Reviewer’s report

Title: Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study

Version: 1 Date: 8 September 2009

Reviewer: Jerome Reiter

Reviewer’s report:

Major Compulsory Revisions

1. The authors state that MICE, NORM, and MIX are not as effective as MICE-PMM, particularly for the variables with the skewed distributions. This comment is based on using MICE (etc.) with logarithmic transformations of the continuous variables. However, logarithms may not be the best choice; for example, perhaps cube roots would give more plausible imputations. I recommend that the authors investigate other transformations, for example the Box-Cox transformations, with MICE (etc.). Otherwise, the comparisons of MICE-PMM to MICE (etc.) are not fair to MICE (etc.), so that the broad recommendation to use MICE-PMM may not be justified from the simulations. If the authors find that logarithms are indeed the best transformations, the authors should report this finding.

2. The authors give practical advice in terms of percentage of missing values. The advice would be more effective if given in terms of fractions of missing information rather than percentages of missing values. For example (as the authors know), 10% missing values could correspond to small fractions of missing information, in which case the choice of imputation method does not have great impact. Or, the 10% missing values could all be high leverage points of a regression, in which case the choice of imputation method matters greatly. I recommend casting the advice in terms of fractions of missing information rather than percentages of missing values.

3. On page 17, the authors state, "It is advisable only to use a CC analysis when fitting a Cox proportional hazards model with a reasonably small amount of missing data, for example when less than 25% of the cases have missing data." The exemplary recommendation of 25% as a guideline for using CC in settings outside this simulation is almost surely not appropriate for general settings; see my comments in point 2 about fractions of missing information. I recommend deleting this exemplary guideline. Also, the word "only" could be moved after "CC analysis" so that it is clear that the statement indicates when CC *could* be applied. As written, it could be interpreted as a statement that analysis *should* be done with CC with small amounts of missing data.

4. On page 21, the authors state, "with a MNAR mechanism, MI should not be performed with 25% or more overall missingness." This is too strong a statement;
after all, analysts have to do something with the missing data! Using MI with several MNAR models enables the analyst to assess the sensitivity of results to the missing data. Perhaps the authors meant that MI with a MAR imputation model should not be performed? If so, I recommend correcting this sentence; if not, I recommend deleting this sentence.

Minor Essential Revisions
5. On page 16 in the first paragraph after the section heading, "Results from imposing....," the authors wrote "... due to the fact that a MNAR mechanism was only imposed...." I believe that this should be a MCAR rather than MNAR.

6. Insert a comma after "25% missingness" and before "for all mechanisms" on page 18.

Discretionary Revisions
7. Related to point 1, it would be helpful to include graphical displays of the distributions of the transformed variables, so that readers can judge the normality assumption. Figure 1 is a reasonable place to put these graphs.

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

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.