Randomized controlled trials and neurosurgery: the ideal fit or should alternative methodologies be considered?

Alireza Mansouri, MD,1,2 Benjamin Cooper, MD,4 Samuel M. Shin, MD,4 and Douglas Kondziolka, MD, MSc3,4

1Division of Neurosurgery, University of Toronto, Ontario, Canada; 2Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada; 3Departments of Neurosurgery and 4Radiation Oncology, NYU Langone Medical Center, New York University, New York, New York

OBJECTIVE Randomized-controlled trials (RCTs) are advocated to provide high-level medical evidence. However, in neurosurgery, there are barriers to conducting RCTs. The authors of this study sought to analyze the quality of neurosurgical RCTs since 2000 to determine the adequacy of their design and reporting.

METHODS A search of the MEDLINE and EMBASE databases (2000–2014) was conducted. The medical subject heading (MeSH) terms used in the search included: “neurosurgery” OR “neurosurgical procedure,” “brain neoplasms,” “infarction” and “decompression,” “carotid stenosis,” “cerebral hemorrhage,” and “spinal fusion.” These studies were limited to RCTs, in humans, and in the English language. The Consolidated Standards for Reporting of Trials (CONSORT) and Jadad scales were used to assess the quality of RCT design and reporting. The standardized median times cited (median citations divided by years since publication) were used to assess impact. A pragmatic-explanatory continuum indicator summary–based scale was used to assess the design of the studies as primarily pragmatic or explanatory.

RESULTS Sixty-one articles were identified, and the following subspecialties were the most common: vascular (23, 37%), followed by functional neurosurgery and neurooncology (both 13, 21%). The following nations were the primary leaders in RCTs: US (25 studies, 41%), Germany (8 studies, 13%), and the United Kingdom (7 studies, 11%). Median sample size was 100 (interquartile range [IQR] 41.5–279). The majority of the studies (40, 66%) had pragmatic objectives. The median number of times cited overall was 69 (IQR 20.5–193). The combined median CONSORT score was 36 (IQR 27.5–39). Blinding was most deficiently reported. Other areas with a relatively low quality of reporting were sample size calculation (34.2% of surgical, 38.5% of drug, and 20% of device studies), allocation concealment (28.9% of surgical, 23.1% of drug, and 50% of device studies), and protocol implementation (18.4% of surgical, 23% of drug, and 20% of device studies). The quality of reporting did not correlate with the study impact. All studies had a median Jadad score ≤ 3. Thirty-three pragmatic studies (83%) and 5 explanatory studies (25%) met the design objectives. All pragmatic studies based on drug and device trials met their objectives, while 74% of pragmatic surgical trials met their objectives.

CONCLUSIONS The prevalence of neurosurgical RCTs is low. The quality of RCT design and reporting in neurosurgery is also low. Many study designs are not compatible with stated objectives. Pragmatic studies were more likely to meet design objectives. Given the role of RCTs as one of the highest levels of evidence, it is critical to improve on their methodology and reporting.


KEY WORDS randomized controlled trial; trial design; publication; CONSORT
The randomized-controlled trial (RCT) is considered to provide one of the highest levels of medical evidence. In particular, RCTs based on a large sample of participants are considered to be of even higher quality and have thus been favored by many epidemiologists. Since their establishment, RCTs have helped answer many important clinical questions through rigorous methods that reduce bias.

One of the pillars of an RCT is the presence of equipoise, which renders the conduct of treatment randomization ethical. In surgical specialties such as neurosurgery, however, decision-making is often based on the lessons taught by those in a position of authority or personal experience, considered to be Level III–V evidence. Therefore, among individuals, clear equipoise often does not exist. Other pertinent issues, such as the concept of the technical learning curve, standardization of a surgical intervention, and patient recruitment are also genuine concerns with surgical RCTs.

Although a well-designed RCT has the potential to overcome many of these barriers, this goal cannot always be attained. In addition to significant resource requirements, rigorous trial design can potentially limit the applicability of its findings to the general population. Pragmatic RCTs (PCTs), as opposed to explanatory (mechanistic) trials, are designed to maintain the prognostic balance established by RCTs while permitting other aspects to proceed as routine. These types of trials represent an alternative strategy. Their flexible nature, however, has been shown to often result in an inadvertent breakdown of randomization, and thus decrease the value of the trial. Others suggest that these limitations should be accepted as realities. Any shortcomings can be remedied by implementing reporting criteria trials that can help readers identify the quality of the evidence.

The Consolidated Standards of Reporting Trials (CONSORT) is one such example. Recent assessments of neurosurgical trials with regard to the quality of reporting using various scales, however, have noted that the overall quality is rather low but perhaps improving over time. This is likely an issue related to both a lack of awareness of these reporting criteria and the inherently flawed design of many neurosurgical RCTs.

