Author's response to reviews

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

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: 3 Date: 4 March 2009

Author's response to reviews: see over
October 3, 2008

Doug Altman, PhD
Curt Furberg, MD, PhD
Jeremy Grimshaw, MBChB, PhD, FRCGP
Peter Rothwell, MD, PhD, FRCP

Editors-in-Chief, Trials

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

We are pleased to submit our revised manuscript entitled “Dealing with Heterogeneity of Treatment Effects: Is the Literature Up to the Challenge”. We appreciate the comments of the editors and reviewers, and believe that our manuscript is significantly improved after careful attention to these comments. A point-by-point description of all changes made is included in this cover letter.

We appreciate your consideration.

Sincerely,

Nicole B. Gabler, MPH, MHA
PhD Candidate, Graduate Group in Epidemiology
University of California, Davis
Center for Healthcare Policy and Research
2103 Stockton Blvd., Grange Building, Suite 2224
Sacramento, CA  95817
From reviewer 1:

(1) Though the authors write (p5, bottom paragraph) that “the most direct approach” of assessing HTE “is the n-of-1 clinical trial, which randomizes treatment episodes in a single patient to different treatment, to identify the best treatment for a single patient,” they seem to ignore the fundamental problem of assessment of individual variability that is associated with usual parallel group trials.

This problem is clearly exposed by Stephen Senn and coauthors in a number of articles (see, e.g., Senn S. Individual response to treatment: is it a valid assumption? BMJ 2004 Oct 23;329(7472):966-8; Senn S. Controversies concerning randomization and additivity in clinical trials. Stat Med. 2004 Dec 30;23(24):3729-53 ($4.8, p3745-6); there also other reference).

In addition to inter-individual variation in treatment response (between patients or subgroups of patients), there may be intra-individual variation. As a matter of principle, this cannot be investigated in parallel group designs, if each patient is exposed to a single treatment only once. What we are used to call subgroup analysis fails to distinguish between two causes of heterogeneity of treatment effects:

(i) variation due to different individual patients (e.g., genetic differences), and
(ii) variation due to intra-individual reasons (e.g., a drug may or may not be working in this patient on a given day).

Most investigators when analyzing subgroups do not think of the latter source of variation. Neither do the authors of this review.

Though I admit that this is primarily a problem of the trials under study, I propose that the authors should read the papers cited above and then discuss this more carefully. There are several places in the MS suited for this: First, on p5 where n-of-1 trials and their limitations are mentioned. Secondly, on p6 where they state that subgroup analyses are perilous: here it must be mentioned as a third point (reported subgroup effects may be spurious because not only of multiplicity, but also of random intra-individual variation). Thirdly, the issue is missed on p15, bottom paragraph beginning with, “The biostatistical literature tends to view subgroup analysis skeptically, often citing the dual problems of multiple statistical comparisons and low power.”

The reviewer makes an excellent point. Subgroup effects may be due to random intra-individual variability, which is impossible to estimate in a parallel group trial. Additional text highlighting this source of variability has been added to the three places suggested by the reviewer. New text (preceded by the old text) reads:

Page 5:

Old:

However, n-of-1 trials are applicable to a relatively small subset of conditions and treatments.

New:

However, n-of-1 trials are applicable to a relatively small subset of conditions and treatments and are subject to random within-patient variability (thus requiring a careful design and repeated crossovers).
Subgroup analysis can be perilous. Real effects can be missed because of inadequate statistical power, and reported effects may be spurious because of the performance of multiple statistical tests.

Subgroup analysis can be perilous. Real effects can be missed because of inadequate statistical power, and reported effects may be spurious because of the performance of multiple statistical tests and/or due to random intra-individual variability.

