Reviewer’s report

Title: Assessment of the consistency and robustness of results from a multicenter trial of remission maintenance therapy for acute myeloid leukemia

Version: 2 Date: 16 November 2010

Reviewer: Ian Marschner

Reviewer’s report:

Overall comments:
This is a nice paper providing a good mix of useful conclusions in a specific context, along with a number of general issues in clinical trial methodology. The paper has two main objectives: (i) to establish the consistency and robustness of results across subgroups from a specific trial (of HDC/IL-2 in AML) and (ii) to illustrate general methodological approaches to assessing consistency and robustness that may be relevant for other trials. I think the paper succeeds with the first objective. I think it only partially succeeds with the second, and have suggested some further discussion below.

Major compulsory revisions:
(1) Nothing major.

Minor essential revisions:
(2) More discussion is needed of how to interpret the suite of statistical analyses that have been proposed for assessing robustness and consistency. The authors have been fortunate in that however one interprets the analyses, this particular trial would be interpreted as consistent and robust. This may not be the case in other studies. For example, 1st paragraph of the Results section: "HRs exceeded 1.00 in 27 subsets out of 28". What is the level of tolerance for this analysis before we start being concerned about inconsistency? Another example is the second paragraph of the Results section and Figure 2: "HRs were larger than 1.00 in 5 out of 7 countries...". Is this what we would expect due to chance variation and how wide a range would we tolerate? Finally, in the section Robustness of Treatment Effects: "the P-value for treatment effect remained significant after elimination of any study center". Some guidance on how to interpret leave-one-out analyses when this does not occur would be useful - does it always mean we have non-robustness or is there a tolerance due to chance variation or other factors? The authors may not have definitive answers to these questions at the moment - they may answer them by pointing to the need for further methodological research in the Discussion (see also point 4 below).

(3) I am not in favour of using an elevated significance level (e.g. 0.1) for tests of heterogeneity (as stated in the Consistency Of Treatment Effects section, 2nd paragraph) to overcome low power. The increase in power is balanced by an
increase in false positives and hence an increase in inappropriate claims of heterogeneity, particularly when multiple tests are being conducted. This is a minor point because it does not affect the results but I don't think this practice should be perpetuated and think the authors should use a conventional level.

Discretionary revisions:

(4) Some of the questions considered under point (2) above, particularly the issue of the tolerable/expected range of treatment effects across countries, have been considered in a recent paper: Marschner IC. Clinical Trials, 2010;10:147-156. The authors should feel under no compulsion to cite this but the methods are designed to aid interpretation of analyses like Figure 2.

(5) The paper includes an excellent example of a validation of surrogacy using a single trial, using methods that are typically applied to meta-analyse multiple trials. There is some confusion in the research community about whether meta-analytic techniques for establishing surrogacy can be applied to strata/centres within a single clinical trial. This study illustrates that they can, and that they can produce quite convincing results - perhaps the authors could emphasise this point more, e.g. in the Discussion.

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.