While RCTs are and will continue to provide some of the highest level of evidence, certain scenarios call for alternative study design strategies. Rather than establish rigid guidelines of design and reporting to create a better match between neurosurgery and RCTs, it may be worthwhile to accept this difference and conduct clinical studies based on designs that reflect the constraints of a particular clinical question. Indeed, the “pragmatic” component of the phrase “pragmatic clinical trials” has served to emphasize the practical and sensible solution that PCT design offers to address some of the logistic barriers posed by the traditional RCT design. Furthermore, many fields in neurosurgery (such as benign primary tumors) have lacked the benefit of evidence-based medicine derived from RCTs, and current practice is, at the very best, based on prospective observational series. These studies are often limited by very small sample sizes, preventing definitive conclusions. Therefore, in certain clinical scenarios, the pragmatic approach to such shortcomings could perhaps be through the use of multicenter, registry-based collection of prospective observational data. Thus, pragmatic registry-based observational studies (PROS), designed through rigorous methodology, can serve as a viable substitute in clinical scenarios in which a well-designed and conducted RCT is simply not feasible. The collection of high-quality prospective data has been facilitated through the implementation of registries, contributing greatly to clinical data in a variety of medical specialties.

In this study, we sought to identify the main surgical RCTs conducted with relevance to the field of neurosurgery since the year 2000. Our objectives were to critically appraise the design and reporting of RCTs in neurosurgery, and to correlate the design of the study (whether explanatory or pragmatic) with its specific objectives to identify the extent of inconsistency. Furthermore, we outline potential strategies to improve the quality of the design and conduct of PROS for situations in which RCTs are not feasible.

Methods

Search Strategy

A search of the MEDLINE and EMBASE (2000 to the present) online databases was initiated on September 9, 2014, to identify RCTs in neurosurgery. The choice of the earlier date boundary was based on the demonstrated improvement in the indexing of studies in databases such as MEDLINE, as a result of the findings and recommendations by Dickersin et al. and Evans and to include a sampling of relatively recent studies. The search was restricted to the English-speaking literature and studies pertaining to humans only. The initial search contained the medical subject heading (MeSH) terms “neurosurgery” or “neurosurgical procedure” and the results were limited to “randomized controlled trials” as the publication type. An analysis of the results from this search suggested that some of the key neurosurgical RCTs had not been identified. Therefore, on September 19, 2014, the following MeSH terms were searched to identify additional studies: “brain neoplasms” (relevant to the surgical aspect of neurooncology), “infarction” and “decompression” (relevant to the overlap between the field of neurology, critical care, and neurosurgery), “carotid stenosis” (relevant to neurosurgeons involved in open and/or endovascular management of carotid disease), “cerebral hemorrhage” (relevant to the overlap between the field of neurology, critical care, and neurosurgery), and “spinal fusion” (primarily relevant to the influx of new devices into this subspecialty). Similar to before, the results for each of these categories were limited to “randomized controlled trials” as the publication type.

The search results were imported into EndNote (version X7.2) and duplicates were removed. Titles and abstracts were reviewed by 1 author (A.M.) to identify relevant neurosurgical studies. Incomplete studies, interim reports of longer trials, and published protocols without any pertinent patient data were excluded. In the case of studies published with varying follow-up durations or different analyses of the same trial, only the initial study was considered; this resulted in the exclusion of 9 studies from...
the final list. To focus the results of our analysis on the quality of reporting of RCTs pertaining to surgical interventions in neurosurgery, studies explicitly pertaining to physiotherapy, pain management, or anesthesia were not included. In “brain neoplasms,” only studies assessing at least 1 form of neurosurgical intervention were included. Therefore, studies solely focused on chemotherapy or radiation-based therapies other than radiosurgery were not included. In “carotid stenosis” only studies assessing at least 1 type of carotid stenting or endarterectomy were included. In “cerebral hemorrhage,” studies assessing solely the role of conservative measures such as cooling were not included. Lastly, in “spinal fusion,” studies pertaining solely to pain management, rehabilitation, and physiotherapy were not included. These strategies were implemented to ensure that RCTs assessing a surgical intervention in at least 1 of the arms of the study were captured. A summary of the search strategy and results is summarized in Fig. 1.

Studies were categorized as “surgical,” “drug,” or “device,” depending on the trial question. The following information was extracted from each individual trial: year of publication, journal name, country of the study’s lead investigator, the subspecialty to which the trial pertained, the sample size, whether the study was funded by industry (if not clear, “unknown” was indicated), whether the study was a multicenter study, and the number of times the study was cited. The Web of Science electronic database was used to obtain the latter.