Inconsistent reporting of subgroup analysis not only impairs recognition of patients who may respond better or worse than average, but also impedes hypothesis generation and stifles future research. Although tests for heterogeneity and interaction have low power for detecting modest but potentially clinically important subgroup differences, they represent a conservative approach and provide a brake on the tendency to over-interpret observed subgroup differences. However, it is important to conduct HTE analysis with caution and not over interpret the results. It is necessary to recognize that post hoc HTE analyses are for hypothesis generation and to aid in the design of future confirmatory studies, that significant effects may be a result of intra-individual variability, and that the results of such analyses should not be used to promote different treatment recommendations.

(2) In general, in my view the trade-off between the chances of analyzing subgroups and the problems associated with subgroup analyses (that are often not even pre-specified) is not worked out appropriately. The authors recommend to increase the number of subgroup analyses. However, they do not sufficiently stress that these have to be done with caution.

We agree with the reviewer that all post hoc subgroup analyses must be interpreted with caution. It is not our intention to advocate for post hoc analyses without clearly stating that these analyses are to be used for hypothesis generation, and later followed by appropriately powered confirmatory analysis. The text below has been altered to reflect this (original text followed by new text):

Inconsistent reporting of subgroup analysis not only impairs recognition of patients who may respond better or worse than average, but also impedes hypothesis generation and stifles future research. Although tests for heterogeneity and interaction have low power for detecting modest but potentially clinically important subgroup differences, they represent a conservative approach and provide a brake on the tendency to over-interpret observed subgroup differences.

Inconsistent reporting of subgroup analysis not only impairs recognition of patients who may respond better or worse than average, but also impedes hypothesis generation and stifles future research. Although tests for heterogeneity and interaction have low power for detecting modest but potentially clinically important subgroup differences, they represent a conservative approach and provide a brake on the tendency to over-interpret observed subgroup differences.
Inconsistent reporting of subgroup analysis not only impairs recognition of patients who may respond better or worse than average, but also impedes hypothesis generation and stifles future research. Although tests for heterogeneity and interaction have low power for detecting modest but potentially clinically important subgroup differences, they represent a conservative approach and provide a brake on the tendency to over-interpret observed subgroup differences. However, it is important to conduct HTE analysis with caution and not over interpret the results. It is necessary to recognize that post hoc HTE analyses are for hypothesis generation and to aid in the design of future confirmatory studies, and their results should not be used to promote different treatment recommendations.

(3) The name of the first author should be written consistently

This has been changed throughout the manuscript.

(4) p10, end of second paragraph: The description of a forest plot (“a graph showing results for each subgroup as a square bisected by a horizontal line representing its confidence interval”) seems to me both clumsy and difficult to understand – why not simply say that the plot for a given outcome shows the confidence intervals of different subgroups.

We agree with the reviewer that our definition may have been overly complicated. The new definition of a forest plot reads, “a graph depicting subgroup results as point estimates [boxes] and confidence intervals [lines]).”

(5) p13, beginning and end of bottom paragraph: What is the bivariate analysis? There is not explained before.

The bivariate analysis was an unadjusted comparison examining the relationship of HTE reporting by study characteristics. The following text was added to the beginning of the data analysis section: “The relationships of HTE reporting with study characteristics were assessed in bivariate analyses by using Pearson $\chi^2$ tests or Fisher’s exact test when sample size was small, while trends (where appropriate) were assessed using the Mantel-Hantzel $\chi^2$ test for trend.”

(6) p19, Author’s contributions: manuscript instead of manuscripts.

This has been corrected in the text.

(7) p15, second paragraph: delete “p-values or”

This has been corrected in the text.

(8) p17, line 3: Perhaps also the statistical ethos tends to be more conservative in Europe.

The sentence in the manuscript has been altered to read “norm” instead of “medical ethos.”

(9) p17, second paragraph: In my view, HTE analysis should be reported not merely because it has been done, not to mention because it is significant, but when it is pre-specified.
We agree with the reviewer that it is necessary to report HTE analysis when it has been pre-specified. The sentence in the manuscript has been modified to reflect this, and now reads, “However, even non-significant data are useful for the purposes of hypothesis generation and, arguably, authors should report any HTE analysis, significant or not, and especially when the analysis is pre-specified.”

