Methodology and reporting of meta-analyses in the neurosurgical literature

John H. Sampson, M.D., Ph.D.1 and Fred G. Barker II, M.D.2

1 Neurosurgery, Duke University Medical Center, Durham, North Carolina; and 2Massachusetts General Hospital, Boston, Massachusetts

Athanasius Kircher, the 17th-century polymath whose interests ranged from volcanoes to decoding Egyptian hieroglyphics, was (we are told) the last person on earth who knew everything—who had mastered all of human knowledge in every field.9 Since his death in 1680, the rest of us have been falling farther and farther behind in our reading. In 2012 PubMed added 12 new articles per day with the major subject heading “neurosurgery” and, for the specialist, still more—for example, 13 new articles each day on “brain neoplasms.” We all need help in keeping up with this information tsunami, and given the realities of publishing, the traditional textbook cannot keep us abreast of the latest advances.

Narrative review articles in medical journals have long been a popular response to this problem, and in the 1970s and 1980s a new quantitative technique of combining information from multiple randomized clinical trials (RCTs) sprang up as an advance on the traditional narrative review. Proponents argued that a systematic search for all relevant trials would reduce subjectivity while providing additional statistical power to answer important questions; in an era of underpowered trials this was important, and it was easy to demonstrate partiality in both narrative reviews and expert opinion.10 Despite initial criticism, the “meta-analysis” caught on and it is now common to see it placed at the top of the evidence-based medicine pyramid, above (that is, more reliable than) the individual RCT.21 In 1989, the first medical textbook based almost entirely on systematic reviews and meta-analyses, Effective Care in Pregnancy and Childbirth, was published.8

Neurosurgeons will have noticed an increasing number of meta-analyses being published on neurosurgical topics and in our journals. Now we have the first examination of the quality of this rapidly growing branch of our literature, using standard instruments specifically designed to evaluate and grade meta-analyses. Klimo et al. collected 72 evaluable papers self-described as meta-analyses that were published in Neurosurgery and the JNS (Journal of Neurosurgery) Publishing Group journals (Journal of Neurosurgery, Journal of Neurosurgery: Pediatrics, Journal of Neurosurgery: Spine, and Neurosurgical Focus) between 1990 and 2012 to ascertain the frequency and quality of this particular type of systematic review.15 The authors used both the 27-item Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist16 and the 11-item Assessment of Multiple Systematic Reviews (AMSTAR) checklist, designed to evaluate the reporting (PRISMA) and the underlying methodology (AMSTAR) of published meta-analyses respectively.20

How did neurosurgery do on these report cards? Regrettably, we failed. In the 72 papers evaluated, Klimo et al. found that on average only 53% of PRISMA items and only 31% of AMSTAR items were completed. Further, only 15% of the papers mentioned using a content checklist such as PRISMA, and no paper reported using a methodology questionnaire. We used literature searches and indexes of papers citing the PRISMA and AMSTAR reports to locate similar studies on meta-analyses published in other medical fields and found that without exception neurosurgery had the worst quality meta-analyses of any medical field on both checklists (Table 1).

Does it matter? After all, the items that make up these checklists are themselves only expressions of expert opinion and have not been rigorously proven to be necessary for a meta-analysis to reach the “right” conclusion—a gold standard that is itself hard to define. But the arguments advanced by the authors of the original checklists, and reviewed by Klimo et al., are largely convincing.15 For example, both checklists mandate that meta-analyses specify the funding source for the meta-analysis itself, and AMSTAR specifies that funding sources also be given for the individual RCTs that make up the meta-analysis. There is certainly ample evidence that funding sources influence both reviewers’ opinions about published medical data and the selective publication, manipulation, or withholding of the primary data itself in RCTs, making it reasonable to require meta-analyses to disclose this as well—particularly in view of the many subjective decisions made behind the scenes in these studies and their
potential impact on the use of commercial products. Meta-analysts can choose whether to report on one drug (nimodipine) or a class of drugs (calcium channel blockers), on one outcome (fusione rates) and not another (cost), on which studies to include or reject, and on whether a graph shows publication bias or does not. It is not unknown for two meta-analyses to reach opposite conclusions based on the same underlying population of trials because of these subjective decisions. Like funding, most of the other items on these checklists also have substantial face validity from direct analogy to other types of medical study, such as RCTs, and are likely to reduce bias and improve reliability and reproducibility in meta-analyses. The checklists make sense and we should adhere to them.