**Study Assessment**

The full text of the relevant studies was reviewed to determine the revised CONSORT and Jadad scores as an assessment of the quality of the reporting of these RCTs. For each of the 22 categories of the CONSORT scale, 2 points were given if the category was addressed fully, 1 point if only partially addressed, and 0 points if the category was not addressed at all; through this strategy, the range of possible CONSORT scores for a given study would be 0 to 44. The CONSORT statement was reviewed by the 3 independent assessors (A.M., B.C., and S.M.S.) prior to scoring of studies to ensure a unified understanding of the headings and appropriate scoring strategy. For each category, the median score obtained by the 3 reviewers was used as the overall score. The Spearman’s correlation analysis was used to assess the interobserver variability for the overall CONSORT scores. A correlation coefficient greater than 0.8 was considered to be satisfactory. One reviewer (A.M.) assessed the studies for the Jadad score. In this 3-category scale, 1 point was given if randomization was noted along with an additional point if the method chosen was appropriate; a point was deducted if the method chosen was inappropriate. Similarly, 1 point was given if blinding was noted, along with an additional point if the method chosen was appropriate; a point was deducted if the method chosen was inappropriate. For the final category, a point was given if all patients initially enrolled had been accounted for. Thus, the range of possible Jadad scores would be 0 to 5.
The 2013 impact factors for the journals publishing the manuscripts identified by our search were identified through an open access website that compiles journal citation report factors for convenient access (http://www. impact-factor.org). Five of these results were selected at random and the impact factor values were confirmed with an independent search to assure the validity of the compiled data, which was found to be true. In 1 of the search results, the original journal Surgical Neurology changed names to World Neurosurgery; for this paper we used the 2009 impact factor, which was the latest available value. The Spearman’s correlation analysis was used to determine the correlation of journal impact factor and the standardized median times cited on the total CONSORT score. The latter factor has been used previously as a surrogate measure for the overall impact of the particular RCT on the medical literature pertaining to the field.30 The Mann-Whitney U-test for nonparametric variables was used to determine a statistical difference between the median score for articles published in the top 3 neurosurgical journals compared with the top 3 medical journals. A p value < 0.05 was considered statistically significant. The statistical program SPSS (version 22.0, IBM) was used for all statistical analyses.

The nature of the question for each trial was used by 1 author (A.M.) to determine whether the objective of the trial was explanatory or pragmatic. Subsequently, an arbitrary scale based on the 10-item pragmatic-explanatory continuum indicator summary (PRECIS) tool31 was used to determine the position of the study on the pragmatic-explanatory continuum as per its actual design. In this scale, all studies would have a starting score of 10. A point was deducted if the design of the study was pragmatic for the given PRECIS item while a point was added if the design was explanatory for the given item. Thus, a fully pragmatic trial would have a score of 0 while a fully explanatory trial would have a score of 20. While the original article by Thorpe and colleagues31 provides a comprehensive overview of the criteria used in this scoring strategy, a pragmatic trial would, in brief, be one wherein: 1) all potential participants are enrolled regardless of baseline risk (for example, no run-in phase has been implemented); 2) the experimental and comparator intervention can be applied according to the best judgment of the clinician without the need for specific expertise or training; 3) the primary outcome of interest can be objectively measured; 4) there are limited to no follow-up/compliance/protocol adherence assessments; and 5) the final analysis is made on an intent-to-treat basis. An example of a primarily pragmatic RCT is the International Surgical Trial in Intracerebral Hemorrhage (STICH) study18 in which all patients with a spontaneous supratentorial hemorrhage within the past 72 hours were eligible and the comparators were best medical therapy versus surgical evacuation of hematoma, both implemented according to the best judgment of the managing clinician. The outcome measured here was objective (death or disability based on the Glasgow Outcome Scale) and it was ascertained primarily through mailed questionnaires with or without confirmation of living status without the need for a formal follow-up. No formal assessment of compliance/adherence was conducted and the final analysis was performed on an intent-to-treat basis. Note that the addition and subtraction of points in this scale is completely arbitrary and done solely to create a sense of the overall directionality for the given trial. The proportion of trial designs meeting the study objective were categorized by the study category.

