Reviewer’s report

Title: Variable selection under multiple imputation using the bootstrap in a prognostic study

Version: Date: 17 November 2006

Reviewer: Patrick P Royston

Reviewer’s report:

General

The authors propose a method of selecting variables in a dataset with missing observations. It is based on combining multiple imputation with bootstrap resampling.

-------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

1. The authors report 4 methods of analysis: MI, B, MI100+B, and MI10+B. The first, 3rd and 4th of these are sensible. However, the B method is based on resampling of just one imputed dataset. I don’t think that can give sensible results, since the stochastic variation in creating this one imputation is ‘frozen’ in the dataset used for the bootstrap sampling. With a different imputation, different results may (almost certainly will) be obtained from the bootstrap investigation. Therefore, I don’t think the authors can validly conclude anything general from the B method.

2. All effects of continuous variables are assumed linear in the imputation modelling and in the substantive model. However no check was apparently made of this important assumption.

3. Few details are given of how the MICE imputation was done. For example, were the distributions of continuous variables with missing data assumed normal? If so, was this assumption checked? If not, does non-normality matter?

4. A different imputation model was created for each variable with missing data, according to whether the correlation of the target variable with potential predictors was > 0.2 or not. This seems to be pretty arbitrary, and I wonder what justification there is for it? Will using different equations for each variable result in incompatible conditional distributions? Will it induce instability in the imputations? Nothing at all is said about these issues, and we are left in doubt as to how reliable the authors’ approach to imputation is.

5. The authors use a nominal P-value of 0.5 for selecting variables. This unusual practice will almost inevitably result in creating large, unstable models whose predictive performance in new data is likely to be questionable. The authors acknowledge the point. The authors attempt to gauge the loss of performance with such models by looking at the bootstrap calibration slope (an internal validation technique). Indeed, with large and complex selected models, the slope turns out to be much less than the nominal target value of 1. Two points. First, I would like to see some results for a more conventional significance level such as 0.05 or perhaps the ‘Akaike level’ of 0.157 (or at least comments on such results). Second, the reported c-index values are certainly inflated due to optimism, and this could have been corrected by using Harrell’s bootstrap approach to estimating the optimism. (The reported calibration slopes were done in this spirit.) Without a correction for optimism, I don’t see how the c-index values can usefully be interpreted. This is important since as things stand, I don’t think it is possible to conclude anything meaningful about how well or badly the model selection procedures performed, particularly with respect to the choice of variable inclusion frequency cutoff. Therefore, I don’t think the last sentence of the Results section is justified.

6. The authors don’t really give a clear idea of what bootstrap inclusion fraction threshold they consider desirable, either in general or in the specific example they study. Note that for independent predictors, a threshold of 50% applies to variables that are just significant at the chosen level. Of course, with correlations among variables, this is only approximate.

7. The sample size (number of events) is not stated, either overall or for the individual studies.
Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

1. Figure 1 is of poor quality (low-resolution plot).
2. I did not see any mention of adjusting for difference in prevalence between studies.

-----------------------------------------------------------------------------------

Discretionary Revisions (which the author can choose to ignore)

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes

Declaration of competing interests:

I have no competing interests.