Reviewer's report

Title: Multiple imputation for estimation of an occurrence rate in cohorts with attrition: a simulation study

Version: 1 **Date:** 31 May 2010

Reviewer: Jerome Reiter

Reviewer's report:

Major Compulsory Revisions

1. The authors state that they used proc MI for imputations, and that they imputed missing E from a logistic regression. My understanding of proc MI is that it imputes from a multivariate normal distribution, not logistic regressions. How did the authors get proc MI to do logistic regression imputation? Or, did they use a sequential imputation strategy like IVEWARE? Or, did they develop their own imputation code and use proc MI to implement the combining rules? I recommend that the authors clarify these issues in the text.

2. The authors' logistic regression for Eij is an incorrect imputation model. The outcomes were generated from a model that specifies one probability for all records with X < .014 and another probability for all records with X > .014. Hence, an imputation model that includes a dummy variable for X < .014 rather than a linear term for X, should give more sensible imputations. If the authors seek to compare a "default" application of MI against a default KM estimator, then this mis-specification is fine, and the authors should explain this in the text. If the authors want to compare MI at its potential best with KM, I recommend that the authors use an imputation model with a dummy variable for X, or perhaps a smooth function of X, rather than a linear one.

Minor Essential Revisions

3. In line 120, the authors state they they impute missing values only when Eij = 0 for previous time periods. I do not understand how this would work with proc MI. It seems like that at any given time point, the imputation model would have to be estimated using only those records that have E_ij = 0 for all previous time points. Can this be done within proc MI, or do you have to set up separate imputation runs for each time point? I recommend some clarification of this issue.

Discretionary revisions:

4. On line 75, the authors ask readers to think of X as age. However, X is defined be N(0, 10) and not to strictly increase with time. Hence, X is not like age. I suggest that the authors simply state that in the IVF example the outcome is delivery and the covariate is age, without calling them E and X.

5. Line 106 states the drop out rates, but it is not possible to know how they are

allocated to the two groups. I suggest adding a sentence after line 106 stating that the allocations for drop out rates are displayed in Tables 1 and 2.

6. Tables 1 and 2 can be combined into one table. The caption can indicate that scenarios 1 - 3 are MCAR and 4 - 7 are MAR.

Level of interest: An article of limited interest

Quality of written English: Needs some language corrections before being published

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

Declaration of competing interests:

I declare that I have no competing interests.