Author’s response to reviews

Title: An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect.

Authors:

Patrick Royston (pr@ctu.mrc.ac.uk)
Mahesh K B Parmar (mp@ctu.mrc.ac.uk)

Version: 2
Date: 9 July 2014

Author’s response to reviews: see over
Reviewer 1

Title: An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect.

Version: 1

Date: 31 March 2014

Reviewer: Honghua Jiang

Reviewer’s report:

The authors proposed a joint test for the treatment comparison, combining the logrank test and Grambsch-Therneau test of proportional hazards due to the fact that non-proportional hazards may be common in clinical trials. When the hazards are not proportional, the joint test provides greater power to detect the treatment difference compared to the logrank test only. However, the joint test may also result in a false claim of significance when actually there is no or little treatment difference but the hazards are significantly non-proportional (for example, treatment effect cross at some time point). So the joint test has major flaw to be used as a method for treatment effect comparison.

Response: the point of the joint test is that it is sensitive to straightforward and also to more “complex” treatment effects. We have included the following text as the penultimate paragraph of the Discussion:

“A key feature of the joint test is that it is sensitive to simple and also to more ‘complex’ treatment effects. In the latter case, assuming the result is not a type 1 error, the test is indicating there is a genuine difference between the survival curves. Even if the overall treatment effect, considered over the entire follow-up time of the trial, is small, the difference between the arms may still be of clinical and/or scientific interest and importance. For example, the difference in the survival curves between the treatment arms may suggest possible mechanisms of action of the treatments.”
Reviewer 2
Reviewer's report
Title: An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect.

Version: 1

Date: 3 March 2014

Reviewer: Walter Gregory

Reviewer's report:
This is an interesting and useful paper, addressing an important issue that is often overlooked in trial design and analysis. While statistical analysis plans usually specify checking that hazards are proportional, it is less clear what to do when they are not, and a variety of ad hoc approaches are often applied. This paper details a logical and systematic approach to dealing with these issues.

Specific comments as follows:

Discretionary revisions:

1. The statistical methods are clear and well detailed in this manuscript. However, it would be helpful to include additional discussion on the circumstances in which they should be used. How often will hazards depart from proportionality? How can this be reasonably evaluated so as to know when the trial should be powered in this way? Should this be done using previous data, or an understanding of the mechanisms of action of the treatments (see next point below)? The authors have documented that there will be some, not inconsiderable, loss of power, or alternatively an increase in sample size, using this approach when the hazards are in fact proportional. Do the authors believe that this loss of power is a reasonable price to pay for being able to accommodate non-proportional hazards, and that this approach should therefore be adopted routinely? They also give an alternative approach of relaxing the significance level of the joint test to achieve the same power as the log-rank test, but say they have a ‘slight preference’ in favour of simply powering according to the joint test. Do they see circumstances in which relaxing the significance level is appropriate? It would be helpful to have such additional information and recommendations.

Response: We thank the reviewer for these salient questions. To address them, following passage has been inserted as the final paragraph of the Discussion:

“We are not suggesting that the joint test be adopted routinely. Primarily, we suggest that the trialist choose the preferred test according to the perceived modes of action of the treatments being compared. If the modes are obviously different, for example surgery versus a more conservative approach such as watchful waiting or a non-surgical therapy, the hazard functions will probably differ markedly in shape and non-PH seems more likely. The joint test may then be a good choice. If rather similar treatments are involved, such as various chemotherapy regimens, non-PH may seem less likely and the logrank test may be best. There may be indications of the extent and nature of
non-PH from earlier trials or, in cancer for example, from other cancer types in which the treatment has been evaluated. Another consideration is judging how close to PH the ensuing survival curves are likely to be. If a treatment effect is expected to emerge relatively soon after randomization, non-PH is likely to be mild and the logrank test will be the more powerful. If the effect emerges much later in follow-up, the joint test is likely to be more powerful.”

Regarding relaxing the significance level to preserve power, we have downplayed this idea as being unlikely to appeal to most trialists. Nevertheless, we have retained the required calculations in the manuscript for information.

2. The authors state (page 6) that ‘In the emerging era in which the PH assumption often seems to fail [6, 7] …’ the two references provided are both of trials in which the PH assumption is inadequate. This is not sufficient justification for stating that the PH assumption OFTEN fails. I suggest that they either give a wider variety of examples, or perhaps instead describe why they think this might be the case, for instance due to the number and widely different types of action of current drugs and treatment approaches, that cause, for instance, growth arrest in tumours, that might lead to unusual patterns of relapse and survival. These are only examples; the authors should consider what they believe to be the reasons behind the current failures of the PH assumption, or at least speculate on such reasons in the discussion.

Response: This is a fair point. We have softened the wording to “... PH assumption may fail ...”. In response to this, we have inserted the following text as the second paragraph of the Background:

“We believe that there may be two reasons that non-proportional hazards are being detected more frequently nowadays. First, phase III trials are generally much larger today giving more power in any given situation to detect non-proportional hazards. Second, with the biological revolution there are many new therapies being evaluated, having different modes of action. For example, monoclonal antibodies have been evaluated for treating many different types of cancer. They are given for a defined period of time (often 1 or 2 years) and then stopped. It is entirely plausible that the effect of the intervention might persist during the treatment period but then diminish gradually afterwards. Such behaviour, which in fact is what has been seen in the examples given above, would lead to non-proportional hazards.”

3. Although it is referenced, it would be helpful (page 6) to have a very brief description/definition of the restricted mean survival time for readers who are not familiar with this approach. Similarly, a very brief description of Cox’s suggestion for modelling a time-dependent hazard ratio (page 10) would be indicated.

A brief description and definition of RMST have been included in the section Graphical presentation. We have added a sentence describing Cox’s suggestion in the section Model for a time-dependent treatment effect.
Minor essential revisions:

4. The authors should give some justification for the number of replicates used in the simulations - 5000 seems relatively small. Standard errors for key parameters are often used to give an idea of the variability from the chosen number of replicates, and therefore to justify this choice.

*Monte Carlo standard errors for power estimates have now been given in Table 4. They do not exceed 0.006.*

Level of interest: An article of importance in its field

Quality of written English: Acceptable

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

Declaration of competing interests:

I declare that I have no competing interests