasons. For the treating physician, blinding prevents preferential use (or omission) of other components of treatment other than the one under study. For the patient, blinding prevents differences in compliance and potential crossover between arms, as well as biased reporting of subjective endpoints such as pain or toxicity. For assessors of endpoints, blinding prevents conscious or unconscious bias in adjudication. Objective studies have shown larger treatment effects and more frequent imbalance in prognostic variables between arms in RCTs that report unclear or inadequate blinding,[22,107] especially when the trial's endpoint is subjective (e.g., pain) rather than objective (e.g., mortality).[126] Blinding is obviously difficult or impossible in many trials that test a surgical treatment, unless sham surgery is used, but this carries both practical and ethical problems of its own.[38,39,64,85] However, in many surgical trials, the primary endpoint can be scored by a central committee that is blinded to treatment allocation, and blinding of both physician and patient can be tested objectively during or after the trial if it might be important.

A third feature of trial conduct important in preventing bias is *completeness of follow-up*. This is because loss to follow-up is not random: patients who are lost will likely differ in prognostic factors, outcome, or both compared with those who comply with follow-up, either for the worse (dissatisfied with results, too sick to attend follow-up evaluation, or dead) or less commonly for the better (asymptomatic so do not bother to return). Additionally, for some surgical procedures, compliance with multidisciplinary follow-up care is also important in ensuring the long-term success of the operation.[98] For example, patients lost to follow-up in a large case series of microvascular decompression were more likely to have experienced hearing loss as a complication of the operation.[9] Younger patients and those who belong to socially disadvantaged groups are frequently found to be at risk of loss to follow-up.[9,34,73,89] An arbitrary rate of 80% completeness of follow-up is sometimes quoted as adequate, but in fact the threshold above which loss to follow-up threatens the validity of the trial depends on the frequency of the

endpoint being tested. When the endpoint is rare and its occurrence is correlated with loss to follow-up, almost any loss could threaten the reliability of the trial's results. Methods for testing the importance of follow-up loss in a specific trial can include a worst-case sensitivity analysis in which all patients lost from the treatment arm are assigned a bad outcome and those lost from the control arm are assigned a good outcome, and the trial results are recomputed. The threat to validity from follow-up loss can also be studied by relating known baseline prognostic factors to the rate of loss or, when interval outcomes are available, by comparing the last known results before loss with those of comparable patients who were not lost.

Another important feature of the most reliable RCTs is the *intent-to-treat* analysis. This refers to analysis of patients grouped by prescribed treatment in a trial rather than by treatment as received.[45] Such exclusions from analysis have been shown to bias trials systematically in the direction of the new treatment being tested.[119] For surgical trials, failure to receive prescribed treatment usually reflects *crossover*, such as failure to undergo a prescribed operation (because the patient becomes too ill for the surgery or because symptoms resolve while waiting) or undergoing the operation outside the trial after randomization to conservative therapy.[100] Failure to use intent-to-treat analysis can cause serious bias when lack of compliance with prescribed treatment results from clinical deterioration, as when patients randomized to receive carotid endarterectomy have a stroke as a complication of a preoperative arteriogram.[102] In other trials, especially those that test operations that are available outside the trial as "standard therapy," frequent use of surgery outside the trial by patients randomized to conservative therapy reduces the trial's ability to measure the effect of the operation itself. Instead, the effect of a "strategy" of prescribing conservative therapy, with surgery allowed if symptoms persist, is what is being tested. This tends to make surgeons unhappy, but as long as it is clearly understood when interpreting the trial, the results are not invalid. From a practical standpoint, to minimize problems arising from patients not receiving protocol treatment, investigators must verify eligibility thoroughly before enrolling patients in an RCT.

SPECIAL PROBLEMS IN SURGICAL RCTs

The relative paucity of RCTs in the surgical literature has prompted some authors to consider special reasons why surgical RCTs are particularly difficult to design or conduct.[78,82,83] Some of these special surgical considerations apply to observational studies as well.

Designing a surgical RCT and persuading both patients and surgeons to participate may be difficult because of a lack of *equipoise* on the patient's or surgeon's part. Equipoise is a genuine uncertainty as to which of two possible treatments is most likely to lead to a good outcome.[40] Both surgeons and patients seem particularly unlikely to submit a treatment choice to random assignment when surgery is being compared with a nonsurgical treatment such as observation or radiation.[112] This may result, on the surgeon's part, from bias attributable to differential compensation, as well as from a reluctance to appear uncertain about the benefits of surgery or fear of losing future referrals from physicians if surgery is not undertaken. Difficulty with equipoise in surgical trials is a complex problem and looms large among reasons why patients and surgeons fail to participate in apparently well-designed RCTs.[113,114,128,129]

Surgical RCTs often have trouble with blinding physicians, patients, and outcome assessors to the treatment assignment. Knowledge of assigned treatment on the surgeon's part may cause patients to receive other effective therapies in a biased fashion and on the patient's part may lead to biased reporting of patient-assessed outcomes such as pain or quality of life. As mentioned above, sham surgery or blinded assessment of outcome by a third party may help to ameliorate these problems.

