Author's response to reviews

Title: Dealing with Heterogeneity of Treatment Effects: Is the Literature Up to the Challenge?

Authors:

Nicole B Gabler (nrbloser@ucdavis.edu)
Naihua Duan (Naihua.Duan@Columbia.edu)
Diana Liao (DLiao@mednet.ucla.edu)
Joann G Elmore (jelmore@u.washington.edu)
Theodore G Ganiats (tganiats@ucsd.edu)
Richard L Kravitz (rlkravitz@ucdavis.edu)

Version: 4 Date: 24 May 2009

Author's response to reviews: see over
May 23, 2009

Douglas G. Altman, DSc
Curt Furberg, MD, PhD
Jeremy Grimshaw, MBChB, PhD, FRCGP
Peter Rothwell, MD, PhD
Editors-in-Chief, Trials

Dear Drs. Altman, Furberg, Grimshaw, and Rothwell:

We are pleased to re-submit our manuscript entitled “Dealing with heterogeneity of treatment effects: Is the literature up to the challenge?” (MS# 1048464801226605). We appreciate the helpful comments from the editors and reviewers, and have attended carefully to them. Please see below for a table detailing our responses and manuscript changes. We believe that the corresponding manuscript is much improved as a result of the reviews and comments.

Thank you for your consideration.

Best regards,

Nicole B. Gabler, MPH, MHA
PhD Candidate, Graduate Group in Epidemiology
Center for Healthcare Policy and Research
University of California, Davis
From Reviewer 1:

(1) In my original review, the main point was that the authors “seem to ignore the fundamental problem of assessment of individual variability that is associated with unusual parallel group trials...”. I requested that the authors discuss the problem of variation due to intra-individual reasons (e.g., a drug may or may not be working in this patient on a given day) more carefully.

It is good that the authors now include refs [12,13] and insert three short sentences on each of three places I proposed. However, instead of this, I’d prefer a thorough discussion of this issue at one place in the beginning of the ms (p. 5/6). What about taking the sentence from the “Response to reviews”; “Subgroup effects may be due to random intra-individual variability, which is impossible to estimate in a parallel group trial.”

Intra-individual variability is an important component of heterogeneity and we appreciate the reviewer highlighting its contribution. An additional explanation of this issue is provided in the introduction. The old text read:

Subgroup analysis can be perilous. Real effects can be missed because of inadequate statistical power [16, 17], and reported effects may be spurious because of the performance of multiple statistical tests (13-16) and/or due to random intra-individual variability [12, 13]. The Consolidated Standards of Reporting Trials (CONSORT) statement warns that subgroup analyses, especially post hoc subgroup comparisons, “do not have great credibility” [18].

The revised text reads:

Subgroup analysis can be perilous. Real effects can be missed because of inadequate statistical power [16, 17], and reported effects may be spurious because of the performance of multiple statistical tests (13-16) and/or due to random intra-individual variability [12, 13]. Random intra-individual variability is especially problematic because it is not possible to estimate this variability in parallel group trials, the most common type of clinical trial design. In parallel group trials, participants are only randomized to one treatment and do not crossover to alternative treatments. As such, it is not possible to estimate any variation that occurs within a participant. In recognition of the drawbacks of subgroup analysis, the Consolidated Standards of Reporting Trials (CONSORT) statement warns that subgroup analyses, especially post hoc subgroup comparisons, “do not have great credibility” [18].

(2) n-of-1 trials (p. 5/6, one of the three places where the point is mentioned): The authors say that “However, n-of-1 trials are...subject to random within-patient variability...”. This is misleading, because in the view of the problem described above it is a strength, not a weakness of the n-of-1 trial that it accounts for within-patient variability. In principle, such trials enable researchers to investigate intra-individual variability, what parallel group