### Results

Upon conducting the outlined search of the literature, 61 suitable articles were identified (Table 1). The range of RCTs per year was roughly similar, averaging approximately 5 studies and peaking at 8 studies in 2007 (Fig. 2). Studies pertaining to vascular neurosurgery comprised 23 of the studies (37%), followed by studies in functional neurosurgery with an additional 13. Of the included studies, 3 were multicenter trials, with the remaining studies being single-site. Studies conducted in the United States comprised the majority of studies (24 of 61; 39%), followed by 14 studies in Europe (23%), 6 in the United Kingdom, 5 in Canada, and 3 in Australia. Of the studies, 12 were industry funded and 49 were not. The median sample size was 32 patients (IQR, 25–69); 40 of 61 included 100 or fewer patients, and 21 studies were multicenter. The median time cited was 69 (IQR, 20.5–193) days, with a median standardized time cited of 10.3 (IQR, 4.0–32). The median SORT score was 29 (IQR, 24–33). Of the studies, 48 were explanatory, 10 pragmatic, and 3 of unknown design.
rosurgery and neurooncology, with 13 studies (21%) each. The majority of RCTs in neurooncology were based on brain metastases and high-grade gliomas; low-grade primary tumors were drastically underrepresented. Surgical studies were found to be the most common type (38 studies, 62%). One-third of the studies were published in neurosurgical journals such as the Journal of Neurosurgery (JNS) with 5 studies (25%), Neurosurgery with 4 studies (20%), and Spine with 2 studies (10%). Among the studies published in nonneurosurgical journals, 6 (15%) were in The Lancet, 5 (12%) were in The New England Journal of Medicine (NEJM), and 2 (5%) were in The Journal of the American Medical Association (JAMA). The US (25 studies, 41%), Germany (8 studies, 13%), and the United Kingdom (7 studies, 11%) were the top 3 nations contributing to this literature. The median sample size for all studies was 100 patients (interquartile range [IQR] 41.5–279); more than half of the studies contained a sample size smaller than 100 patients. The majority of the studies (45, 74%) were multicenter and not funded by industry (36, 59%). Based on the study questions, the majority of the studies (40, 66%) appeared to have pragmatic objectives. The median number of times cited overall was 69 (IQR 20.5–193).

Each study was assessed independently to identify the CONSORT score for each category. The interobserver agreement was strong (Spearman’s correlation coefficients: A.M. and S.M.S. = 0.81; A.M. and B.C. = 0.80; B.C. and S.S. = 0.92). The combined median CONSORT score was 36 (IQR 27.5–39). The extent of fulfillment of each item in the CONSORT scale, based on study type, is summarized in Table 2. The areas with the greatest deficiency regarding quality of reporting pertained to blinding, in which a large proportion of studies did not fulfill any of the criteria; surgical studies, in particular, failed to report blinding in 65.8% of studies. Other areas with a relatively low quality of reporting with none of the criteria met pertained to the categories of sample size calculation and justification (34.2% of surgical, 38.5% of drug, and 20% of device studies), allocation concealment (28.9% of surgical, 23.1% of drug, and 50% of device studies), and protocol implementation (18.4% of surgical, 23% of drug, and 20% of device studies). The majority of studies were of good quality with regard to reporting clearly defined eligibility criteria, study objectives, and statistical methods. In addition, the baseline characteristics of the patients were in general well-documented, along with a high frequency of reporting all adverse events.

Spearman’s correlation coefficient between total CONSORT score and the journal impact factor was 0.66, reflecting a strong correlation. The top 3 medical journals, based on impact factor, were NEJM (54.42), Lancet (39.2), and JAMA (29.98; Table 3). The top 3 neurosurgical journals were JNS (3.2), Neurosurgery (3.0), and Spine (2.45). The median CONSORT score of articles published in these medical journals was significantly greater than that of studies published in the neurosurgical journals (Table 3). The standardized median times cited was not correlated with the median CONSORT score.

The major design components of all study types, based on the median score of the Jadad items of appropriateness of randomization and blinding methods, along with accounting for all participants, were weak across various types of trials, with each having a median Jadad score of 3 or less (Table 4). Among all study types, many of the trials either did not involve blinding or failed to implement the methods appropriately. Although the sample size of device studies may have been too low to draw firm conclusions, these study types appeared to be associated with the weakest design quality overall.

The median PRECIS-based score for studies with pragmatic objectives was 6 (IQR 4–6), and was 8 (IQR 6–11.5) for studies with explanatory objectives. While this difference was significant, a substantial amount of overlap between the scores was observed (Fig. 3 left). Among the 40 pragmatic studies, 33 (83%) met the design objectives; 3 of these (9%) were overall neutral (PRECIS-based
## TABLE 2. Percentage of neurosurgical RCTs meeting items in the CONSORT checklist* (continued)