From reviewer 2:

(10) It is not appropriate to state “subgroup-only analyses represent missed opportunities on the path to understanding HTE: with minor effort, studies that reported subgroup-only analyses could have conducted formal HTE analyses and provided direct information on HTE.” -- If a subgroup analysis does not show difference at all, there is no need to conduct a formal HTE analysis. The purpose is to understand whether there is any treatment difference between subgroups, instead of performing and reporting formal HTE analysis. A high p-value with an estimate of wide confidence intervals from HTE analysis won’t provide any more information on HTE than subgroup analysis. Same comments apply to page 15, second paragraph. It is not reasonable to imply that every article needs to report a HTE.

Actually the distinction between the two types of analysis is not very useful here, without considering the purpose of studies. If one goal of the study is to test the difference among subgroups, HTE should be conducted. If an interaction is significant, then the authors should look at the difference among subgroups. If the study is not designed to do this, the study should not be criticized for not doing this.

On the other hand, I agree that a test of an interaction is a better way to formally compare the subgroups. However, many studies are not designed to make formal comparison but more interested in just obtaining an estimate for each group.

We agree with the reviewer that the distinction between subgroup analysis and HTE analysis is not always clear. Our goal is clarify the distinction. If a significant interaction exists (as determine by a formal test of interaction), then subgroup-specific results should be reported since the treatment effect is different in Group X and Group Y. However, in this study, subgroup-only analyses were those that reported subgroup-specific effects without ever conducting a formal test of interaction. In this case, concluding that there are no differences between subgroups is essentially conducting HTE analysis by ‘eyeball’ and may lead to incorrect decisions. It is easy to conduct formal HTE analyses. If researchers are reporting subgroup-specific results because they are ‘interested in just obtaining an estimate for each group’, then we argue that it makes sense to take the analysis one step further and conduct a formal HTE analysis, even though it may not provide any more information for that specific study. It is important for researchers to conduct formal HTE analyses in order to help generate new hypotheses that can be tested in future (appropriately powered) confirmatory studies.

It is not our intention to criticize researchers for not conducting these HTE analyses, but instead we are trying to encourage them to conduct and report HTE analyses for hypothesis generation. To that effect, we had added the following:

“Subgroup-only analyses represent missed opportunities on the path to understanding HTE; with minor effort, studies that reported subgroup-only analyses could have conducted formal HTE analyses and
provided direct information on HTE. The additional step (of conducting a formal HTE analysis as opposed to a subgroup-only analysis) is important because it will provide important hypothesis-generating information for future studies.”

(11) Discussion, “The results suggest that reporting on HTE is far from routine, and was only marginally better in 2004 than in 1994” – It is not appropriate to imply that reporting HTE needs to be routine, without considering the goal and the scope of the study. Based on logistic regression, there was no significant differences among years, either.

We appreciate the reviewer’s insightful comment and agree with the assessment. While it should not be implied that HTE reporting should be routine, we do believe that it should be used more commonly in order to facilitate hypothesis generation. The text of the discussion has been altered as follows: “The results suggest that reporting on HTE occurs in less than one-third of studies published in prominent general medical journals, and was only marginally better in 2004 than in 1994.”

(12) “In addition, when the purpose of HTE analysis is hypothesis generation, it may make sense to accept a more lenient standard of statistical significance (i.e., 0.10).” -- This is an inappropriate suggestion: one may be willing to conduct another study to test an interaction with p-value = 0.10 in an earlier study, however, it would unnecessarily increase Type I error by using a different standard for significance, considering HTE are much likely post hoc tests.

We do agree with the reviewer that there is a tradeoff between power and Type I error, but we do not entirely understand the comment outside this point. Examining post hoc analyses at the 0.10 level can be used to identify plausible exploratory, which can then be assessed at the 0.05 level in an appropriately powered second study.