Timing of surgical trials is often problematic. Although medical treatments are relatively easily standardized, surgeons gradually develop better skills as they acquire experience (the learning curve[92,125]), and the treatment itself can evolve over time (as when better coiling technology becomes available). This can mean that RCTs with lengthy accrual periods can become out of date and irrelevant to practice before they are completed or analyzed.

SYSTEMATIC REVIEWS AND META-ANALYSES

In many descriptions of the hierarchy of evidence, a meta-analysis of RCTs without significant heterogeneity occupies the highest position.[51] This reflects the concept that when several well-designed trials have produced generally concordant results, those results gain in credibility over a single trial. Some special problems that arise in collecting and combining data from individual trials deserve attention.

A meta-analysis is usually understood to be a mathematical synthesis of quantitative results from more than one individual trial. Meta-analysis involves four steps: formulating an answerable question, collecting data, examining its quality, and finally combining it to give a single estimate of treatment effect.

Questions that suggest a meta-analysis as the proper means of study often concern the effect of a treatment for which more than one reported trial already exists, especially when the existing trials are too small to produce a definitive answer individually or when there is apparent conflict between their results. The unit under observation in a meta-analysis is the individual research trial, so instead of specifying the study patient population and sampling methods, a protocol describes a strategy for identifying all relevant studies, both published and unpublished, as well as formal

criteria to determine which of the identified studies can be included. Some subjective judgment often enters here, for example, when trials differ in patient population or in treatment details, and bias is minimized through specifying these choices in a written protocol before the search is done. Combining trials that are too different in patient population or treatment method is like comparing apples and oranges.

Data for the meta-analysis are collected from a systematic review of published and unpublished evidence that bears on the study question using established methods.[76] The goal is to identify all evidence, both published and unpublished, that bears on the chosen question. Unpublished data are important to capture because trials that are never published or are delayed in publication (whether because of the author's inaction or the peer-review process) are more likely to have negative results (the "file-drawer" problem). Reviews limited to published results are thus skewed toward trials with positive results, an effect known as *publication bias*.[81,95]

With identified studies in hand, the meta-analyst extracts data from the primary trials. In extracting data, a basic choice is whether to rely on published data only or to seek data on individual patients from the trialists who conducted the primary studies (an *individual patient data* meta-analysis). Individual patient data meta-analyses have much greater power for identifying the influence of patient characteristics on treatment effects, but require the cooperation of the original trialists and are very resource intensive.[13,115]

Trials differ in quality as well, mandating another subjective choice for the analyst. Most meta-analyses are limited to RCTs, which is a simple and sensible quality judgment in itself. Incorporating nonrandomized studies into a meta-analysis, especially when there are no randomized studies to act as an internal gold standard, carries all the bias of the original studies into the analysis and can produce seriously misleading conclusions. This is because the mathematical process of meta-analysis can isolate and magnify any consistent biases in the original studies just as efficiently as it can detect weak treatment effects that are present in the studies—the "garbage in, garbage out" problem. Many standardized instruments for assessing the "quality" of both randomized[59,124] and observational treatment[111] trials have been described, based on various details of the trials' methods and reporting, but none have been shown to correlate predictably with larger or smaller treatment effect estimates. Despite this, researchers often include a comparison of treatment effect estimates derived from high- and low-quality trials as a sensitivity analysis.

The final step in meta-analysis, the statistical process of combining the study results to yield a single unified conclusion, is based on the hypothesis that studies addressing a similar question are drawn from a possible population of similar studies that should produce answers varying in a predictable fashion around the "true" answer. Almost any measure of treatment effect, such as odds ratios or risk differences, can be combined in a meta-analysis.[28,31] The two basic methods used to actually combine the individual trial results are the *fixed-effects* method, which assumes that all trials provide individual estimates of the same treatment effect differing only by random chance, and the *random-effects* method, which assumes that the true treatment effect might differ slightly among trials (a more conservative and hence safer assumption).[31] After a summary measure of treatment effect is constructed, with appropriate confidence intervals, researchers compute a measure of the *heterogeneity* present in the analysis.[29] Heterogeneity is the degree to which trials' results differ from one another in excess of what would be expected from the play of chance. A large amount of heterogeneity among trials means that a summary measure of treatment effect cannot be confidently applied to all the patients and/or treatments included in the individual trials. Sometimes this can suggest a search for differences among trials that explain the observed heterogeneity.[117,118] For example, a treatment's efficacy might differ in specific patient populations defined by age, tumor type, or symptom severity at presentation.