Will this fix our problem? After all, most of these studies were published long before the checklists were invented; some neuroscience journals already require them; we (as authors, reviewers, and editors) can promise to do better in the future. But Klimo et al. outline another problem with neurosurgical meta-analyses, probably a more important one, and certainly one that will be harder to fix. Meta-analysis was invented to combine the evidence from multiple “experiments”—in modern medical terms, from RCTs. As such, the method has a formidable theoretical underpinning and a fair track record of success as judged by correct predictions about large RCTs conducted subsequent to the meta-analysis itself. But most neurosurgical meta-analyses, it seems, are not based on randomized trials, and many are not based on trials at all—they are collections of case series and case reports that the authors have subjected to statistical tests that are often not appropriate for this use. Klimo et al. found that just 13 (18%) of the 72 papers they studied were based on RCTs alone, and another 12 included at least one RCT (total 35%). Again, we compared these figures to those reported in other medical and surgical fields, and again neurosurgery had the worst results (Table 1). The next lowest value we found, from a collection of meta-analyses in dentistry, had twice as high a proportion of RCT-based meta-analyses as neurosurgery does. Many other surveys of meta-analyses that we reviewed, but could not include in our table, did not report the proportion based on RCTs, apparently because of a tacit or explicit assumption that all meta-analyses are based on RCTs. Because Klimo et al. limited their study to journals of the American Association of Neurological Surgeons and of the Congress of Neurological Surgeons, including the highest-impact neurosurgical journals and thus perhaps to the highest-quality neurosurgical meta-analyses, the true proportion of meta-analyses written by and read by neurosurgeons that are based on RCTs is probably even lower than the 20% that Klimo et al. report.

Given the tiny number of RCTs published in neurosurgery (about 1% of our literature), this embarrassing result should come as no surprise. But it’s important to understand the types of problems in the underlying literature that meta-analysis is capable of correcting—what meta-analysis can and can’t do. When several RCTs are performed that ask the same question, involving the same types of patients, all of whom are eligible for both treatments, and using treatments that are reasonably the same from study to study, there will of course still be a variation in their results due to the play of chance, especially if the individual studies are small. This statistical variation might even be larger than the treatment effect itself, so that studies on a truly effective therapy may (taken in isolation) fail to reach a conventionally “statistically significant” threshold for declaring therapeutic success and may even have apparently conflicting results. When the individual trials each provide an unbiased (though imprecise) measurement of the actual treatment effect, meta-analysis can successfully identify the “correct” answer to the investigators’ question. When the individual trials have a consistent bias, meta-analysis will blindly return the bias itself as the estimate of treatment effect. Klimo et al. illustrate this with a craniohypophyseal example: they assume tumors resected transphenoidally will be smaller than those resected through open craniotomy and point out that a crude comparison of gross-total resection rates or visual results is likely to be skewed beyond the point of usefulness by the inclusion of tumors that would not have been eligible for transphenoidal removal in the craniotomy group. We leave to the interested reader the challenge of finding all of the neurosurgical meta-analyses that do calculate these exact comparisons for endoscopic versus open treatment of these and other paraseptal tumors—we found 10 in an admittedly unsystematic search. But the basic point is well established: when a nonrandomized study resembles an RCT in design and conduct (both treatments conducted concurrently, and all patients eligible for both treatments), the results are likely to be similar to RCT results; when controls are historical, or patients are not eligible for one of the treatments, the comparison will be biased. Perhaps—we have no way of

<table>
<thead>
<tr>
<th>Authors &amp; Year</th>
<th>Specialty</th>
<th>PRISMA</th>
<th>AMSTAR</th>
<th>RCT-All</th>
<th>RCT-Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhandari et al., 2001</td>
<td>orthopedics</td>
<td>NR</td>
<td>NR</td>
<td>45%</td>
<td>85%</td>
</tr>
<tr>
<td>Bader &amp; Ismail, 2004</td>
<td>dentistry</td>
<td>NR</td>
<td>NR</td>
<td>39%</td>
<td>NR</td>
</tr>
<tr>
<td>Mrkobrada et al., 2008</td>
<td>renal</td>
<td>NR</td>
<td>NR</td>
<td>66%</td>
<td>NR</td>
</tr>
<tr>
<td>Fleming et al., 2013</td>
<td>orthodontics</td>
<td>64%</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Gagnier &amp; Kellam, 2013</td>
<td>orthopedics</td>
<td>68%</td>
<td>54%</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Momeni et al., 2013</td>
<td>hand surgery</td>
<td>NR</td>
<td>7 items (64%)</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Klimo et al., 2014</td>
<td>neurosurgery</td>
<td>55%</td>
<td>31%</td>
<td>20%</td>
<td>38%</td>
</tr>
</tbody>
</table>

* NR = not reported.
knowing—the comparison will be so biased as to be use-
less or even deceiving.