(13) Page 17, first limitation could well invalidate the results from large studies. There are numerous cases that large trials would generate multiple studies with major findings published in best journals, and other results (including subgroup/HTE) published in more specialized journals. Sometimes, the results are too many to be published in one article. Sometimes, articles were devoted to subgroup analysis only. Therefore, the estimate of prevalence would be biased here. Did the author do a search to check whether the large trials have other publications?

We agree that this is a limitation. We performed a sensitivity analysis assuming that the articles most likely to report HTE elsewhere are large (>500 participants) multicenter trials. Twenty-three trials that did not report subgroup or HTE analysis fall into this category. Even if all of these trials had reported HTE, the percent of articles reporting HTE would only increase to 36%. Including the additional 5 trials that included >500 participants (but were not multicenter), and the 13 that included 250-499 participants (and were multicenter), would only increase the HTE reporting percentage to 42%. We acknowledge that our estimate of HTE reporting as detailed in the manuscript is mostly likely an underestimate, but we can estimate the upper limit of HTE reporting at 42%.

(14) Rewrite sentence: “The prevalence of HTE is unknown, but many observers believe it is nontrivial” – prevalence of HTE could not be clearly defined in the first place.
The sentence has been rewritten as requested: “The prevalence of HTE is unknown and perhaps unknowable, but highly variable treatment response rates for many common conditions suggest it is substantial.”

(15) Provide a description on the position/standards of CONSORT on subgroup analysis as it is used to interpret differences among time points.

The revised CONSORT standards made a clear delineation between HTE and subgroup analysis, and cited an interaction test as the appropriate method of assessing interaction. The statement went on to discuss how inferring interaction from subgroup-specific p-values can lead to spurious findings and is incorrect. For this reason, we hypothesized that HTE analysis would increase in 2004 (compared to 1999 and 1994), with a concomitant decrease in subgroup analysis.

To clarify CONSORT’s position, the following sentence was added to the manuscript: “The revised CONSORT standards made a clear distinction between subgroup and HTE analysis, citing a test of interaction as the correct, and stronger, analytic technique. The Statement emphasized the incorrectness of comparing subgroup-specific p-values as a method of inferring treatment heterogeneity. Thus, we expected to see in increase in HTE analysis subsequent to 2001, with a concomitant decrease in subgroup analyses.”

(16) Page 6, “these studies do not provide information on HTE per se.” – they do provide information on HTE, just not a very quantitative way.

We altered the sentence to read as, “These studies do not provide quantitative information on HTE per se.”

(17) Page 6, last sentence, in individual studies, it is a test of interaction, not called “test of heterogeneity”.

Tests of heterogeneity are a specific type of interaction test, in that it examines interaction between patient characteristics and treatment response. Tests of interaction, on the other hand, can be between two covariates (neither of them treatment). It is for this reason that we use the term ‘test of heterogeneity.’

(18) Page 11, Data analysis: Pearson’s chi-square test instead of Wald test for categorical variables in the contingency tables? Clarify “differences across odds ratios”? Clarify “we separately examined the association of covariates other than sample size in trials above and below the median sample size (262 subjects).”?

We agree with the reviewer that Pearson’s chi-square test should be used instead of Wald test for categorical variables; this was mislabeled in our original manuscript and the labeling has been corrected in the methods section and in Table 2.

Additionally, we agree that the descriptor, “differences across odds ratios” may be misleading. This text has been changed to, “Significance of study characteristics in relation to use of HTE analysis was
assessed with a Wald $\chi^2$ test.”

The sentence, “we separately examined the association of covariates other than sample size in trials above and below the median sample size (262 subjects)” has been changed to, “To further explore HTE reporting characteristics, we separately examined the association between study characteristics (other than sample size) and HTE reporting in articles above and below the median sample size (262 subjects).”

(19) Provide information on how many articles have multiple trials – such articles are likely not reporting HTE due to space limitation.