Published meta-analyses can be sought by searching PubMed using publication type "meta-analysis" in combination with specific subject terms. As more evidence accumulates on a given clinical question, a meta-analysis needs to be revised and updated, although most meta-analyses published in standard medical journals have no preplanned means of keeping results current. The Cochrane Collaboration has produced systematic reviews on a broad range of medical topics using standardized techniques[14,18,24,77,104]; these are updated regularly by the original analysis team or their successors. The organization also provides protocols and software to aid in performing new meta-analyses at www.cochrane.org.

MOVING FROM EVIDENCE TO RECOMMENDATION

Although in most versions of the hierarchy of evidence (*Table 4.1*) a meta-analysis of randomized trials occupies the top rung, in some, the very best evidence is said to come from an "N-of-1" RCT.[51] In this type of trial, one patient receives each of two treatments in an alternating fashion, and outcomes while the patient is taking the two different agents are compared.[62] Obviously, there are many surgical questions for which a crossover design is impossible because of the permanent nature of the treatment. For drug therapies, the reason that this design is considered so strong is that, unlike evidence from RCTs with restrictive entry criteria, the evidence gained from an N-of-1 trial can confidently and immediately be applied to the one person in whom the trial was conducted. In other words, some RCTs lack *external validity* or *generalizability* because the conditions of the

trial do not resemble everyday practice and the patients who entered the trial were highly selected rather than reflecting the entire community or persons with the disease in question. In an N-of-1 trial, external validity usually does not become an issue because the results are never used to generalize to the population at large.

RCTs that are intended to have maximal external validity (i.e., that test a therapy delivered under real-world conditions) are said to measure treatment *effectiveness*, whereas RCTs that test whether a therapy works under ideal conditions test *efficacy*. Although this seems straightforward, the characteristics of the two types of RCT are open to some debate. A consensus study of epidemiology experts identified, among others, that characteristics of effectiveness studies included primary care populations, less stringent eligibility criteria, patient-relevant endpoints, and intent-to-treat analysis.[41]

Another approach to increasing the external validity of an RCT is to supplement its results with observational results on the same therapy in a broad patient population.[16,17,50,70,122] Large existing databases, such as administrative databases or case registries, can offer a cost-effective means of monitoring treatment efficacy in unselected populations. Unfortunately, there are significant barriers to using most large databases for this purpose, especially administrative databases. These barriers include inaccurate coding of diagnoses and procedures,[10] difficulties ensuring adequate risk adjustment based on observed and unobserved prognostic factors and comorbidities,[56] difficulty distinguishing presenting signs and symptoms from complications of treatment,[15] and lack of long-term follow-up information in many databases. Conversely, because most clinical trials are powered to show the benefit of a treatment, their ability to identify rare but serious treatment morbidity is usually poor. Large databases offer higher sensitivity for detecting these rare treatment toxicities or patient safety events such as retained foreign bodies after surgery.

As noted above in the discussion of systematic reviews and meta-analyses, combining results from several similar studies of the same or related treatments is a means of broadening the applicability of a treatment decision or recommendation. The GRADE Working Group is a consensus group of experts who has developed a system for guideline developers to rate the quality of the evidence that underlies their recommendations.[2] This approach involves making separate ratings for quality of evidence for each patient important outcome based on five factors that limit quality of evidence: study limitations, consistency, directness and precision of evidence, and the possibility of publication bias.[52–54] The GRADE system explicitly recognizes the problems in applying evidence-based medicine to individual patients, which is the final step in actually using the evidence obtained through scientific trials in daily practice.

CONCLUSION

One striking characteristic of medical evidence is the diversity of study designs that are used to generate it. This chapter addresses many of the most common study designs, each of which has both strengths and weaknesses. Given the lack of RCT evidence for most neurosurgical treatments and decisions, fluency in evaluating nonrandomized evidence will continue to be important for neurosurgeons for the foreseeable future. A commitment to designing and conducting RCTs of novel procedures in the future will be increasingly necessary, both because of our own desire to offer the best care to patients and (most likely) because of formal requirements by government regulatory agencies and payers. Understanding the conditions under which nonrandomized evidence can take the place of RCTs is a challenge for the future.

Disclosure

The author has no personal financial or institutional interest in any of the drugs, materials, or devices described in this article.

