In 1996, *Lancet* editor Richard Horton angered a generation of academic surgeons by equating surgical research with “comic opera.” He challenged the field to improve; and an important part of this, he believed, would be for surgeons to turn away from the case series that traditionally cluttered our journals toward performing more randomized controlled trials (RCTs). In this issue of the *Journal*, Mansouri et al. report on their examination of the quality of reporting of a cohort of RCTs on neurosurgical procedures published between 2000 and 2014. Sixty-one RCTs published in a variety of neurosurgical and nonneurosurgical journals were found. The most frequent subject domains were vascular, functional, and tumor neurosurgery; spine surgery RCTs were excluded by design. Overall, the reporting quality of the RCTs was poor, particularly for those studies that were published in neurosurgical journals. Major deficiencies were present in reporting of blinding, allocation concealment, justification of sample size, and to a lesser degree, in completeness of subject follow-up. The number of RCTs published per year was low in comparison with other medical specialties. RCTs had high impact: on average they were cited more than 10 times per year since publication. The authors’ major conclusion was that neurosurgeons should explore alternate means of answering clinical questions, specifically registry studies.

The authors’ conclusions about the numbers of RCTs on neurosurgical topics and their quality of reporting are likely to be sound, both because of the careful methods used and because their findings are consistent with similar previously published studies. In 2004, Vranos et al. found 108 RCTs published during 1966–2004 that contained a neurosurgical procedure in at least 1 trial arm; approximately half addressed spinal surgery, so the rate of publication was less than 3 RCTs per year (slightly more than 1 per year, excluding spine studies). Mansouri et al. found an average of 5 RCTs per year from 2000 to 2014. In a study limited to 3 major neurosurgical journals, Gnanalingham et al. found a modest increase in RCTs reported, from 0% of all articles in 1982 to 2% in 2002, a trend confirmed by a 1966–2006 survey that used EMBASE as well as MEDLINE and was not limited to neurological journals. Although Mansouri et al. found no increasing trend in numbers of neurosurgical RCTs between 2000 and 2014, the longer-term trend does appear to be positive. This would be good news, because presently neurosurgery seems to have fewer RCTs in its literature and knowledge base than almost any other medical or surgical specialty. For example, a study on major orthopedic journals found 4%–6% of all articles to be RCTs (2006–2010), and total numbers of orthopedic RCTs found in a systematic review (2008–2011) were 40 per year. Cardiothoracic surgery journals contain about 6% RCTs, general surgery journals about 7%, and general medical journals about 12%. In specialty spine journals the rate is more than 4%. One consequence of this neurosurgical poverty is that our “meta-analyses” are less likely to include data from RCTs than those in any other specialty. The same is probably true for our daily clinical decisions.

So neurosurgical RCTs are rare birds; but can we perhaps be happy with their quality, if not their quantity? Here, too, we will be disappointed. Neurosurgical RCTs had low quality scores on two widely-used metrics, especially when published in our own journals. Again, while direct cross-study comparisons are not possible for these scores, due to subjectivity in the scoring scales and incomplete reporting of results, prior studies of RCTs in neurosurgery and spine surgery also showed major deficiencies in RCT quality measures. In some contexts, it has been shown that “bad reporting does not mean bad methodology,” i.e., authors used high quality methods in their trials but failed to report them. Notably, though, these studies originated in US Cooperative Cancer Groups in which tight supervision of trial quality is structurally guaranteed. It is more likely that the deficiencies noted by Mansouri et al. are real, but journals could address this by publishing trial protocols as a supplement to trial reports. Still worse news is that Mansouri et al. scored neurosurgical
RCTs using a scale based on the standard Consolidated Standards for Reporting of Trials (CONSORT) criteria, not the more rigorous nonpharmacological extension specifically designed for trials that evaluate procedures. If the nonpharmacological scale had been used, the deficiencies found would likely have been even deeper and more widespread. The extra criteria in the CONSORT extension address factors such as skill of participating surgeons and completeness of description of the procedure studied, and any measures of compliance with the protocol definition of the procedure. The deficiencies found by Mansouri and coauthors in neurosurgical RCTs affected blinding (adequately performed and reported in less than 10% of neurosurgical RCTs), adequate concealment of treatment allocation (important to prevent investigators from cheating on randomization; adequate in 23%–50% of trials), and sample size calculation (adequate in 20%–38% of trials). There is objective evidence that failures in blinding and allocation concealment are associated with biased results of RCTs.

Factors that make surgical RCTs more difficult to perform than medical trials are well known, and some of these factors can help explain this study’s findings. Blinding is a particular challenge in surgical trials and had the lowest compliance rate of the CONSORT criteria in this RCT cohort. Yet surgery has a powerful placebo effect that should not be ignored in trial design, particularly for this RCT cohort. However, registries capture results of standard medical practice, in which patient and physician choices determine treatment selection. This complex process invariably creates treatment and comparison “control” groups that are imbalanced for every variable we could possibly measure (as statistician Marks Nester’s “applied statistician’s creed” reminds us, “no two populations are identical in any respect”). Some of these baseline imbalances can be ignored, because they are in variables that are not prognostic for outcome (for example, zodiac sign). Some are known prognostic variables, and we can plan ahead in a registry to collect the necessary information to adjust our outcomes using multivariate risk adjustment or propensity scores to make a fair treatment comparison.

Treatment groups in registries are also unbalanced for unknown prognostic factors as well, and statistical techniques cannot adjust these differences away (unless the unknown confounders are perfectly correlated with known factors, an assumption too strong to be plausible). For example, patients with glioblastoma who are married live about 20% longer, and those undergoing both surgery and radiation are more likely to be married: living more than 15 miles from the treatment center carried a risk of death about two-thirds lower than those who lived closer in a review of cancer treatment trials, largely using ineffective drugs; and eligibility for treatment accounted entirely for the apparent treatment benefit of brachytherapy for glioblastoma when finally tested using RCTs. Such prognostic factors can be too difficult to define or too expensive to archive in a registry, or so unexpected that we do not consider collecting them at all. The difficulty of defining some hidden confounders makes them practically intangible. For example, patients who adhere better to prescribed placebos—ineffective by definition—have significantly lower rates of cardiac mortality in myocardial infarction prevention trials, less bone loss, lower rates of hip fracture in fracture prevention trials, and lower mortality from cancer and from all causes in hormone replacement trials. How can we measure this quality in a surgical registry? Randomization, however, balances every prognostic factor, both known and unknown. This is why effective randomization is of paramount importance in an unbiased treatment comparison.

Some additional problems with neurosurgical RCTs, including some of those identified by Mansouri et al., are not really avoided in registry studies either. For example, if blinding and allocation concealment are poorly conducted in neurosurgical RCTs, they are never part of a registry study at all. Crossovers in an RCT have their equivalent in every patient in a registry, since treatment selection by patient and physician choice is the norm in standard practice. Even the difficulty establishing equipoise necessary for an RCT is not such a clear advantage for registries. If there truly is no equipoise in a community for a treatment decision, registry comparisons will be "apples..."