Reviewer's report

Title: Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis

Version: 1 Date: 28 March 2014

Reviewer: Wilfried Grossmann

Reviewer's report:

The paper considers an interesting approach for combining data from EHR with community level factors for finding risk factors for obesity at the community level. As statistical method logistic regression is used.

As statistician I cannot evaluate the medical implications of the results but in my opinion the approach making use of different data sources in retrospective observational studies is important and of interest in many areas of evidence based medicine and by no means limited to the considered question of obesity.

Overall the paper is well written and describes data preparation and the method for model development quite clear. Limitations of the approach are clearly stated.

However, I would propose to go one step further in the analysis and put it into the framework of data mining. This would mean that the authors use for model assessment the method of splitting the data into a training set, a validation set and a test set.

(See for example the textbook Hastie et al.: The Elements of Statistical Learning, chapter 7, Springer).

The frequently recommended sizes are 50% for training, 25% for validation and 25% for testing. Besides the training and the validation step, which was combined in the paper by using the method of variable selection for the logistic model you can apply the model to an independent test set. Application of the model for independent test data would give a better idea about the generalization power of the model.

Due to the large amount of data available this approach should work quite well in this case.

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 have no competing interests