REFERENCES

1. Andrews DW, Scott CB, Sperduto PW, Flanders AE, Gaspar LE, Schell MC, Werner-Wasik M, Demas W, Ryu J, Bahary JP, Souhami L, Rotman M, Mehta MP, Curran WJ Jr: Whole brain radiation therapy with or without stereotactic radiosurgery boost for patients with one to three brain metastases: Phase III results of the RTOG 9508 randomised trial. **Lancet** 363:1665–1672, 2004.
2. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D, Hill S, Jaeschke R, Leng G, Liberati A, Magrini N, Mason J, Middleton P, Mrukowicz J, O'Connell D, Oxman AD, Phillips B, Schunemann HJ, Edejer TT, Varonen H, Vist GE, Williams JW Jr, Zaza S: Grading quality of evidence and strength of recommendations. **Br Med J** 328:1490, 2004.
3. Austin PC: A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. **Stat Med** 27:2037–2049, 2008.
4. Barker FG 2nd: Phineas among the phrenologists: The American crowbar case and nineteenth-century theories of cerebral localization. **J Neurosurg** 82:672–682, 1995.
5. Barker FG 2nd: Comment [on Sanai N, Berger MS: Glioma extent of resection and its impact on patient outcome]. **Neurosurgery** 62:765–766, 2008.
6. Barker FG 2nd, Amin-Hanjani S: Changing neurosurgical workload in the United States, 1988–2001: Craniotomy other than trauma in adults. **Neurosurgery** 55:506–517; discussion 517–508, 2004.
7. Barker FG 2nd, Amin-Hanjani S, Butler WE, Hoh BL, Rabinov JD, Pryor JC, Ogilvy CS, Carter BS: Age-dependent differences in short-term outcome after surgical or endovascular treatment of unruptured intracranial aneurysms in the United States, 1996–2000. **Neurosurgery** 54:18–28; discussion 28–30, 2004.
8. Barker FG 2nd, Curry WT Jr, Carter BS: Surgery for primary supratentorial brain tumors in the United States, 1988 to 2000: The effect of provider caseload and centralization of care. **Neuro-oncology** 7:49–63, 2005.
9. Barker FG 2nd, Jannetta PJ, Bissonette DJ, Larkins MV, Jho HD: The long term outcome of microvascular decompression for trigeminal neuralgia. **N Engl J Med** 334:1077–1083, 1996.
10. Benesch C, Witter DM Jr, Wilder AL, Duncan PW, Samsa GP, Matchar DB: Inaccuracy of the International Classification of Diseases (ICD-9-CM) in identifying the diagnosis of ischemic cerebrovascular disease. **Neurology** 49:660–664, 1997.

