Reviewer's report

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

Version: 1 Date: 14 November 2006

Reviewer: Xiaowei YANG

Reviewer's report:

General

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

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

1. Multiple Imputation and Missing Data
   More literature review or discussion should be added regarding the missingness mechanism and MI. For example, ignorability is an important concept, thus should be discussed. There are many model-based algorithms that can be used for creating multiple imputations, and most of them require the assumption of ignorability.

2. Imputation with MICE
   When applying the MICE algorithm, it would be better that the full conditional distributions be used, where all the other predictors be enclosed to make predictions on the current prognostic indicator. If there is collinearity problem, it should be managed beforehand. The procedure as you described may face convergence problems because the algorithm jumps between spaces with changing-dimensions.

3. Variable Selection with Bootstrap
   In the paper of Austin and Tu (2004), which you referred, the bootstrap method with stepwise selection schemes was first introduced. Even in that paper, no enough theoretical justification or simulation-based evaluation was presented. I am very interested to compare the performance of this algorithm with other sophisticated methods, e.g., criterion based methods (e.g., AIC, BIC) or MCMC-based Bayesian selection methods (e.g., SSVS, GVS). If you could provide a simulation study for this purpose, that would be very helpful.

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

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. In your merged data set, I noticed that there are also missing values on the outcome variable, how did you handle them?

2. There are many places with improper use of languages. I suggest find a good proof reader.

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

Discretionary Revisions (which the author can choose to ignore)

1. A limitation with MICE is that large number of parameters are encountered when the number of prognostic indicator is as large as 30. An alternative is Schafer's (1997) Data Augmentation algorithm based on the general location model with log-linear constraints.

2. Your justification of MAR (missing at random) is not adequate, because there are many missing values that were not due to the design of the study (e.g., Education, Smoking). You aimed to study the influence of imputation variation and sampling variation on variable selection using only this practical data. Unfortunately, you did not know the "real" missingness mechanism, the "real" prognostic predictors, and the population variance. Therefore, many of your conclusions based on this data analysis cannot be simply generalized. Once again, simulated data sets are necessary for this goal.
What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

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

Statistical review: Yes

Declaration of competing interests:
I declare that I have no competing interests.