So: garbage in, garbage out. This has been one of the
criticisms of meta-analysis from the beginning, and it still
holds true. When the input to the meta-analysis is Level IV evidence (case studies and case series), the results are
Level IV evidence. While only a few of the studies ex-
amined by Klimo et al. were based entirely on individual
cases collected by authors using a systematic review, our
experience as readers and reviewers is that this is a type
of study whose popularity is virtually exploding. While
the intent to make a silk purse out of many individual
sows’ ears is understandable, studies that use this design
should not be called meta-analyses, because they do not
combine the results of separate, individual analyses each
comparing the same two treatments, but instead pool
results of individual cases. While the distinction may
initially seem academic, it is actually an important one.
When authors compare two treatments within a single in-
stitution or cooperative multicenter study, we can expect
some basic comparability between the patients receiving
the two treatments: they came from the same referral
base, to the same institution(s), during the same interval;
their diagnoses and treatments were classified as similar
by investigators with access to primary clinical data; their
general treatment aside from the intervention under study
is likely to have been consistent; their outcome is likely
to have been graded using consistent outcome scores by
the same observers; we can reasonably ask whether the
patients were a consecutive series or a highly selected
subgroup; and we can ascertain how much important data
is missing and if necessary adjust for it.

When the patients have been culled from the litera-
ture, none of these assumptions is likely to be true and
most cannot be tested. Instead, authors combine groups
of patients treated during different eras, using many dif-
ferent and usually nonstandard outcome scales, rated by
different teams of investigators at many different insti-
tutions, perhaps at different time points after treatment.
We have no sense of individual patient eligibility for both
treatments. We know that surgical patients reported on
in the literature have better outcomes than those who are
not reported on, a form of publication bias, and that the
smallest series report the most skewed results.1 But we
usually have no way to be sure whether the published cas-
es in these reviews are a large fraction of those in which
patients suffered from the condition (and hence repre-
sentative) or a tiny tip of the iceberg (and possibly very
unrepresentative). We find multiple outcomes reported on
patient cohorts even though only a minority of patients
had information on each outcome (less than 10% in some
examples), a form of selective outcome reporting bias in
the primary papers that is known to skew heavily toward
overoptimistic results.14 We find success rates between in-
dividual series that range wildly after what is described
as identical treatment—in one neurosurgical example,
ranging from 2% to 50% for the main endpoint in one arm
of the comparison and from 0% to 45% in the other arm—simply added up in the two arms and compared us-
ing a chi-square test with no adjustment for heterogeneity
between series. This is a seriously inappropriate statisti-
cal method when heterogeneity is present, as it usually is,
and is highly likely to yield a misleading result. Unfortu-
nately, tests for heterogeneity lose power as the number of
cases in each included report decreases, approaching zero
in the limiting instance of reports of only a single case.
Based on our knowledge of biological, clinical, and meth-
odological heterogeneity between individual small case
series, we should be cautious and analyze results of such
reviews as if containing significant heterogeneity even
when formal statistical tests fail to prove its presence.

These pooled reviews, which are not meta-analyses in
the original sense, could better be called literature-
based synthetic cohort studies or synthetic case series
to indicate both their composite origin and their essential
nature as artificial cohorts assembled from published case
series. Reviewers and readers should expect synthetic
case series to meet the same minimum standards of qual-
ity we would expect from a single-institution case series,
or at least a candid discussion of how they fall short of
that standard. In addition synthetic case series analyses
should always include a quantification and discussion of
between-study heterogeneity, as well as the recognition
that if it is too large (or if tests for it are too underpow-
ered) the individual case results can’t be combined sta-
tistically as originally planned. Authors of such studies
should contact the authors of the original published se-
ries and case reports as necessary and ask them to supply
missing data, update patient outcomes, and potentially
supply additional cases before conducting their analysis.
In most cases, it should be possible for authors to esti-
mate and report what proportion of actual incident cases
of the condition being studied have reached literature
publication and are included in the review. For example,
if cavernous malformations in a particular rare location
are the object of the study, what proportion of all cav-
ernous malformations do they represent (from unselected
case series) and what is the annual incidence of cavern-
ous malformations in a relevant population? Reviewers of
such articles need the necessary experience in this area to
enforce such standards.

The logical place for such studies in our literature
should be in examining truly rare conditions for which
individual institutions’ series are inadequate and no ap-
licable prospective registry or administrative database
is available. Conclusions about demographics and over-
al outcome will be the strongest results of such studies.
Treatment comparisons will generally be difficult to trust
in the absence of the protection provided by randomiza-
tion or single-series internal comparisons.

