Author's response to reviews

Title: A Richly Interactive Exploratory Data Analysis and Visualization Tool Using Electronic Medical Records

Authors:

Chih-Wei Huang (weigo7729@gmail.com)
Richard Lu (richardlu@tmu.edu.tw)
Usman Iqbal (usman.iqbal85@gmail.com)
Shen-Hsien Lin (local.king525@gmail.com)
Phung-Anh (Alex) Nguyen (alexnthhp@gmail.com)
Hsuan-Chia Yang (itpharmacist@gmail.com)
Chun-Fu Wang (cfwang@ucdavis.edu)
Jianping Li (kelli@ucdavis.edu)
Kwan-Liu Ma (ma@cs.ucdavis.edu)
Yu-Chuan (Jack) Li (jack@tmu.edu.tw)
Wen-Shan Jian (jjtmutw@gmail.com)

Version: 5
Date: 2 August 2015

Author's response to reviews: see over
Author's response to reviews

Title: A Richly Interactive Exploratory Data Analysis and Visualization Tool Using Electronic Medical Records

Authors:
    Chih-Wei Huang (weigo7729@gmail.com)
    Richard Lu (richardlu@tmu.edu.tw)
    Usman Iqbal (usman.iqbal85@gmail.com)
    Shen-Hsien Lin (local.king525@gmail.com)
    Phung Anh (Alex) Nguyen (alexnthhp@gmail.com)
    Hsuan-Chia Yang (itpharmacist@gmail.com)
    Chun-Fu Wang (cfwang@ucdavis.edu)
    Jianping Li (kelli@ucdavis.edu)
    Kwan-Liu Ma (ma@cs.ucdavis.edu)
    Yu-Chuan (Jack) Li (jaak88@gmail.com)
    Wen-Shan Jian (jj@tmu.edu.tw)

Version: 2 Date: 2 Aug 2015

Author's response to reviews: see over
Reviewer #1 (Dr. Hong Kang)

Comments to the Author:

In this article, using CKD as a model disease, the authors proposed a visual mining system which support exploratory data analysis of EMR data. The knowledge of relevant comorbidities that CKD patients develop over time are revealed by Sankey diagram, which makes sense of the EMR data. Overall, the authors have done extensive and solid investigation and well answered the question that how to make sense of the data. There are a few comments.

Major Compulsory revisions:

It is interesting that the system allow the user to define a set of factors by selecting independent codes or aggregating correlated ones based on their knowledge. The advantage is if the user wants to study the clinical trajectories of other disease, they can define a list of relevant factors and apply the same process to set up time stages, cluster patients, as well as explore the structure of the cohort trajectories. However, my concern is that, there would be a very high requirement for user’s experience of handling such complicated model independently, since most of them maybe healthcare providers without enough machine learning skills.

1. Thanks for your comments. So far, this may not automatically transform any dataset to build the interactive diagram. However, we would like to provide our data structure to users so that they can prepare and transform the dataset accordingly. Besides, we addressed the data analysis process and interactive visualization approach by using Chronic Kidney Disease patients’ cohort data in our study. This diagram is user friendly as users can easily interact with our web-based application by adjusting the group sizes, hovering the mouse cursor on the nodes or transitions, and clicking the nodes or transitions. This tool can answer users’ research questions by showing the information on the diagram. The CKD visualization software is built for non-technical healthcare provider and researchers. No machine learning skills are required. The user only needs to supply basic domain knowledge. The visualization software produces a Sankey diagram corresponding to the selection criteria that the user chooses. A Sankey diagram, because it is visual, makes exploring complex EMR data easier by clearly showing trends over time between patient cohorts.
I also concern about the stability of the system when using on other EMR data besides CKD, I suggest the authors to apply the system on another disease in order to support your point mentioned in the conclusion part that the analysis process is generalizable to any other disease that a user wishes to follow over time and can work with different clustering and filtering algorithms.

1. Thanks for your suggestion. It could be reported in further our study as CKD was a pilot study.

2. We used CKD as an example for this pilot study. For the next step, we will apply our system on other diseases to validate its usability and functionality.

Minor Essential Revisions:

1. The figures present by Sankey diagram are very important for readers to understand the value of the system. Please make more detailed legend for these figures.

   1. Thanks for your suggestion. We have revised and add more detailed information in the legends for the figures on line 612-642 page 24-25.

2. As the system is a web-based application, I suggest the authors to provide its link in the paper.

   1. Thanks for your suggestion
   2. We have now added the link for readers to access the website of Spectrum Sankey State Timeline (http://sankey.phr.tmu.edu.tw/) in the manuscript on line 393 page 17.

3. There’re several grammatical mistakes in the article, for instance, line 178, “Hence, our system also the user to rank and filter the associations based on their statistical importance.” Moreover, please reduce the exactly same sentences occurred in the parts of abstract, introduction and conclusion.