We agree with the reviewer that within-patient variability may be measured in an n-of-1 trial (and is impossible to measure in a parallel group trial), and that this is a strength of the n-of-1 trial design. However, we also believe that it is important to recognize that chance (randomness) may also affect the results of the n-of-1 trial, therefore it is necessary to take within-patient variability into consideration when analyzing n-of-1 trials data. When
<table>
<thead>
<tr>
<th><strong>designs don’t.</strong></th>
<th>combining n-of-1 trials to estimate population HTE, a hierarchical Bayesian random effects model takes within-patient variability into consideration.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(3) “Bivariate analyses”</strong> (p. 11, 14): From a statistical point of view, I would not describe simple contingency table analyses as “bivariate”. A bivariate analysis is one with a bivariate outcome, modeled in terms of a couple of parameters that may influence the outcome. Please replace “bivariate analysis” by “contingency table (analysis)”.**</td>
<td>We appreciate the reviewer highlighting this source of confusion and have changed “bivariate analysis” on pages 11 and 14 to “two-way contingency table analysis.”</td>
</tr>
<tr>
<td><strong>From Reviewer 2:</strong></td>
<td><strong>We agree with the reviewer that the scope of our study was broad. However, we believe that the broad scope of our study does have value for researchers and readers and provides them with a useful synopsis of HTE exploration and reporting in the field at large. Prior publications of this topic have typically been limited to a specific discipline (such as cardiology) or to a specific journal; a larger scope allows for a better description of the situation at large.</strong></td>
</tr>
</tbody>
</table>
| (1) While it might be interesting to look at the analysis of heterogeneity of treatment effects in the literature, however, the formulation of the research question (prevalence of HTE analysis...in the general medical literature...trends over time) is way too broad to make the results of the study very useful. The studies identified are not equipped to answer this question. a. Prevalence of HTE analysis, disregarding whether it is appropriate or not, or the goal of the study, is not meaningful. (Why would you consider a study of n=10 for HTE). Also prevalence of HTE analysis itself is very vague, without defining any scope to study. b. It is limited to consider the five big journals as “general medical literature”. These journals are among the very best and likely publish the best-designed trials, however, HTE is usually not considered as criteria of study quality. c. Three years’ data are not adequate to assess trend over time. More time points are needed to get a reliable estimate for trend over time. | We appreciate the reviewer’s concern that HTE analysis might be inappropriate for some study, such as studies with very small sample sizes. Motivated by the same concern, we conducted the sensitivity analysis to provide an estimate of HTE reporting among those studies that would be best-suited to report HTE (sample size of at least 250 participants, with at least 100 participants per arm). In this analysis, we found that only 40% of trials report HTE analysis and an additional 31% report subgroup-only analysis. Highlighting that HTE is (1) often not conducted and (2) conducted incorrectly approximately one-third of the time underscores our study conclusions. We advocate for a greater frequency of HTE analyses (when conducted correctly) not because we believe that the prevalence should reach a benchmark value, but because HTE exploration is necessary for hypothesis generation. And, while we agree with the reviewer that HTE is not a criterion for study quality, we do believe that these well-designed studies should include exploratory HTE analyses for hypothesis generation. We appreciate the reviewer’s caution that three
years will not provide adequate data to assess trends over time. However, please note that the three years selected for this study are not three consecutive years, but rather span a ten-year period (1994, 1999, and 2004) during which important developments in the medical literature took place, such as the first publication of the SORT Statement in 1994 (1) and the CONSORT Statement in 1996 (2). To reliably examine a trend, we agree that we could benefit from selection more than three data points. However, our purpose in collecting this data was to assess whether researchers were heeding the CONSORT recommendations. If so, we expected to see an increase in appropriate HTE analysis with a concomitant decrease in (inappropriate) subgroup analysis over the three time periods. For this purpose, we believe the three years selected are informative.


(2) This study is not a systematic review, but used a convenience type of sample of articles to look at HTE analysis in selected journals and selected years. For a systematic review, one is generally expected to find all evidence relating to the research question. Due to the broad question, it is very unlikely to carry out such a search. The author needs to formulate a more defined and targeted research questions.

We agree with reviewers 2 and 3 that this study may not be a systematic review in the usual sense, in that we did not conduct an exhaustive search of all articles. To avoid any further confusion, we have removed the words ‘systematic review’ from the title and all other places in the manuscript.

(3) The way to select studies is awkward – why odd numbered months? Why randomly dividing the articles into 10 batches, then randomly selecting 7 batches? In what sense to you mean that it is a probability sample? What is more, what will this sample be representative? Again, due to the ill formulated research question, it would be hard to develop meaningful searching/sampling strategies.

