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.

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.

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.

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.

**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