Twelve articles reported on more than one study. Of these, ten articles reported on two trials each, and two articles reported on four trials each. Five of the twelve articles reported HTE for at least one of the included trials.

Additional information about the number of articles that reported on more than one trial has been added to the results section:

“Twelve articles reported on more than one trial: ten articles reported on two trials each, and two articles reported on four trials each.”

(20) Page 13. “Despite increased recognition of the value of multivariable risk indices in HTE analyses (8) only three studies (37-39) evaluated outcomes of treatment stratified by multivariable risk.” – Reference 8 was published much later than references 37-38.

While we did cite one of the more recent articles on the strengths of multivariable risk indices, numerous previous examples exist in the literature. These additional citations have been added to the text. All of the examples below were published prior to 2004 (and one before 1999).


Kent DM, Ruthazer R, Selker HP: Are some patients likely to benefit from recombinant tissue-type plasminogen activator for acute ischemic stroke even beyond 3 hours from symptom onset? Stroke 2003, 34:464-467.


(21) Logistic regression: did the authors try to build a parsimonious model with only significant/important predictors? Page 14, first paragraph, the purpose of the study is to look at trend
over time, the results on time were not discussed here.

The variable selection for our logistic regression model was based on theoretical importance and past literature. The analysis was pre-planned and specified prior to fitting the data.

While examining trends over time was one of the purposes of the study, our logistic regression did not indicate that any significant trends over time existed. The only two significant results (sample size and journal) are discussed on p 14. Text has been added to indicate that the time trend was not significant.

(22) Discussion, clarify “This review of 319 RCTs published in five prominent general medical journals is the most comprehensive to date, and the only one that examines trends over time”? Of what?

The sentence has been altered to read, “This review of 319 RCTs published in five prominent general medical journals is the most comprehensive to date, and the only one that examines trends of HTE reporting over time.”

(23) Page 17, line 2 “the dominant medical ethos??”

This is a good point. Based on another reviewer’s comment we have changed the phrasing to “norm.”

(24) Page 5, the distinction between quantitative and qualitative HTE is too loose, not considering the precision of estimated effects.

We respect the reviewer’s position. However, the distinction between quantitative and qualitative HTE is based on the values of the true parameter, not the estimated parameter. This conceptual definition is not affected by precision because it is based on the value of true parameters. When estimates are being used to determine qualitative or quantitative HTE, the definition then becomes a hypothesis that requires testing.

(25) Clarify Allen Roses’ statements.

Allen Roses’ statement is used as a support and example of the previous sentence, where we state that ‘highly variable response rates for many common treatments’ indicate that HTE is substantial. Another way of putting this is that NNTs (number needed to treat) for common treatment almost always exceed 2, meaning that fewer than half of treated individuals actually benefit. Appropriate HTE analysis may help elucidate groups of individuals in which average drug response rates are high (or low).

(26) Page 5, rewrite “which randomizes treatment episodes in a single patient to different treatments”.

As requested, the sentence was rewritten as, “which assigns individual patients to receive alternative treatments in a randomly predetermined sequence.”

(27) One potential interesting predictor of HTE would be the limitation on number of words for each paper from different journals – if a journal has a more restricted word limit, the authors would have no space to report results on subgroups.
We agree with this statement, and argue on p 17 that the use of a Forest Plot may be a method to present HTE results without sacrificing available text. The five journals included in our review have similar requirements for article length, according to information for authors provided by each journal.

(28) Use more common language and avoid words like “perilous”, “in a bind” … etc.

We believe “perilous” best conveys our meaning, but agree “in a bind” is colloquial and was replaced with “presents difficulties.”

(29) Appropriate standards for subgroup analysis reporting should be encouraged; however, it is hard to make connection with the study design and results.

While we fully agree that our results do not directly speak to reporting standards, they lay the foundation by (sketching) the magnitude of the problem and highlighting its importance.