11. Bennett AH, Godlee RJ: Excision of a tumour from the brain. **Lancet** i:13, 1884.
12. Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. **N Engl J Med** 342:1878–1886, 2000.
13. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI: Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. **Stat Med** 21:371–387, 2002.
14. Bero L, Rennie D: The Cochrane Collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. **JAMA** 274:1935–1938, 1995.
15. Best WR, Khuri SF, Phelan M, Hur K, Henderson WG, Demakis JG, Daley J: 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** 194:257–266, 2002.
16. Black N: Why we need observational studies to evaluate the effectiveness of health care. **Br Med J** 312:1215–1218, 1996.
17. Black N: What observational studies can offer decision makers. **Horm Res** 51(Suppl 1):44–49, 1999.
18. Bombardier C, Esmail R, Nachemson AL: The Cochrane Collaboration Back Review Group for spinal disorders. **Spine** 22:837–840, 1997.
19. Börgel J, Olschewski H, Reuter T, Miterski B, Epplen JT: Does the tuberous sclerosis complex include clivus chordoma? A case report. **Eur J Pediatr** 160:138, 2001.
20. Bullock MR, Lyeth BG, Muizelaar JP: Current status of neuroprotection trials for traumatic brain injury: lessons from animal models and clinical studies. **Neurosurgery** 45:207–217; discussion 217–220, 1999.
21. Centers for Disease Control and Prevention: Kaposi's sarcoma and *Pneumocystis* pneumonia among homosexual men—New York City and California. **MMWR Morb Mortal Wkly Rep** 30:305–308, 1981.
22. Chalmers TC, Celano P, Sacks HS, Smith HJ: Bias in treatment assignment in controlled clinical trials. **N Engl J Med** 309:1358–1361, 1983.
23. Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. **N Engl J Med** 342:1887–1892, 2000.
24. Counsell C, Warlow C, Sandercock P, Fraser H, van Gijn J: The Cochrane Collaboration Stroke Review Group. Meeting the need for systematic reviews in stroke care. **Stroke** 26:498–502, 1995.
25. Curran WJ Jr, Scott CB, Horton J, Nelson JS, Weinstein AS, Fischbach AJ, Chang CH, Rotman M, Asbell SO, Krisch RE, Nelson DF: Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. **J Natl Cancer Inst** 85:704–710, 1993.
26. Curran WJ Jr, Scott CB, Weinstein AS, Martin LA, Nelson JS, Phillips TL, Murray K, Fischbach AJ, Yakar D, Schwade JG, Corn B, Nelson DF: Survival comparison of radiosurgery-eligible and -ineligible malignant glioma patients treated with hyperfractionated radiation therapy and carmustine: A report of Radiation Therapy Oncology Group 83-02. **J Clin Oncol** 11:857–862, 1993.
27. Cushing H: The surgical aspects of major neuralgia of the trigeminal nerve. A report of twenty cases of operation on the Gasserian ganglion, with anatomic and physiologic notes on the consequences of its removal. **JAMA** 44:773–778, 860–865, 920–929, 1002–1008, 1088–1093, 1905.
28. Deeks JJ: Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. **Stat Med** 21:1575–1600, 2002.
29. Deeks JJ, Altman DG, Bradburn MJ: Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis, in Egger M, Davey Smith G, Altman DG (eds): *Systematic Reviews in Health Care: Meta-analysis in Context.* London, BMJ, 2001, ed 2.
30. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, Petticrew M, Altman DG: Evaluating non-randomised intervention studies. **Health Technol Assess** 7:iii–x, 1–173, 2003.
31. DerSimonian R, Laird N: Meta-analysis in clinical trials. **Controlled Clin Trials** 7:177–188, 1986.
32. Djalilian HR, Hall WA: Malignant gliomas of the cerebellum: An analytic review. **J Neurooncol** 36:247–257, 1998.
33. Dronkers NF, Plaisant O, Iba-Zizen MT, Cabanis EA: Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. **Brain** 130:1432–1441, 2007.
34. Edwards P, Fernandes J, Roberts I, Kuppermann N: Young men were at risk of becoming lost to follow-up in a cohort of head-injured adults. **J Clin Epidemiol** 60:417–424, 2007.
35. Enkin MW, Jadad AR: Using anecdotal information in evidence-based health care: Heresy or necessity? **Ann Oncol** 9:963–966, 1998.
36. Fisher M: An unusual variant of acute idiopathic polyneuritis (syndrome of ophthalmoplegia, ataxia and areflexia). **N Engl J Med** 255:57–65, 1956.
37. Florell RC, Macdonald DR, Irish WD, Bernstein M, Leibel SA, Gutin PH, Cairncross JG: Selection bias, survival, and brachytherapy for glioma. **J Neurosurg** 76:179–183, 1992.
38. Frank SA, Wilson R, Holloway RG, Zimmerman C, Peterson DR, Kieburtz K, Kim SY: Ethics of sham surgery: Perspective of patients. **Mov Disord** 23:63–68, 2008.
39. Freed CR, Greene PE, Breeze RE, Tsai WY, DuMouchel W, Kao R, Dillon S, Winfield H, Culver S, Trojanowski JQ, Eidelberg D, Fahn S: Transplantation of embryonic dopamine neurons for severe Parkinson's disease. **N Engl J Med** 344:710–719, 2001.
40. Freedman B: Equipoise and the ethics of clinical research. **N Engl J Med** 317:141–145, 1987.
41. Gartlehner G, Hansen RA, Nissman D, Lohr KN, Carey TS: A simple and valid tool distinguished efficacy from effectiveness studies. **J Clin Epidemiol** 59:1040–1048, 2006.
42. Gillum LA, Johnston SC: Influence of physician specialty on outcomes after acute ischemic stroke. **J Hosp Med** 3:184–192, 2008.
43. Glasziou P, Chalmers I, Rawlins M, McCulloch P: When are randomised trials unnecessary? Picking signal from noise. **Br Med J** 334:349–351, 2007.
44. Gnanalingham KK, Tysome J, Martinez-Canca J, Barazi SA: Quality of clinical studies in neurosurgical journals: Signs of improvement over three decades. **J Neurosurg** 103:439–443, 2005.
45. Gravel J, Opatrny L, Shapiro S: The intention-to-treat approach in randomized controlled trials: Are authors saying what they do and doing what they say? **Clin Trials** 4:350–356, 2007.
46. Green S, Benedetti J, Crowley J: *Clinical Trials in Oncology*. London, Chapman & Hall/CRC, 2002, ed 2.
47. Greenberg HS, Ensminger WD, Chandler WF, Layton PB, Junck L, Knake J, Vine AK: Intra-arterial BCNU chemotherapy for treatment of malignant gliomas of the central nervous system. **J Neurosurg** 61:423–429, 1984.
48. Greenhalgh T: Narrative based medicine: Narrative based medicine in an evidence based world. **Br Med J** 318:323–325, 1999.
49. Greenland S: An introduction to instrumental variables for epidemiologists. **Int J Epidemiol** 29:722–729, 2000.
50. Guller U: Surgical outcomes research based on administrative data: Inferior or complementary to prospective randomized clinical trials? **World J Surg** 30:255–266, 2006.
51. Guyatt G, Rennie D, Meade MO, Cook DJ (eds): *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. New York, McGraw-Hill, 2008, ed 2.
52. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, Schunemann HJ: Going from evidence to recommendations. **Br Med J** 336:1049–1051, 2008.
53. Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ: What is "quality of evidence" and why is it important to clinicians? **Br Med J** 336:995–998, 2008.
54. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ: GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. **Br Med J** 336:924–926, 2008.
55. Higashida RT, Smith W, Gress D, Urwin R, Dowd CF, Balousek PA, Halbach VV: Intravascular stent and endovascular coil placement for a ruptured fusiform aneurysm of the basilar artery. Case report and review of the literature. **J Neurosurg** 87:944–949, 1997.

