Alpha, Beta, Meta: Guidelines for Assessing Power and Type I Error in Meta-Analyses

Edward J. Mascha, PhD

“If I have seen further…it is by standing on the shoulders of giants.”
—Isaac Newton

Minimizing the chance of a false-positive finding, or type I error, should be a fundamental goal whenever a research hypothesis is tested. However, unless steps are taken to control it, the chance of type I error is increased when multiple testing is done, for example, comparing multiple treatment arms, analyzing multiple primary and secondary outcomes, subgroup analyses, or repeated testing of the data. Type I error is also more likely with small studies because the treatment effect estimates tend to be more unstable, especially for binary outcomes.

In their recent article in Anesthesia & Analgesia, Imberger et al.1 highlight that the chance of a type I error is also increased in meta-analyses because results are repeatedly updated over time. They further show that type II error (β), the probability of missing true findings, is an issue with many meta-analyses—that is, they are underpowered. Guidelines would thus be helpful for reviewers and editors to be able to see at a glance whether a meta-analysis appears to be well-powered or not and whether there appears to be an elevated chance of a type I error. We offer such guidelines in this article.

We further propose to expand current scoring systems for the quality of a meta-analysis such as the AMSTAR (A Measurement Tool to Assess Systematic Reviews) checklist (amstar.ca) or the Centre for Evidence Based Medicine (Oxford) systematic review critical appraisal data sheet (http://www.cebm.net/critical-appraisal/) to include assessment of power and control of a type I error for the meta-analysis as a whole. Particularly, we propose adding the following questions to the checklists:

1. Does the analysis account for multiple testing?
2. Is the analysis adequately powered?

We know that the overall chance of falsely rejecting the null hypothesis, say total α, increases with the number of hypotheses. For example, with 10 comparisons, all done using a significance criterion of P < 0.05, there is a 40% chance of making at least 1 false-positive conclusion! Common methods to control type I error are Bonferroni correction for multiple comparisons using an α level that is small enough (e.g., some journals/editors recommend P < 0.01 or smaller for observational studies) and group sequential methods for clinical trials.

GUIDELINE 1

Guideline 1: Research reports that test multiple hypotheses or repeat analyses but make no mention of methods to control type I error most likely have an elevated chance of reporting false-positive results. This holds for both single trials and meta-analyses.

In a single clinical trial in which interim analyses are conducted after each new group of patients is randomized (hence the name “group sequential” design), mathematical functions are used to “spend” some of the prespecified α at each analysis, making sure to cumulatively not spend more than the overall prespecified level (usually 0.05) over the course of the trial.2-4 This is done by increasing the P value “bar” (i.e., requiring a smaller P value) for significance at any particular analysis beyond what it would be for a single analysis. As well, type II error is often spent over time to monitor futility.

Combining studies into a meta-analysis over time is analogous to accumulating data over time in a single-study group sequential design but with more expected heterogeneity of effects. However, when meta-analyses are updated over time to add new studies, current practice uses the customary 0.05 significance criterion each time, thus increasing the overall chance of a type I error. As discussed by Imberger et al., one promising solution is trial sequential analysis (TSA)5;6 other related approaches are emerging as well.7-8 TSA applies the α-spending principle from group sequential design to a meta-analysis to control the type I error (say at 5%). A TSA projects the overall sample size for the meta-analyses that the research community should target, based on expected variability and the treatment effect of interest (the latter might be based on an average of the planned effect sizes for the included clinical trials in the meta-analysis or on a current consensus from the research.

From the Departments of Quantitative Health Sciences and Outcomes Research, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication August 10, 2015.

Funding: Anesthesia Institute.

The author declares no conflicts of interest.

Reprints will not be available from the author.

Address correspondence to Edward J. Mascha, PhD, Cleveland Clinic, 9500 Euclid Ave., P77, Cleveland, OH 44195. Address e-mail to maschae@ccf.org.

Copyright © 2015 International Anesthesia Research Society

DOI: 10.1227/ANE.0000000000001093

1Intuitively, total alpha is 1 minus the probability of not making a type I error (1 - αi) on any K tests conducted. Thus, total α = 1 - (1 - αi)^K, where αi is P value criterion at iih analysis and K is number of analyses. So with 10 comparisons, all done at the 0.05 significance level, total α = 1 - (1 − 0.05)^10 = 0.40.
community on what is clinically important). The TSA gives the research community information on whether a particular meta-analysis should be considered to be conclusive or whether more studies are needed.

Clearly, TSA proposes a fairly drastic, and arguably refreshing, change to how meta-analyses are planned and conducted. A main change is that initial studies done for a research question would have a much higher bar for significance than later ones, when more information would be available; this is analogous to a group sequential design for a single trial. For example, using the conservative Lan and Demets spending function (which mimics the less flexible O’Brien-Fleming method), as recommended by Imberger et al. for TSA, would have a P value criterion for significance of $P < 0.000014$ at 25%, $P < 0.003$ at 50%, and $P < 0.018$ at 75% of the total required sample size instead of the traditional $P < 0.05$. Less stringent criteria early on (e.g., a function mimicking the more aggressive Pocock spending function), with the penalty of an increase in the maximum required $n$, could also be used.

