Reviewer's report

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

Version: 2 Date: 3 December 2008

Reviewer: ROchelle Fu

Reviewer's report:

Major Compulsory Revisions

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.

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.

3. The way to select studies is awkward – why odd number months? Why randomly dividing the articles into 10 batches, then randomly selecting 7 batches? In what sense do 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.

4. 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 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 for 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.

5. The authors should have applied more restricted criteria for selecting studies. For example, instead of a sensitivity analysis on sample size, there should be a pre-set sample size for study inclusion. The smallest sample size of the included studies is 6 and 10% studies have sample size less than 37 and 30% less than 124 – such studies are very limited for HTE and usually not appropriate for HTE. For example, one may suspect the quality of those studies with median n = 37 that examined up to 9 covariates. Even if for relatively large studies, when the event outcome is rare, the ability of doing HTE is also very limited.

6. 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.

7. “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.

8. 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?

- Minor Essential Revisions

1. 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.

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

3. Page 6, “these studies do not provide information on HTE per se.” -- they do provide information on HTE, just not a very quantitative way.

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

5. 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).”?

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


8. 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.

9. 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?

10. Page 17, line 2 “the dominant medical ethos??”

- Discretionary Revisions

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

2. Clarify Allen Roses’ statements.

3. Page 5, rewrote “which randomizes treatment episodes in a single patient to different treatments”.

4. 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.

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

6. Appropriate standards for subgroup analysis reporting should be encouraged, however, it is hard to make connection with the study design and results.
**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.