<table>
<thead>
<tr>
<th>CONSORT Variable</th>
<th>Surgical Studies</th>
<th>Drug Studies</th>
<th>Device Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant flow</td>
<td>68.4</td>
<td>69.2</td>
<td>60</td>
</tr>
<tr>
<td>Recruitment</td>
<td>23.7</td>
<td>15.4</td>
<td>40</td>
</tr>
<tr>
<td>Baseline data</td>
<td>97.4</td>
<td>84.6</td>
<td>80</td>
</tr>
<tr>
<td>Numbers analyzed</td>
<td>57.9</td>
<td>69.2</td>
<td>30</td>
</tr>
<tr>
<td>Outcomes &amp; estimation of effect size &amp; precision</td>
<td>52.6</td>
<td>69.2</td>
<td>30</td>
</tr>
<tr>
<td>Ancillary analyses</td>
<td>86.8</td>
<td>76.9</td>
<td>30</td>
</tr>
<tr>
<td>Adverse events</td>
<td>81.6</td>
<td>61.5</td>
<td>90</td>
</tr>
<tr>
<td>Data interpretation</td>
<td>55.3</td>
<td>84.6</td>
<td>70</td>
</tr>
<tr>
<td>Generalizability</td>
<td>44.7</td>
<td>76.9</td>
<td>50</td>
</tr>
<tr>
<td>Overall interpretation in the context of available evidence</td>
<td>78.9</td>
<td>84.6</td>
<td>70</td>
</tr>
</tbody>
</table>

* All data given as percentages unless otherwise indicated.

In the CONSORT checklist, the percentage of neurosurgical RCTs meeting items in the checklist was assessed across different study types. The median total score was calculated for each item, with the interquartile range (IQR) provided. The table highlights the percentage of studies meeting criteria in various categories such as title & abstract, introduction & background, participants, interventions, objectives, outcomes, sample size calculations & justifications, random sequence generation, allocation concealment, implementation of methods, blinding, and statistical methods. For example, in the category of participant flow, 68.4% of surgical studies, 69.2% of drug studies, and 60% of device studies met all criteria. The overall design of studies was more consistent with a pragmatic study (Fig. 3 right), indicated by higher percentages of pragmatic studies (80%) compared to explanatory studies (25%) for most items. All pragmatic studies based on drug and device trials met a score of 10, and 5 (15%) had an overall design that was more consistent with an explanatory study. Among the 21 explanatory studies, 5 (24%) met the design objectives; 4 (19%) were overall neutral, and 11 (52%) had an overall design that was more consistent with a pragmatic study (Fig. 3 right). Irrespective of the study type (surgical, drug, or device), a greater proportion of pragmatic studies were designed to meet their objectives as compared with the explanatory studies (80% vs 25%; Fig. 3 right).
perhaps alternative trial designs that are complementary to neurosurgical specialty and RCT trials, suggesting that some of the shortcomings in the match made between the actual design have also been identified; many studies with justifications for sample sizes emerging as major issues. Furthermore, discrepancies between study objectives and medical RCTs, with blinding, allocation concealment, and regard to both the design and the reporting of neurosurgical RCTs is low in comparison with other medical fields in general. We have demonstrated areas of deficiency with certain topics (such as pain management) were excluded. This information highlights explanatory/mechanistic objectives failed to meet this goal. In this study, we have conducted a high-yield search of the English literature from the year 2000 onwards to identify and assess the quality of neurosurgical RCTs. This was not a systematic review of the relevant literature, and certain topics (such as pain management) were excluded. Our analysis has shown that the prevalence of neurosurgical RCTs was significantly higher than that of the top 3 neurosurgical journals (p < 0.05). * The combined median CONSORT score of the top 3 medical journals was 604 significantly higher than that of the top 3 neurosurgical journals (p < 0.05).

**Discussion**

In this study, we have conducted a high-yield search of the English literature from the year 2000 onwards to identify and assess the quality of neurosurgical RCTs. This was not a systematic review of the relevant literature, and certain topics (such as pain management) were excluded. Our analysis has shown that the prevalence of neurosurgical RCTs is low in comparison with other medical fields in general. We have demonstrated areas of deficiency with regard to both the design and the reporting of neurosurgical RCTs, with blinding, allocation concealment, and justifications for sample sizes emerging as major issues. Furthermore, discrepancies between study objectives and actual design have also been identified; many studies with explanatory/mechanistic objectives failed to meet this goal with regard to their design. This information highlights some of the shortcomings in the match made between the neurosurgical specialty and RCT trials, suggesting that perhaps alternative trial designs that are complementary to RCTs rather than serving as substitutions (such as PROS) should be considered and implemented.