The sampling scheme was devised to result in a random sample of randomized controlled trials from the included journals and years. We assumed that an RCT would have equal probabilities of publication in both even and odd months, and only pulled from odd months to decrease our search burden. The RCTs were randomized into 10 batches (and then 7 batches randomly selected) because we were uncertain how many articles we would be able to review within our capacity; by batching the sample and selecting batches randomly from the
From Reviewer 3:

(1) “Systematic review” should be removed from the title. Perhaps something like “a literature sample” would be better.

We agree that referring to our study as a systematic review may not be appropriate, as our search was not exhaustive. We thank the reviewer for the suggestion, and have removed the term “systematic review” from the title and the manuscript text.

(2) I found the identification (page 8) of the trial cumbersome and would like to know why the authors did not use an RCT filter as part of the search strategy?

Initially, we did use an RCT filter to identify RCTs eligible for inclusion in our study. However, we felt that our specific sampling frame (5 journals, odd months, known years) provided us with a unique opportunity to be as thorough as possible and examine individual articles by hand prior to exclusion. It is interesting to note that our RCT filter and individual article examination resulted in the same number of RCTs eligible for inclusion (n=541).

(3) The authors may want to note that all journals except NEJM endorsed CONSORT in 1994. NEJM came on board in 2004 or 2005.

We’d like to thank the reviewer for highlighting this important historical development. As we noted in our response to Reviewer #1, item 1c, we chose the three specific years (1994, 1999, and 2004) with the goal to capture the impact of the SORT and CONSORT Statements. Please note that while 4/5 of the journals endorsed SORT/CONSORT in 1994, we believe that it is unlikely the articles published in 1994 were impacted directly by the SORT/CONSORT Statements, given the usual lag in publications. For an RCT to be published in 1994, it was likely completed in 1993 or earlier – before the recommendations were endorsed by the journals and also before the recommendations were first published in 1994/1996. For this reason, we hypothesized that we would observe an increase in HTE analysis (defined as those analyses utilizing an interaction test or a test of heterogeneity), and a decrease in subgroup analysis (where no test for interaction or heterogeneity is used). In 2001, the
<table>
<thead>
<tr>
<th>CONSORT recommendations elaborated on their initial recommendations, and further highlighted the correctness of using a test of interaction in subgroup analyses.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4) On page 15 (and figure 2) the authors propose some reporting guidance. How was this guidance developed – like the recent STROBE statement or something more informal? If the latter the authors might want to note this as a potential limitation of the guidance. Similarly, the authors might want to elaborate on how the proposed reporting guidance fits into the current CONSORT reporting guidance for randomized trials – particularly the CONSORT explanatory paper.</td>
</tr>
<tr>
<td>We appreciate the reviewer highlighting this limitation. Our recommendations are not based on a formal methodology. We have added the following text to the limitations section of the discussion to explain this: “Our recommendations for authors and editors are based on an informal procedure, and should be interpreted in light of this limitation. Further refinement of the recommendations may be necessary before adoption by editors and authors.”</td>
</tr>
<tr>
<td>Our recommendations underscore the CONSORT recommendations that proper HTE analysis requires a test for interaction or heterogeneity and that all analyses should be labeled as a priori or post hoc. However, our recommendations regarding clinical significance of results is a departure from the recommendations in CONSORT. We agree with the reviewer that it would be helpful to have this differences labeled in the text and have corrected the text to read:</td>
</tr>
<tr>
<td>“We present complementary guidelines in Figure 2, and emphasize the role of journal editors in setting appropriate standards for subgroup analysis reporting. There are three major principles to our guidelines. The first and second underscore the CONSORT recommendations [18] that (1) proper HTE analysis requires statistical tests of interaction or heterogeneity and (2) that all variables examined for HTE should be labeled as prespecified or post hoc and reported in the body of the paper or in an electronic appendix. Finally, we advocate for authors explicitly considering the clinical and statistical significance of results obtained, and provide recommendations for future research and clinical practice.”</td>
</tr>
</tbody>
</table>