56. Iezzoni LI: Assessing quality using administrative data. **Ann Intern Med** 127:666–674, 1997.
57. Ioannidis JPA, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, Contopoulos-Ioannidis DG, Lau J: Comparison of evidence of treatment effects in randomized and nonrandomized studies. **JAMA** 286:821–830, 2001.
58. Irish WD, Macdonald DR, Cairncross JG: Measuring bias in uncontrolled brain tumor trials—To randomize or not to randomize? **Can J Neurol Sci** 24:307–312, 1997.
59. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, McQuay HJ: Assessing the quality of reports of randomized clinical trials: Is blinding necessary? **Control Clin Trials** 17:1–12, 1996.
60. Jenicek M: *Clinical Case Reporting in Evidence-based Medicine*. London, Arnold, 2001, ed 2.
61. Joensuu H, Roberts PJ, Sarlomo-Rikala M, Andersson LC, Tervahartiala P, Tuveson D, Silberman S, Capdeville R, Dimitrijevic S, Druker B, Demetri GD: Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. **N Engl J Med** 344:1052–1056, 2001.
62. Jull A, Bennett D: Do n-of-1 trials really tailor treatment? **Lancet** 365:1992–1994, 2005.
63. Kassir RR, Rassekh CH, Kinsella JB, Segas J, Carrau RL, Hokanson JA: Osteosarcoma of the head and neck: Meta-analysis of nonrandomized studies. **Laryngoscope** 107:56–61, 1997.
64. Kim SY, Frank S, Holloway R, Zimmerman C, Wilson R, Kieburtz K: Science and ethics of sham surgery: A survey of Parkinson disease clinical researchers. **Arch Neurol** 62:1357–1360, 2005.
65. Kirby S, Brothers M, Irish W, Florell R, Macdonald D, Schold C, Cairncross G: Evaluating glioma therapies: Modeling treatments and predicting outcomes. **J Natl Cancer Inst** 87:1884–1888, 1995.
66. Kivela T: Trilateral retinoblastoma: A meta-analysis of hereditary retinoblastoma associated with primary ectopic intracranial retinoblastoma. **J Clin Oncol** 17:1829–1837, 1999.
67. Klein-Geltink JE, Rochon PA, Dyer S, Laxer M, Anderson GM: Readers should systematically assess methods used to identify, measure and analyze confounding in observational cohort studies. **J Clin Epidemiol** 60:766–772, 2007.
68. Klungel OH, Martens EP, Psaty BM, Grobbee DE, Sullivan SD, Stricker BH, Leufkens HG, de Boer A: Methods to assess intended effects of drug treatment in observational studies are reviewed. **J Clin Epidemiol** 57:1223–1231, 2004.
69. Kordower JH, Cochran E, Penn RD, Goetz CG: Putative chromaffin cell survival and enhanced host-derived TH-fiber innervation following a functional adrenal medulla autograft for Parkinson's disease. **Ann Neurol** 29:405–412, 1991.
70. Krishnan E, Fries JF: Measuring effectiveness of drugs in observational databanks: Promises and perils. **Arthritis Res Ther** 6:41–44, 2004.
71. Kunz R, Oxman AD: The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. **Br Med J** 317:1185–1190, 1998.
72. Lamont EB, Hayreh D, Pickett KE, Dignam JJ, List MA, Stenson KM, Haraf DJ, Brockstein BE, Sellergren SA, Vokes EE: Is patient travel distance associated with survival on phase II clinical trials in oncology? **J Natl Cancer Inst** 95:1370–1375, 2003.
73. Lanoy E, Mary-Krause M, Tattevin P, Dray-Spira R, Duvivier C, Fischer P, Obadia Y, Lert F, Costagliola D: Predictors identified for losses to follow-up among HIV-seropositive patients. **J Clin Epidemiol** 59:829–835, 2006.
74. Laperriere NJ, Leung PM, McKenzie S, Milosevic M, Wong S, Glen J, Pintilie M, Bernstein M: Randomized study of brachytherapy in the initial management of patients with malignant astrocytoma. **Int J Radiat Oncol Biol Phys** 41:1005–1011, 1998.
75. Lee-Jones L, Aligianis I, Davies PA, Puga A, Farndon PA, Stemmer-Rachamimov A, Ramesh V, Sampson JR: Sacrococcygeal chordomas in patients with tuberous sclerosis complex show somatic loss of TSC1 or TSC2. **Genes Chromosomes Cancer** 41:80–85, 2004.
76. Lefebvre C, Clarke MJ: Identifying randomised trials, in Egger M, Davey Smith G, Altman DG (eds): *Systematic Reviews in Health Care: Meta-analysis in Context*. London, BMJ, 2001, pp 69–108.