The most common RCTs were in the subspecialty fields of vascular, oncological, and functional neurosurgery. This is likely both a reflection of the higher prevalence of disorders in these fields and the overlap with other medical specialties such as neurology and medical/radiation oncology. In neurooncology, no surgical RCTs pertaining to primary low-grade lesions such as meningiomas, vestibular schwannomas, or low-grade gliomas were identified. This is perhaps a reflection of both the lower prevalence and the longer time course needed to determine outcomes, compared with metastases and high-grade gliomas. While spinal disorders are relatively common in the general population, there is little overlap between neurosurgery and other medical specialties, particularly given that we had opted to not include studies on pain management and physiotherapy. Similar to previous findings, the US, Germany, and United Kingdom continue to be the greatest contributors of RCTs; this is likely a reflection of a large population base, higher standards for conducting trials, and greater resources for conducting trials. Although the number of RCTs with relevance to neurosurgery is likely higher than that found in our study, the trend of approximately 4–5 major RCTs per year without any evidence of growth, and the steady pattern noted through a comparison with prior assessments of the neurological literature, is concerning. This is in stark contrast to the overall landscape in the literature in which RCTs in general are being published at an approximate rate of 200 per week, according to an estimate in the year 2000; neurosurgery clearly lags behind. In terms of the quality of reporting, however, similar analyses to ours have identified significant deficiencies in other specialties such as oncology, gynecology, orthopedic surgery, and plastic surgery.

Bhandari and colleagues assessed the quality of RCTs in orthopedics using a scale different from CONSORT and found shortcomings in the quality of reporting blinding and allocation concealment. A study by Taghinia and colleagues assessed the quality of RCTs in plastic surgery, using the CONSORT statement, over a 20-year period (1986–2006) and found that the majority of the tri-

**TABLE 3. Comparison of the medical and neurosurgical journals with the top 3 highest impact factors using the combined median CONSORT score**

<table>
<thead>
<tr>
<th>Journal</th>
<th>Impact Factor</th>
<th>No. of Studies Contributed</th>
<th>Journal Median CONSORT Score</th>
<th>Combined Median CONSORT Scores*</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEJM</td>
<td>54.42</td>
<td>5</td>
<td>38</td>
<td>39</td>
</tr>
<tr>
<td>Lancet</td>
<td>39.2</td>
<td>6</td>
<td>38.5</td>
<td>39</td>
</tr>
<tr>
<td>JAMA</td>
<td>29.98</td>
<td>2</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>JNS</td>
<td>3.2</td>
<td>5</td>
<td>30.5</td>
<td>30</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>3.0</td>
<td>4</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>Spine</td>
<td>2.45</td>
<td>2</td>
<td>33.5</td>
<td>33.5</td>
</tr>
</tbody>
</table>

* The combined median CONSORT score of the top 3 medical journals was significantly higher than that of the top 3 neurosurgical journals (p < 0.05).

**TABLE 4. Percentage of neurosurgical RCTs meeting items in the Jadad checklist**

<table>
<thead>
<tr>
<th>Jadad Variable</th>
<th>Surgical Studies</th>
<th>Drug Studies</th>
<th>Device Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median total score, max score 5 (IQR)</td>
<td>3 (2–4)</td>
<td>3 (2–3)</td>
<td>2 (1.75–3)</td>
</tr>
<tr>
<td>Randomization (max score 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Randomized &amp; method appropriate</td>
<td>60.6</td>
<td>61.5</td>
<td>40</td>
</tr>
<tr>
<td>Randomized, no elaboration of method</td>
<td>36.8</td>
<td>38.5</td>
<td>60</td>
</tr>
<tr>
<td>Inappropriate method of randomization</td>
<td>2.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Blinding (max score 2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blinded &amp; method appropriate</td>
<td>10.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Blinded, no elaboration of method</td>
<td>31.6</td>
<td>7.7</td>
<td>40</td>
</tr>
<tr>
<td>Inappropriate method of blinding</td>
<td>57.9</td>
<td>92.3</td>
<td>60</td>
</tr>
<tr>
<td>Account of all patients (max score 1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All patients accounted for</td>
<td>92.1</td>
<td>84.6</td>
<td>40</td>
</tr>
<tr>
<td>All patients not accounted for</td>
<td>7.9</td>
<td>15.4</td>
<td>60</td>
</tr>
</tbody>
</table>

* All data given as percentages unless otherwise indicated.
als did not adequately report allocation concealment and sample size calculations/justifications; in fact, these fields were rated lower in this review than our assessment of neurosurgical RCTs. The quality of the methodology, assessed through the Jadad score, was low in these studies as well. Therefore, it is clear that the issue of poor quality reporting and methodology of RCTs is pervasive and applies to many medical specialties. In surgical specialties, the emerging theme appears to be the issue of sample size calculation/justification, blinding, and allocation concealment. The causes for these deficiencies are multifactorial.

