Reviewer's report

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

Version: 2 Date: 3 November 2008

Reviewer: Gerta Rucker

Reviewer's report:

Summary of the ms

The authors present a review looking at a random sample of 319 RCTs published in five major medical journals. The outcome of interest is whether the articles reported an analysis of potential heterogeneity of treatment effects (HTE) (defined as a formal test of treatment-by-covariate interaction), subgroup analysis only, or neither. The authors conclude that HTE is too often ignored or incorrectly analyzed.

Major Compulsory Revisions

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 treatments, 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 references).

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 analysing 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 sceptically, often citing the dual problems of multiple statistical comparisons and low power".

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.

Minor Essential Revisions

The name of the first author should be written consistently (List of authors: Nicole Bloser Gabler; Affiliations: Ms. Gabler; Address: Nicole Bloser; Author's contributions: NG).

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.

p13, beginning and end of bottom paragraph: What is the bivariate analysis? This is not explained before.

p19, Author's contributions: manuscript instead of manuscripts.

Discretionary Revisions

p15, second paragraph: delete "p-values or"

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

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.

Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.