For instance, research on whether a particular anesthetic regimen can reduce 30-day mortality over another would typically require thousands of patients. To have 90% power at the 0.05 significance level to detect a 20% relative reduction from 5% mortality would require 18,150 patients for a single trial with no interim analyses. Conservative spending functions for efficacy and futility would raise the maximum $n$ to approximately 19,470. In a meta-analysis context, a first trial of 1000 patients would only be 5% of the required maximum. Adding a second trial of 5000 patients would only reach approximately 30% of the information target. Early investigations would thus be encouraged to highlight the fact that further studies are needed before the meta-analysis should be considered definitive. TSA could be used to project a recommended target sample size, perhaps under various effect size and heterogeneity scenarios.

Imberger et al. studied a random sample of 50 meta-analyses with dichotomous (yes/no) outcome variables that reported statistically significant findings, with striking results. Applying TSA, they report (Imberger et al., Table 4) that only 38% (19 of 50) of the meta-analyses preserved the type I error at 5% or less when powered for detecting a relative risk of 20% between groups, suggesting that 31 of the 50 significant results may be false-positives. Furthermore, only 12% (6 of 50) had 80% power to detect a 20% reduction, meaning that although significant at $P < 0.05$, the meta-analyses should have included more patients. In addition, although power increased with the number of studies, participants, and outcomes, even the largest studies were generally underpowered, and only 60% to 70% of the largest were still statistically significant when type I error was controlled using the TSA method (Imberger et al., Table 5).

Imberger et al. report only on meta-analysis with positive (significant) results. One can safely assume that even a larger proportion of nonsignificant meta-analyses is underpowered, an assessment of those studies as well would be enlightening!

Note that Imberger et al. use the term “imprecision” to imply studies having evidence of a type I error. We are using “precision” in the more usual sense to simply indicate how narrow the confidence interval is—more narrow, driven by more data and less variability, is more precise.

**GUIDELINE 2**

*Guideline 2:* A meta-analysis should be considered definitive only if the total sample size is at least as large as expected for a well-powered clinical trial (Fig. 1) or else if the signal is so strong that an efficacy (or futility) boundary is crossed using a TSA (or analogous approach) to protect type I error for multiple looks over time.

How can one tell at a glance whether a study is well-powered? Required sample size (either single study or meta-analysis) with a binary outcome is a function of 4 things: control group incidence, relative risk, power, and significance level. Figure 1 displays the sample size per group required to have 90% power at the 0.05 significance level to detect various relative risks given reference group incidences of 10%, 20%, and 30% with the outcome. For example, even with a high 30% reference group incidence (bottom line in Fig. 1), approximately 5000 per group are needed to have 90% power to detect a 10% relative reduction (relative risk of 0.90), whereas only approximately 1000 per group are required to detect a 20% reduction. With more likely reference group incidences of 10% or 20%, the required sample size increases substantially for a given relative risk.

Thus, when assessing whether a research report is sufficiently powered, we should first note the highest observed incidence for either group (assuming an effect in either direction is being tested for). Then consider what relative risk makes sense for the research hypothesis and refer to Figure 1 to see whether the observed per-group sample size is adequate. For example, with a primary outcome of major infection after surgery, one might observe approximately a 15% incidence in the worst group, and a relative reduction of approximately 20% or more (relative risk of 0.80) would be clinically important. Figure 1 indicates that between 1000 and 2000 patients per group would give adequate power; reports with <1000 per group would be underpowered. For a meta-analysis, the required numbers would be higher because of between-trial heterogeneity, so these are minimum requirements.

The total number of observed events can be nearly as indicative as the number of patients in determining whether a study with a binary outcome is sufficiently powered, as shown in Table 5 in the article by Imberger et al. and others. Table 1 provides the required total number of events and total patients needed to detect various relative risks at specific control group incidences. Notice that if the total number of events is <100 (say for an entire meta-analysis), then the study could only be sufficiently powered to detect huge relative risks of 0.5 or greater, and with 100 to 200 events, only relative risks of 0.60 or greater. Because usually a relative risk of 0.80 (20% reduction) is quite clinically important, at least 500 events would typically be needed for adequate power at control group incidences of <50%.

Small research studies are generally a concern, especially with binary outcomes, because they are unstable (a change of 1 or 2 data points can change conclusions) and thus prone to being both underpowered and to giving spurious positive results. However, including small studies along with larger ones in a meta-analysis is not much of a concern because...
none of them individually is given much weight. Findings from a meta-analysis are driven by the largest studies. Thus, the mentioned guidelines for power in a meta-analysis are for the aggregate numbers and not for scrutinizing the individual studies.