Many of the concepts in surgery in general, but particularly neurosurgery, have been passed down from mentors to mentees without significant consideration of possible equipoise between alternative management strategies for an individual practitioner. Therefore, from the perspective of principle, the conduct of a trial to assess the difference between two interventions would be unethical. Although Freedman has argued that equipoise within the clinical community is sufficient to justify a trial, the inherent bias associated with conducting a trial by individual surgeons cannot be ignored. In our assessment, two-thirds of surgical trials and more than 30% of drug/device trials did not meet any of the CONSORT criteria for reporting on blinding. While some of the trials had simply not noted their methods of blinding, a concerning majority had simply implemented an incorrect method of blinding based on the Jadad criteria. Although the quality of reporting on allocation concealment was slightly better, a minority of trials had met all of the CONSORT criteria for reporting on blinding. While some of the trials had simply not noted their methods of blinding, a concerning majority had simply implemented an incorrect method of blinding based on the Jadad criteria. Although the quality of reporting on allocation concealment was slightly better, a minority of trials had met all of the CONSORT criteria for reporting on blinding. While some of the trials had simply not noted their methods of blinding, a concerning majority had simply implemented an incorrect method of blinding based on the Jadad criteria. Although the quality of reporting on allocation concealment was slightly better, a minority of trials had met all of the CONSORT criteria for reporting on blinding. While some of the trials had simply not noted their methods of blinding, a concerning majority had simply implemented an incorrect method of blinding based on the Jadad criteria.

The median sample size for all trials in our study was relatively low, with the majority of studies based on samples smaller than 100. In contrast to conditions in other medical specialties addressed through RCTs, such as infectious diseases and chronic illnesses, the neurosurgical specialty is confronted with relatively more rare conditions. Combined with the fact that clinical epidemiologists consider larger RCTs with narrower confidence intervals to be of greater importance and value, many neurosurgical questions are either not amenable to an RCT or their quality is considered too low to be of much value. It may be argued that neurosurgical RCTs should be based on more rigorous design, and neurosurgical jour-
nals, the primary source of knowledge for the neurosurgical community, should enforce reporting guidelines such as CONSORT. In one of the most rigorously designed RCTs applicable to neurosurgery, the North American Symptomatic Carotid Endarterectomy Trial (NASCET), surgeons selected to participate were screened based on their perioperative morbidity and mortality rates prior to participation. While this ensures patient safety within the trial and meets the ethical objectives of research, the findings of such a design may not be generalizable to every neurosurgeon performing a carotid endarterectomy in the community. Furthermore, a well-designed RCT requires a significant investment of essential resources. The overall time lag between the inception of a trial idea and the publication of its individual or pooled results can be significant, potentially limiting the applicability of its findings at the time of publication.

The concern regarding the time lag from inception of an idea (such as a novel device or a new surgical technique) to its establishment as the standard of care should not be interpreted as a need to forego rigorous criteria, such as those demanded by the US FDA, Health Canada, and the European Medicines Agency. These safeguards have been established to deliver high-quality care with patient safety as the highest priority. Such safety and efficacy studies are most appropriately addressed through explanatory RCTs, and other designs are likely to be a compromise in quality. An example of an explanatory RCT design that has significantly impacted the standard of care for carotid stenosis has been the NASCET trial (not in our cohort). Once a device or a surgical technique has been approved by the FDA (as an American example), however, other agencies such as the Centers for Medicare and Medicaid Services demand the evidence that highlights patient-important outcomes, cost-effectiveness, and relevance to daily practice. To provide such evidence in favor of an intervention’s effectiveness, a pragmatic design objective would be more appropriate.

The majority of the studies identified in our search were judged to have a pragmatic objective. Based on a PRECIS-based scale, many of these studies were designed to meet their objectives while the cohort of studies with explanatory objectives was generally not designed for this purpose. While it is recognized that the pragmatic-explanatory concept is a continuum rather than a dichotomous variable, our observation likely reflects the general objective and reality of neurosurgical trials in that they tend to have a pragmatic objective and are better suited to meet that objective. Some of these pragmatic studies (such as STICH) have indeed affected the practice of neurosurgery pertaining to the surgical management of spontaneous intracranial hematomas. However, with obvious exceptions, successfully devising a protocol that appears to effectively meet the objectives does not imply that the methodology and reporting quality of these PCTs are not deficient. Considering the term “pragmatic” as a strictly methodological definition applicable to any design that is practical and reflective of the true landscape of the field, a well-designed and implemented pragmatic observational study would be a realistic adaptation to the issue relevant to RCTs in various surgical specialties, including neurosurgery. Combined with the aforementioned shortcomings of PCTs and traditional RCTs, it may be beneficial to consider PROS as a design that is better suited to a great proportion of neurosurgical questions, and therefore to consider it as a complementary approach when RCTs are not feasible.