While the quality of meta-analyses in the neurosurgi-
cal literature may be improving, at least in the journals
Klimo et al. examined, our current standards of peer re-
view still do not adequately address these common de-
fects in design and analysis, and the conclusions of pub-
lished meta-analyses must be evaluated in the context of
these findings. We feel this should apply as well to the
many papers in our journals that perform a statistical syn-
thesis on published data but are not self-identified as me-
ta-analyses or even as systematic reviews, a group of pa-
pers that is actually much larger than the cohort Klimo et
al. have examined. The authors urge a standardization of
meta-analyses beginning with a clear definition of what a meta-analysis is, requiring both authors and reviewers to use both the PRISMA and AMSTAR checklists and requiring authors to include experts in the statistical methodology of meta-analysis at the authorship level when undertaking systematic reviews. We agree with the authors that their study has identified a crisis in our literature that our journals should address, before a potentially useful tool loses its value.

(http://thejns.org/doi/abs/10.3171/2013.10.JNS13724)

**Disclosure**

The authors report no conflict of interest.

**References**


**Response**

PAUL KLIMO JR., M.D., M.P.H., 1,2 CLINTON J. THOMPSON, M.S., 3 BRIAN T. RAGEL, M.D., 4,5 AND FREDERICK A. BOOP, M.D. 6

1Semmes-Murphey Neurologic & Spine Institute; 2Department of Neurosurgery, University of Tennessee Health Science Center, Memphis, Tennessee; 3School of Public Health and Health Services, The George Washington University, Washington, DC; and 4Department of Neurosurgery, Oregon Health & Science University, Portland, Oregon

We are deeply grateful for the eloquent and prudently crafted editorial by Drs. Sampson and Barker. It is clear they invested much time, energy, and thought into formulating a response to our findings and recommendations. There is very little we can add to what they have already written, but we would like to offer the following points for consideration.

Systematic reviews and meta-analyses are a reflection of the quality of literature that is available on a particular topic. Therefore, it is not surprising that many of the meta-analyses in neurosurgery are based on studies other than randomized controlled trials. Randomized controlled trials are rare in neurosurgery for a number of reasons: 1) the rate of patient accrual may be too low; 2) they are time-consuming and challenging to properly design; 3) approval by the institutional review board may be prohibitively frustrating; and 4) they may create ethical dilemmas. Ultimately, an RCT may not be the best study design from a financial or logistical standpoint to answer a particular clinical question, despite it being the gold standard for assessing the efficacy of a therapeutic intervention. Thus, meta-analyses based on observational studies need to be conducted with greater emphasis on reporting and methodological rigor in order to minimize bias.

Drs. Sampson and Barker correctly identify the bur-
Editorial

growing popularity of studies in the neurosurgical literature that appear to be more sophisticated than a narrative review but fall significantly short of a systematic review or meta-analysis. Their description of this type of report as a “literature-based synthetic cohort study” or a “synthetic case series” is spot-on. Currently, there are efforts underway to produce such “pooled reviews” on a massive scale. Multiinstitutional, diagnosis-specific data repositories—the most notable example being the National Neurosurgery Quality and Outcomes Database (N2QOD)—are just that. If clinical research papers are published using information from databases such as N2QOD (separate from their function to establish benchmarks in quality), the level of evidence cannot be graded any higher than their basic building block—a Level IV case series. Although there may be uniform collection of preoperative data and postoperative outcomes, it is still a pooled collection of cases performed by various surgeons at various institutions across various time periods. There is no a priori question posed. There is no a priori attempt to control or standardize variables that could have a dramatic impact on patient outcome (particularly, in degenerative spinal pathology, the first functional N2QOD module), such as preoperative symptom severity, radiographic data, preoperative treatment(s), surgical technique, surgical ability, and postoperative treatment(s), to name a few. Controlling for potential bias or confounders prior to study onset is methodologically preferable. By collecting massive amounts of data that can be analyzed by statisticians, we fear that such repositories may circumvent desperately needed prospective cooperative clinical trials. Excellent examples of recent collaborative studies include the evaluation of ultrasound-guided shunt insertion and the effectiveness of lumbar discectomy and single-level fusion for spondylolisthesis.\textsuperscript{1,2} What N2QOD has demonstrated, thus far, is that neurosurgeons are willing to invest considerable resources, financial and otherwise, to collect large amounts of data. Parallel efforts, in our opinion, should continue to be directed at designing appropriate multiinstitutional clinical trials in an effort to address the many questions to which we as neurosurgeons and our patients need answers. In summary, adherence to available and proven reporting and methodology checklists, in conjunction with a collective effort on the part of the academic neurosurgical community to pursue evidence-based medicine clinical studies, will yield the highest quality and most clinically useful systematic reviews and meta-analyses.

References


Please include this information when citing this paper: published online January 24, 2014; DOI: 10.3171/2013.10.JNS13724.