77. Levin A: The Cochrane Collaboration. **Ann Intern Med** 135:309–312, 2001.
78. Lilford R, Braunholtz D, Harris J, Gill T: Trials in surgery. **Br J Surg** 91:6–16, 2004.
79. Loke YK, Price D, Derry S, Aronson JK: Case reports of suspected adverse drug reactions—Systematic literature survey of follow-up. **Br Med J** 332:335–339, 2006.
80. Lustig RA, Scott CB, Curran WJ: Does stereotactic eligibility for the treatment of glioblastoma cause selection bias in randomized studies? **Am J Clin Oncol** 27:516–521, 2004.
81. Macaskill P, Walter SD, Irwig L: A comparison of methods to detect publication bias in meta-analysis. **Stat Med** 20:641–654, 2001.
82. McLeod RS: Issues in surgical randomized controlled trials. **World J Surg** 23:1210–1214, 1999.
83. McLeod RS, Wright JG, Solomon MJ, Hu X, Walters BC, Lossing A: Randomized controlled trials in surgery: Issues and problems. **Surgery** 119:483–486, 1996.
84. Miettinen OS: *Theoretical Epidemiology*. New York, Delmar, 1985.
85. Miller FG: Sham surgery: An ethical analysis. **Am J Bioeth** 3:41–48, 2003.
86. Minor LB, Solomon D, Zinreich JS, Zee DS: Sound- and/or pressure-induced vertigo due to bone dehiscence of the superior semicircular canal. **Arch Otolaryngol Head Neck Surg** 124:249–258, 1998.
87. Mixter WJ, Barr JS: Rupture of the intervertebral disc with involvement of the spinal canal. **N Engl J Med** 211:210–215, 1934.
88. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP: Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? [see comments]. **Lancet** 352:609–613, 1998.
89. Murnaghan ML, Buckley RE: Lost but not forgotten: Patients lost to follow-up in a trauma database. **Can J Surg** 45:191–195, 2002.
90. Mylonas C: Research in neurosurgical journals. **Br J Neurosurg** 6:41–45, 1992.
91. Nelson HD, Humphrey LL, Nygren P, Teutsch SM, Allan JD: Postmenopausal hormone replacement therapy: Scientific review. **JAMA** 288:872–881, 2002.
92. O'Malley BW Jr, Grady MS, Gabel BC, Cohen MA, Heuer GG, Pisapia J, Bohman LE, Leibowitz JM: Comparison of endoscopic and microscopic removal of pituitary adenomas: Single-surgeon experience and the learning curve. **Neurosurg Focus** 25:E10, 2008.
93. Palesch YY, Tilley BC, Sackett DL, Johnston KC, Woolson R: Applying a phase II futility study design to therapeutic stroke trials. **Stroke** 36:2410–2414, 2005.
94. Pearce JM: Broca's aphasiacs. **Eur Neurol** 61:183–189, 2009.
95. Pham B, Platt R, McAuley L, Klassen TP, Moher D: Is there a "best" way to detect and minimize publication bias? An empirical evaluation. **Eval Health Prof** 24:109–125, 2001.
96. Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC: Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. **Int J Epidemiol** 36:847–857, 2007.
97. Plotkin SR, Stemmer-Rachamimov AO, Barker FG 2nd, Halpin C, Padera TP, Tyrrell A, Sorensen AG, Jain RK, di Tomaso E: Hearing improvement after bevacizumab in patients with neurofibromatosis type 2. **N Engl J Med** 361:358–367, 2009.
98. Pontiroli AE, Fossati A, Vedani P, Fiorilli M, Folli F, Paganelli M, Marchi M, Maffei C: Post-surgery adherence to scheduled visits and compliance, more than personality disorders, predict outcome of bariatric restrictive surgery in morbidly obese patients. **Obes Surg** 17: 1492–1497, 2007.
99. Prados MD, Gutin PH, Phillips TL, Wara WM, Sneed PK, Larson DA, Lamb SA, Ham B, Malec MK, Wilson CB: Interstitial brachytherapy for newly diagnosed patients with malignant gliomas: The UCSF experience. **Int J Radiat Oncol Biol Phys** 24:593–597, 1992.
100. Prasad KS, Gregson BA, Bhattathiri PS, Mitchell P, Mendelow AD: The significance of crossovers after randomization in the STICH trial. **Acta Neurochir Suppl** 96:61–64, 2006.
101. Reiser SJ: The clinical record in medicine. Part 1: Learning from cases. **Ann Intern Med** 114:902–907, 1991.
102. Robless P, Halliday A: Vascular surgical society of Great Britain and

