Reviewer's report

Title: Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia

Version: 1  Date: 25 February 2015

Reviewer: Zhongyang Zhang

Reviewer's report:

The authors based their assessment of different model selections methods on a concrete dataset and provided valuable guidance on how to choose statistically and clinically sound model selection methods in epidemiology research. This manuscript is well organized and written, while it may be further improved if the authors can address a few minor issues.

Minor Essential Revisions:
1) What is the distribution of the dependent variable HRQoL? Does it follow normal distribution well or is a proper transformation needed for it to be applicable in the context of OLS?

2) It seems that the whole dataset was used to tune the regularization parameter lambda in penalized regression by cross-validation and also used to assess the stability of selected subsets of variables by bootstrap. While this strategy is suitable for the purpose of identifying important predictors in exploratory or hypothesis generating analysis, which is a usual aim in epidemiology, it may still raise up concerns in that the lambda tuning step with cross-validation has already implicitly accounted for model uncertainty. Therefore, it may over estimate the model stability to bootstrap the whole dataset once more. The results may be changed (perhaps not very substantially for this specific dataset) if the whole dataset is divided into "training" and "testing" sets: one is used for tuning lambda and the other for assessing model stability. Similarly, the p-value cut-off in stepwise regression with LRT can be tuned in "training" set and model stability be assessed in "testing" set. This treatment may potentially lead to a slightly more rigorous comparison of different methods. It should be noticed that the sample size of "testing" set needs to be the same in assessing model stability of these methods.

3) 95% CIs in step regression can be also estimated by bootstrap and listed side by side with those estimated by asymptotic distribution for comparison, since they were estimated by bootstrap in penalized regression methods.

Discretionary Revisions:
1) How many bootstrap iterations were used in each analysis?

2) It would be helpful to improve the readability if the authors can visualize the independent variables by grouping relevant ones together in tables (Table 1,
Additional Files 2-5) and figures (Figures 1-4) as what they did in the second paragraph of Subsection Dataset, where these variables were grouped into different sections in the structured questionnaire.

3) The correlation matrix in Additional File 5 may be color-coded in the form of heatmap to make it more intuitive.

4) Typo in the second paragraph of subsection Dependent variable and independent correlates (Line 158): "cut-of" should be "cut-off".

5) In the captions of Figures 1-4, (A), (B), etc. are "subfigures" rather than "figures", or just simply remove "Figure" before (A), (B), etc.

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.