Reviewer’s report

Title: Comparison of logistic regression and classification and regression tree model: An application in identifying high-risk populations in alternative tobacco product use

Version: 2
Date: 4 October 2014

Reviewer: Mohammad Reza Reza Baneshi

Reviewer’s report:

Major Compulsory Revisions

1) One important issue that should be clarified from the beginning is that whether the authors aim to address a statistical question (i.e. comparison of two modeling techniques), or a health related question (i.e. who is more likely to use cigarette + ATP).

If the aim is the first one, then I would like the authors to take into account the sampling variation. Only one training and one validation set is selected, and Final results suggest higher c-index for regression model. However, to provide a more realistic picture, several samples should be selected. In addition, current results are restricted to a data set with a fix sample size, number of independent variables, … If the aim is to provide statistical messages, then issues that affect performance of these two techniques should be changed to compare their performance under different circumstances.

On the other hand, if the aim is simply to provide information for health professionals, then some sections of the manuscript need revision. For example, discussion section is written as it aims to give statistical messages.

Minor Essential Revisions

1) As I pointed above, the motivation of the study is not clear. In the background section, more details about advantages and disadvantages of tree models are required. Furthermore, the specific aim of the study should be clarified.

2) Authors noted missing rate was low and imputation was not necessary. It is recommended to report missing rate.

3) What do authors mean by ‘neighborhood models’ and ‘incremental prognostic value’? In regression modeling, it seems to me that authors followed AIC approach, and then added variables that had added value. But from which prospect (i.e. R-square, C-index?)

4) In addition, authors noted that logistic regression is a parametric approach which requires distributional assumptions. Then they noted that, to release
distributional assumptions, they wish to apply tree methods. However, logistic regression can deal with continuous, binary, and categorical independent variables. So this method is not restricted to distributional assumptions. Only the dependent variable should be binary.

5) I think there are much better justifications to use tree methods. These methods reveal complex interactions; provide pictorial easily interpretable evidence; and can be considered as complementary to regression models. I suggest the concentrate on these issues.

6) In the results section, figure 1 is not necessary. Furthermore, I was not able to print figure 4.

7) How discrimination score was measured?

8) I prefer to see sensitivity and specificity of models. C-index is less sensitive to the number of variables.

9) Please report the same statistics for all models. For example, in the abstract c-index for logistic and misclassification rate for tree models are reported. This makes the comparison of models difficult.

10) If I understand well, in logistic regression the probability of event has been calculated for all subjects. These probabilities are then used to estimate C-index. On the other hand, tree model does not estimate any probability. It simply allocates the subjects into two groups. So not sure how C-index was estimated for tree model.

I prefer to divide the subjects into two groups based on logistic regression as well, and to see the cross-table of logistic and tree models. This provides information on number of subjects classified into different groups by means of two models.

10) The final conclusion does not match the findings. Authors noted that tree was classified the participants more efficiently. However, its c-index was lower than the regression model.

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