Ireland: Carotid angiography is used more selectively in the asymptomatic carotid surgery trial. **Br J Surg** 86:690–691, 1999.
103. Rosling L, Rosling M: Pneumonia causes panic in Guangdong province. **Br Med J** 326:416, 2003.
104. Schierhout G, Roberts I: The Cochrane Brain and Spinal Cord Injury Group. **J Neurol Neurosurg Psychiatry** 63:1–3, 1997.
105. Schroeder BA, Wells RG, Starshak RJ, Sty JR: Clivus chordoma in a child with tuberous sclerosis: CT and MR demonstration. **J Comput Assist Tomogr** 11:195–196, 1987.
106. Schulz KF: Subverting randomization in controlled trials. **JAMA** 274:1456–1458, 1995.
107. Schulz KF, Chalmers I, Hayes RJ, Altman DG: Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. **JAMA** 273:408–412, 1995.
108. Selker RG, Shapiro WR, Burger P, Blackwood MS, Arena VC, Gilder JC, Malkin MG, Mealey JJ Jr, Neal JH, Olson J, Robertson JT, Barnett GH, Bloomfield S, Albright R, Hochberg FH, Hiesiger E, Green S: The Brain Tumor Cooperative Group NIH Trial 87-01: A randomized comparison of surgery, external radiotherapy, and carmustine versus surgery, interstitial radiotherapy boost, external radiation therapy, and carmustine. **Neurosurgery** 51:343–355; discussion 355–357, 2002.
109. Shapiro WR, Green SB, Burger PC, Selker RG, VanGilder JC, Robertson JT, Mealey J Jr, Ransohff J, Mahaley MS Jr: A randomized comparison of intra-arterial versus intravenous BCNU, with or without intravenous 5-fluorouracil, for newly diagnosed patients with malignant glioma. **J Neurosurg** 76:772–781, 1992.
110. Shrieve DC, Alexander E 3rd, Black PM, Wen PY, Fine HA, Kooy HM, Loeffler JS: Treatment of patients with primary glioblastoma multiforme with standard postoperative radiotherapy and radiosurgical boost: Prognostic factors and long-term outcome. **J Neurosurg** 90:72–77, 1999.
111. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J: Methodological index for non-randomized studies (minors): Development and validation of a new instrument. **ANZ J Surg** 73:712–716, 2003.
112. Solomon MJ, McLeod RS: Should we be performing more randomized controlled trials evaluating surgical operations? **Surgery** 118:459–467, 1995.
113. Stadhouder A, Buskens E, de Klerk LW, Verhaar JA, Dhert WA, Verbout AJ, Vaccaro AR, Oner FC: Traumatic thoracic and lumbar spinal fractures: Operative or nonoperative treatment: Comparison of two treatment strategies by means of surgeon equipoise. **Spine** 33: 1006–1017, 2008.
114. Stadhouder A, Oner FC, Wilson KW, Vaccaro AR, Williamson OD, Verbout AJ, Verhaar JA, de Klerk LW, Buskens E: Surgeon equipoise as an inclusion criterion for the evaluation of nonoperative versus operative treatment of thoracolumbar spinal injuries. **Spine J** 8:975–981, 2008.
115. Steinberg KK, Smith SJ, Stroup DF, Olkin I, Lee NC, Williamson GD, Thacker SB: Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies [see comments]. **Am J Epidemiol** 145:917–925, 1997.
116. Syin D, Woreta T, Chang DC, Cameron JL, Pronovost PJ, Makary MA: Publication bias in surgery: Implications for informed consent. **J Surg Res** 143:88–93, 2007.
117. Thompson SG: Why sources of heterogeneity in meta-analysis should be investigated. **Br Med J** 309:1351–1355, 1994.
118. Thompson SG, Sharp SJ: Explaining heterogeneity in meta-analysis: A comparison of methods. **Stat Med** 18:2693–2708, 1999.
119. Tierney JF, Stewart LA: Investigating patient exclusion bias in meta-analysis. **Int J Epidemiol** 34:79–87, 2005.
120. Tilley BC, Galpern WR: Screening potential therapies: Lessons learned from new paradigms used in Parkinson disease. **Stroke** 38:800–803, 2007.
121. Vandenbroucke JP: In defense of case reports and case series. **Ann Intern Med** 134:330–334, 2001.
122. Vandenbroucke JP: Observational research, randomised trials, and two views of medical science. **PLoS Med** 5:339–343, 2008.
123. Venning GR: Validity of anecdotal reports of suspected adverse drug reactions: The problem of false alarms. **Br Med J (Clin Res Ed)** 284:249–252, 1982.
124. Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA: The art of quality assessment of RCTs included in systematic reviews. **J Clin Epidemiol** 54:651–654, 2001.
125. Welling DB, Slater PW, Thomas RD, McGregor JM, Goodman JE: The learning curve in vestibular schwannoma surgery. **Am J Otol** 20:644–648, 1999.
126. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA: Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. **Br Med J** 336:601–605, 2008.
127. Yoshimoto Y: Publication bias in neurosurgery: Lessons from series of unruptured aneurysms. **Acta Neurochir (Wien)** 145:45–48, 2003.
128. Young J, Harrison J, White G, May J, Solomon M: Developing measures of surgeons' equipoise to assess the feasibility of randomized controlled trials in vascular surgery. **Surgery** 136:1070–1076, 2004.
129. Young JM, Solomon MJ, Harrison JD, Salkeld G, Butow P: Measuring patient preference and surgeon choice. **Surgery** 143:582–588, 2008.