Through the implementation of registries, PROS can be particularly helpful in establishing the effectiveness of a given intervention or to provide surveillance of harm in the case of rare events; this would provide relevant evidence to the Centers for Medicare and Medicaid Services and equivalent agencies. The widespread acceptance of prospective observational studies has led to the establishment of guidelines for the design of prospective registries. Furthermore, criteria for the design and evaluation of prospective observational studies have also been established. Registries that are amenable to multicenter studies are particularly helpful.

Multicenter PROS enable the analysis of outcome from a greater pool of patients who are managed based on the routine practice of the selected centers. This approach overcomes some of the inherent limitations of randomizing a large sample of patients for an RCT. Furthermore, this design seeks to obtain answers that are relevant to daily practice given that the typical patient is “enrolled” rather than the ideal patient. Depending on the particular question, centers are selected based on the interventions offered: each center would have the flexibility of offering the intervention it is most equipped and/or trained for, rather than one that they are not familiar with or do not prefer. Therefore, this approach would minimize preferential patient enrollment or outcome assessment. It is evident that this strategy does not apply to questions addressing the effect of an intervention compared with placebo. Furthermore, as with other nonrandomized studies, the lack of randomization of patients and blinding in PROS can introduce bias. Furthermore, the practice at various institutions can be greatly heterogeneous. Therefore, strategies must be implemented to improve this methodology.

The deviation of the concept of PROS from PCTs and traditional RCTs should not be interpreted as the former not being adherent to any quality-control guidelines. First and foremost, the investigators involved in the study must establish proof that the current question cannot be reasonably addressed through a well-designed and implemented RCT; this includes the concepts of equipoise, blinding, resource availability, and other aforementioned concerns. Furthermore, other factors such as the required sample size, planned statistical analyses, and clearly defined objectives and end points should be established upfront. These and other considerations have been comprehensively outlined in the 2012 International Society for Pharmacoeconomics and Outcomes Research (ISPOR) position statement. The ISPOR group further suggests that a detailed and accurate protocol be established and registered upfront to increase the validity of the study and its findings. A commitment to publishing the results, regardless of whether a positive or negative outcome was attained, is also essential. In addition, guidelines must be implemented to track, report, and act upon all adverse events. Protocols for the protection of patient privacy are also essential. Furthermore, while
patients enrolled at each participating center to take part in the study are not randomized, there are no restrictions on randomizing participating centers to determine the ones that are to be included in the final analysis. Under ideal circumstances, this strategy would balance some of the baseline characteristics of the centers involved. The results of this ideal scenario would reflect a realistic pool of typical patients managed for a particular condition under routine intervention settings. The success of this approach would depend on the involvement of a large number of centers, which is not always feasible. Such strategies would be particularly relevant for addressing clinical queries pertaining to pathologies for which there is a dearth of RCTs, such as meningiomas and vestibular schwannomas.

The current study was not a systematic review of all neurosurgical RCTs in the stated time interval and it is likely that some studies may not have been included. Furthermore, the CONSORT, Jadad, and PRECIS-based scales involve subjective assessments. In addition, the PRECIS-based scale was arbitrary and this approach has not been attempted to assess the design of other studies. However, the purpose of this study was to determine the general trend of neurosurgical RCTs with regard to the quality of design and reporting, and correlate these with previous assessments. Furthermore, the CONSORT scores were assessed by 3 independent reviewers with a high correlation, and the corroboration by previous studies is reassuring as well.

The current study does not imply that RCTs are not applicable to neurosurgery and should therefore be replaced by PROS. Rather, our objective has been to demonstrate that not all neurosurgical questions can be answered through the conduct of an RCT and that alternative complementary approaches must be considered. Furthermore, we have provided strategies on developing prospective observational trials that adhere to as many of the principles of sound clinical research as possible and convey high-quality clinical findings that are applicable to routine daily practice in situations where a well-designed RCT is simply not feasible. The pace and quality of neurosurgical research can be improved.

References


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

Author Contributions
Conception and design: Kondziolka, Mansouri. Acquisition of data: Mansouri, Cooper, Shin. Analysis and interpretation of data: Mansouri, Cooper, Shin. Drafting the article: Kondziolka, Mansouri. Critically revising the article: Kondziolka, Mansouri. Reviewed submitted version of manuscript: Kondziolka, Mansouri. Statistical analysis: Mansouri. Study supervision: Kondziolka.

Correspondence
Douglas Kondziolka, Departments of Neurosurgery and Radiation Oncology, Suite 8R, NYU Langone Medical Center, New York, NY 10016. email: douglas.kondziolka@nyumc.org.