Reviewer's report

Title: Multivariable Risk Prediction Can Greatly Enhance the Statistical Power of Clinical Trial Subgroup Analysis

Version: 1 Date: 3 August 2005

Reviewer: Tim Peters

Reviewer's report:

General

1. Overall I found this a clear and helpful manuscript, with useful messages both to trialists and consumers of trial results, covering for the most part realistic scenarios to provide examples to support the authors' general argument. My concerns relate mainly to variations on the theme of a tendency for the authors to somewhat over-state their case (at least in this referee's opinion, even as one broadly sympathetic to their proposal), especially given some reservations about quite how broad some of their assumptions/scenarios really are in practice. The comments below could all be readily addressed by the authors.

-----------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

1. The main ways in which I felt the authors were over-stating their case are in terms of: (a) their over-confidence that multivariable risk models would be available (notwithstanding Table 4 – see below); (b) the assumption that even where available these models would be applicable/relevant in any given situation; (c) their implicit presumption that such tools if available would be useable in practice for future patients (in fact this receives virtually no attention in the paper); (d) the concluding sentence in the Abstract, which despite the first two words could be interpreted as meaning that risk-stratified subgroup analyses should be a component of all clinical trials. The following points cover these issues in detail.

-----------------------------------------------------------------------------------

2. The Abstract should generally be more balanced in respect of these concerns, and in particular the last sentence should encompass the fact that subgroup analyses should be embarked upon with caution in all instances (regardless of the statistical methods used and their power). Where they are performed they should reflect reasonable prior expectation of differential effects (preferably, as pointed out in the Introduction, with clear distinction between influences of scale and more fundamental, clinical, effect-modification). The importance of this is emphasised by the point that it is debatable to say the least as to whether clear prior expectation can be phrased in terms of a specific risk assessment as opposed to either a general concept of ‘baseline risk’ and/or to individual components/risk factors. The Discussion already is much better than the Abstract in respect of a balanced perspective, though some of these comments also apply there.

3. Central to the authors' argument is that the multivariable risk-stratified approach is not data-driven in the sense that the risk score should be derived independently of the trial data set itself. This seems to me crucial to avoid the worst excesses of false positive risk that might offset any power advantages. First, this point could itself be more explicit in the paper, especially as the comment at the end of page 5 is only correct if the risk stratification is independently derived. In any case, in the second sentence of paragraph 2 on page 4 the word “preferably” seems contradictory to this point and indeed the opening phrase of the sentence itself and other parts of the paper.
4. In the next paragraph the point regarding the reanalysis of the GUSTO data is well made, but seems to ignore the practical issue of when the “independently derived and validated model” was available. There is a danger of appearing wise after the event in arguing (however implicitly) from this example that the original analysis should have heeded the authors’ conclusion in their Abstract. More generally, in my view the statement “easily identifiable” at the end of the following paragraph is an over-statement, since in practice this is surely a complex issue covering relevance, interpretability and practicability etc. as well as availability. Likewise in the opening paragraph of the Methods section (and the penultimate line of page 12) as well as in the second sentence of the Discussion, I felt that the availability and predictive ability of risk prediction tools was overstated in respect of randomised trials generally, and their problems rather ignored.

5. The only major concern I have regarding the methods themselves is that in our experience in this type of work, 2000 iterations for the simulations seems a somewhat small number. In part this is presumably limited by the package used (the implication from the references is that this was Stata, but this should be explicit in the paper), and indeed we had to use a more powerful programming language for a sufficient number of iterations. Admittedly the acceptable margin of error around the estimates (of power here) will vary according to their magnitude — and the general messages of this paper are unlikely to be greatly affected by the imprecision resulting from just 2000 iterations. However, the number used should be justified in the paper, at the very least in general terms and preferably also with some indication of the precision of the derived estimates of power. Only if this is adequate for the authors’ argument should the number used in this paper be considered sufficient.

__________________________________________________________

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

1. The statement in the first half of the first sentence of the Abstract read oddly — do the authors not mean “… treatment benefits apply to specific clinical subgroups” rather than “all subgroups”?

2. Tables 2 and 3 and the accompanying text were helpful and clear if a little unoriginal. The evidence from Table 4 was not wholly reassuring to this reader — for instance the ‘reverse weighting’ used by the authors did not seem that extreme to me. There might be benefits in considering other scenarios here.

3. While perhaps a personal matter of semantics, I found the use of the term “baseline” somewhat confusing in this paper. At times it is used in what I consider to be the most helpful in the context of a clinical trial — namely, observations made at the start of the study (prior to randomisation). However, in the Glossary, the figure and at numerous points in the text, the term is used as a synonym for the control ‘state’/group in a trial. This seemed to me to be an unnecessary potential confusion, since of course there should in general be no expectation that the control group remains static from (true, pre-randomisation) baseline to the (post-randomisation) follow-up at which time the outcome is assessed. I found the rest of the Glossary very clear and helpful.

__________________________________________________________

Discretionary Revisions (which the author can choose to ignore)

1. Generally the style is very good, but there are a number of occasions when a semi-colon would be much clearer as well as more correct before the word “however” (i.e. suggest “…; however, …” rather than “…, however, ….”). Those I spotted were: abstract line 2; last line page 10; page 12, para 2, line 8; page 13, para 2, line 4; and page 14, para 1, line 19. Similarly, before “then” in page 4, para 2, line 8. In addition, ideally the authors should be consistent about hyphenation of “risk-stratified”, there’s a superfluous full stop at the end of the first paragraph on page 7 and a duplicated “each” on
Lastly, other typos were: word “Area” missing from last footnote of Table 1; apostrophe missing from “subject’s” in Panel B of the figure; just above this, “RFs” not “RF”; and a full stop missing from sentence 2 of the figure legend.

2. Reference 21 has my initials (TJ) missing. More fundamentally, it’s unclear why the word “(Abstract)” is given following the full reference. This may be because this was what the authors had available but it seems unnecessary. While of course I would support reference to the full HTA report, in addition it may be helpful to reference a perhaps more widely accessible publication that resulted from our work on this report – Brookes ST, Whitley E, Davey Smith G, Mulheran PA, Egger M, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004;57:229-236. Naturally I acknowledge an interest in the citation!

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:
I declare that I have no competing interests.