Heterogeneity of the treatment effect is expected and typically observed in meta-analyses. Correspondingly, a meta-analysis will need at least some more patients than a single large trial to address the same research question, all else being the same. Meta-analysis methods account for the degree of heterogeneity of treatment effect among the trials. Likewise, the TSA approach to control type I error uses the observed heterogeneity to date and also elicits the overall expected degree of heterogeneity as an input parameter. Because it is unknown in the early stages of assessing a research hypothesis, one should consider several possible levels of heterogeneity when conducting a TSA; the unlikely “zero heterogeneity” will give results very similar to a single trial group sequential design.

Interestingly, the maximum projected sample size for a meta-analysis can vary greatly as each new trial is added. This is because of the typical heterogeneity across trials, meaning that the observed treatment effect from one trial to the next can vary considerably; depending on how large the newly added trial is, this may strongly affect the maximum required $n$. A future area of research may be to study methods to obtain a less variable “maximum” sample size over time and still adjust for heterogeneity and type I error.

Reviewers and conductors of meta-analyses may also wonder if the increased type I error is more of an issue if a meta-analysis is conducted after each sequential trial is completed compared with if only a single meta-analysis is conducted after several studies are published on a topic. This is relevant because as Imberger et al. point out, the Cochrane Collaboration recommends that all systematic reviews be updated every 2 years. One can argue that because each study addresses the same research question and makes an inference about it, there is a chance for type I error in simply publishing the multiple individual studies. However, additional “summary” inference is done for each meta-analysis, and thus, the number of meta-analyses done on the topic also seems relevant to the overall type I error. It is my understanding that the TSA approach does not consider how many previous meta-analyses have been done when addressing the type I error for the current meta-analysis. How to make the optimal type I error adjustment across the research question may also be an issue for further study and debate.

Finally, aside from whether a meta-analysis result is statistically significant or not, precision of the estimated treatment effect as measured by width of the confidence interval should be given considerable weight in

---

**Table 1. Events and Sample Size as a Function of Relative Risk and Control Group Incidence (Required Events/Required $n$)**

<table>
<thead>
<tr>
<th>Control group incidence</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>87/1156</td>
<td>155/1923</td>
<td>309/3622</td>
<td>774/8596</td>
<td>3433/36,125</td>
</tr>
<tr>
<td>0.2</td>
<td>80/526</td>
<td>141/873</td>
<td>279/1637</td>
<td>697/3867</td>
<td>3072/16,161</td>
</tr>
<tr>
<td>0.3</td>
<td>72/316</td>
<td>126/522</td>
<td>250/976</td>
<td>619/2291</td>
<td>2710/9506</td>
</tr>
<tr>
<td>0.4</td>
<td>65/211</td>
<td>112/347</td>
<td>221/645</td>
<td>542/1503</td>
<td>2349/6179</td>
</tr>
<tr>
<td>0.5</td>
<td>57/148</td>
<td>98/242</td>
<td>191/448</td>
<td>465/1030</td>
<td>1987/4182</td>
</tr>
</tbody>
</table>

Required total number of events and total number of patients needed to detect various relative risks at specific reference group incidences with 90% power at the 5% significance level. Numbers are based on a single trial. Meta-analyses need more events and total $n$ (perhaps substantially more) to attain the same power, accounting for $\alpha$ spending in trial sequential analysis and between-study heterogeneity.
determining whether the results are conclusive. Early crossing of an efficacy boundary, even if using stringent criterion, may give confidence intervals too wide for the results to be considered definitive. Standard confidence intervals for a meta-analysis account for the heterogeneity across trials in the treatment effect. Importantly, TSA further adjusts meta-analysis confidence intervals (i.e., makes them even wider) in attempts to control type I error at the desired level over repeated testing, as is done in group sequential results for a single clinical trial. This is a good thing, and no doubt there will be more research on the best ways to make this adjustment. It cannot be stressed enough, however, whether for TSA or any assessment of treatment effect, that a $P$ value is not sufficient or even the most important result; an estimated treatment effect and confidence interval give more information.

Indeed, in conducting meta-analyses, we earnestly strive to “see further” by compiling results of previous investigations, some being “giants” in size and some not. However, simply conducting a meta-analysis does not mean that the research question has been definitively answered. Aside from the important issues of study quality and combinability, primary concerns are whether the meta-analysis itself is well-powered and, if significant, whether the chance of a false-positive conclusion has been adequately controlled. New and exciting tools such as TSA encourage the research community to work together in the planning and conducting of meta-analyses over time and hopefully to create standard methodology to help clarify and agree when a research question has been definitively answered.

**DISCLOSURES**

**Name:** Edward J. Mascha, PhD.

**Contribution:** This author designed and conducted the study, analyzed the data, and wrote the manuscript.

**Attestation:** Edward J. Mascha approved the final manuscript.

**This manuscript was handled by:** Franklin Dexter, MD, PhD.

**REFERENCES**