Reviewer's report

Title: A Richly Interactive Data Visualization Tool for Cohort Studies Using Electronic Medical Records

Version: 3 Date: 9 June 2015

Reviewer: Linda Moniz

Reviewer's report:

This paper describes a highly user-driven data mining process that should result in useful visualizations of cohorts from electronic medical records.

Major Compulsory Revisions:

1. The authors need to keep the tool and rewrite the paper to make a more comprehensible explanation of the research. Revise the notation and the explanation of the method and include visualization that indicates the true power of the tool. Label the graph axes, please. More detail is provided in the following comments.

The paper is rife with malapropisms, incomplete and non-parallel constructions, poor grammar, and sentences that are missing key parts. The paper itself is difficult to read through and comprehend.

The paper describes a visualization tool, but visualizations in the paper are poorly labelled and cluttered. They reportedly describe key features of the data that have been discovered by the visualization tool, but the y-axis is in all cases unlabeled; it is impossible to interpret the meaning of the plots. The figure captions are minimal.

Explanations of the tool are contradictory. The data mining aspects of the tool seem to add little power over what would be available in, for example, SAS. It is not possible to tell easily from the paper that the work is innovative or merely a user-driven rearrangement of data.

That said, there is possibly some hidden power of the tool that is rather poorly explained.

The tool requires the user to provide the “factors.” Each factor is a single criterion (in the text, criteria is used as singular; please repair) that can distinguish patients into cohorts. In the text, the single factor cohort is not clear; the paper seems to indicate in some places that multiple factors are necessary to distinguish cohorts. The explanation of the factor/cohort organization is unclear; in some places the paper indicates factors are single criteria, in others it indicates the user needs to aggregate them with considerable domain knowledge in order to find cohorts.

The cohorts are then ordered somehow into time stages, but again the paper is
unclear and the notation provided does not clarify the selection. These time stages are supposed to represent stable comorbidities but again the notation is a hindrance and does not clarify the selection.

After this rather confusing organization of patients into discrete time-staged cohorts, the paper describes a potentially very powerful construction. The description of the process to get to the construction needs to be overhauled so it is at least cogent, but the construction of the cohort trajectory network G could be very useful in uncovering hidden relationships in the data. The clustering, similarity, and entropy measures are not particularly innovative, but the network itself, if it could be described in an understandable way, is worthy of developing as the central idea of the paper.

The tool as described is not really a data mining tool; the user must provide the data mining. The visualizations provided yield very little insight, but the development of other visualizations that demonstrate the properties of the cohort trajectory network are potentially very useful.

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

**Quality of written English:** Not suitable for publication unless extensively edited

**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 interest.