Reviewer's report

Title: Stratification of Coronary Artery Disease Patients for Revascularization Procedure Based on Estimating Adverse Effects

Version: 2

Date: 5 November 2014

Reviewer: Anna L Buczak

Reviewer's report:

Reviewer’s report:
Overall this is an interesting paper that addresses an important problem.

Major Compulsory Revisions

1. All the abbreviations should be described before they are used:
   a. For example: p.2, line 38: BMS is used but not defined.
   b. Many such instances can be found in the text

2. P.4., l30: “TVR could save up to $205 per year”. Is it for a specific country or a specific hospital?

3. P.5, l. 44-45: “The complete data set comprised 2377 patients who underwent 2733 interventions”. Figure 2 talks about a total of 3958 interventions. Make sure the figures and the text agree and describe the same data set.

4. P.5., l.45 – p.6 l.1: “of which 913 were lost to follow-up”. What does it mean that interventions were lost to follow-up ? Follow-up was not preformed ? Be clear so it is easier to read.

5. P.24-25: Table 1: features in the table are C1-C3 but in the text underneath they are referred to as D1-D3.

6. P.7, l. 6-7: “Prev(Restenosis | BMS)… among BMS receivers in the overall population”. What is the overall population you are referring to ? You can refer to Fig 2 or Fig 3, if needed. You have to be careful when you do that, as later you say that you are using only part of the data for training.

7. P.7. l. 14-16: “If the rate of hazardous events is similar For BMS and DES…” – this looks like an assumption. Any reference describing this is indeed the case?

8. P.7. l.16-17: “… if Des and CABG treatment mostly differ in the rate of hazardous events …”– this looks like an assumption. Any reference describing this is indeed the case?

9. P.7. l. 27 – you are talking here about using decision trees. Decision trees are usually built using an algorithm like C4.5 on the training data. However on p.7. l.42-43 you say “We constructed all possible trees …” It is unclear why would you construct all possible trees when C4.5 (and similar) algorithms do that already without exhaustively constructing all the trees. This is the most important problem I see with the method employed in the paper.
10. Figure 4: You show there Composite 1, 2 and n. How many Composites do you have? 27?

11. P.8. l27-29: “... classifiers for restenosis... and hazardous events were trained...” On which data were they trained? On the training set?

12. P.8. l. 32: “We obtained classifiers ... for each patient group”. How many patient groups are we talking about?

13. p.9, l. 25-27. “to determine the best partitioning into patient subgroups ... selects the best performing decomposition of the entire data set... Again, on which data was this performed? Entire data set? training data?

14. Figure 5: What is the percentage on y-axis referring to? If this is the percentage of models that achieve a given AUC? If yes, all the dark bars should add to 100% and all the white bars should add to 100%. And they definitely do not.

15. P.12, l.18-19: 10-fold cross validation was performed and somewhere else in the paper is says that 300 different cross-validations were performed. It also says that overall 133-185 classifiers were constructed. So was it 133-185 classifiers per one fold of cross-validation? Meaning 300*10*(133 to 185) classifiers? Make sure you describe somewhere how many classifiers have been built.

16. P.12 l.45 –p.13, l1-2: “...probability of restenosis was 6.45(6.2%) ...” What do the numbers in brackets refer to?

17. P.14, l.43-44. The costs of $590 and $28,683 are these costs for a particular insurance company? The cost are in $, so I am assuming these are for US but it is well known that the same procedure can have different costs in different hospitals, and even the same hospital bills a different amount for the same procedure different insurance companies based on the agreement they have.

18. Tables 1-6 – Table captions should be above the corresponding tables, not below.

Minor Essential Revisions
1. P.4, l. 21 change “BMS have” to “BMS has”
2. P.4, l. 32: “They used unbalanced data”. Who is “they”, Amin et al. or Cavender & Ellis ?
3. P.5, l. 1: change “there model” to “their model”
4. P.8. l. 17: “= M= 4” change into “N= M= 4”
5. P.11, l.17: change “[0; 1]” to “[0, 1]”

Level of interest:An article of importance in its field

Quality of written English:Acceptable

Statistical review:No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:

I declare that I have no competing interests