   1. We have revised our manuscript with those grammatical mistakes in the revised article.
Reviewer #2 (Dr. Zhen Zhang)

Comments to the Author:

**Major Compulsory Revisions**

1. *Please make the research question clear-- developing a visualization tool for exploratory analysis with the demonstration of its application? or evaluating the usefulness of this tool in clinical research?*

   We developed a visualization tool for exploratory analysis and described how the system can be used to begin answering clinical research questions. Chronic kidney disease is the prototype disease that we use to explain how the software works. We have revised our title and the manuscript and pointed out the research question more clearly in the “Background” on line 91-100 page 4.

2. *The clearly stated research question will help determine the title as well. The current title "A Richly Interactive Data Visualization Tool for Cohort Studies Using Electronic Medical Records" makes it sound like the purpose of the tool is specifically for cohort studies, however, after reading the article, it seems that the tool is meant to be used to facilitate "exploratory analysis where the user may not have a well-defined hypothesis".*

   We have modified the title as the reviewer suggested to “A Richly Interactive Exploratory Data Analysis and Visualization Tool Using Electronic Medical Records.” This tool will help users to explore chronic diseases’ evolution over time with or without a well-defined hypothesis.

3. *The article needs to be better organized by headings such as the motivation to develop the tool, the technical details of the tool, the end user interface of the tool, the validation of the tool, the application of the tool, and the evaluation of the tool.*

   Thanks for reviewer’s comments. We have now organized our revised version.

   For “the motivation to develop the tool,” we mentioned in *Background* (Line 91-100 Page 4).

   For “the technical details of the tool,” we mentioned in
Methods-System Design section.

1. For “the end user interface of the tool,” we mentioned in Results-Explore Cohort Structures section.

1. For “the validation of the tool,” we mentioned in Results-Exploring Associated Relationships section.

1. For “the application of the tool,” we mentioned in Results-Responsiveness section.

1. For “the evaluation of the tool,” we mentioned in Methods-Pilot Study section.

Minor Essential Revisions

1. Please define terms used in the article: homogeneous subset, uniform subgroups, visual mining, visual rendering, human-curated visual evaluation, visual aggregation, time stage.

We have defined those terms in the revised version of paper, please also find the index of each definition below:

a. Inhomogeneous subset: A dataset whose items are not easily categorized together. Due to those high-dimensional data, such as electronic medical records, would lower the homogeneity between data items, we used a divide-and-conquer process as a solution to uniform the data.(Line260-264 Page11;Line 457-460 Page 19)

b. Uniform subgroups: A subset of a larger set of patients who share a common characteristic. For decreasing the confusion of those similar terms, such as “subsets” and “subgroups”, we have removed the “subgroup” term in our manuscript.

c. Visual mining: A term for any patterns or knowledge gained from visually inspecting large datasets. It’s a technique for visualizing the multidimensional data. In this study, we propose the visual mining process for adopting electronic medical records data visualization.(Line 462-463 Page 21)

d. Visual rendering: Rendering is the process of generating an image from a software program. In the future work, we will also work on better visual rendering for optimizing our system.

e. Human-curated visual evaluation: In our study, the visual evaluation is based on human-curation. Once the user specifies the time windows based on their domain knowledge, our system
partitions the patient records accordingly. This is a human-assisted task because homogeneity, especially on the semantic level, is often judged best by human. In this study, we used CKD as a model disease, and a CKD cohort was assembled by automated correlational analysis and human-curated visual evaluation. (Line 283-290 Page12-13)

f. Visual aggregation: We have removed this term to reduce unclear sentence.

g. Time stage: We have separated “stage” and “windows” to refer two different objects. For proper usage, we used the “stage” to refer the patients’ CKD condition, and the “window” to refer the time duration or interval in the article. (Line 283-290 Page12-13)

2. Please explain more about how to understand the information displayed on those figures.

1. We have revised and add more detailed information in the legends for the figures on line 612-642 page 24-25.
Reviewer #3 (Dr. Linda Moniz)

Comments to the Author:

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.

   1. We have revised and add more detailed information in the legends for the figures on line 612-642 page 24-25. We have also labeled properly in figure 3.

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.

   1. We have revised the whole manuscript, and copy edited by native English professional language editor. The revision part were highlighted to stress our work.

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.

   1. We have revised and add more detailed information in the legends for the figures on line 612-642 page 24-25. We have also labeled properly in figure 3.

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.

1 We have now revised our manuscript and explained clearly that this tool has great power to understand disease trends in a user-friendly way. We’ve explained these user operations in Pilot Study section. (Page 10-14)

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.

1 We have revised our manuscript and now explained more clearly about the selection criteria. Multiple factors are necessary to distinguish cohorts. We keep this option open for users to decide which factors they would like to take into account while grouping and to find the research question. So, visualizations can be done based on a single criterion or multiple criteria, depending on researchers’ interests. We have described more detailed information on Line 268-280 Page 11-12)

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.

1 We have revised our manuscript to make it more clearly about selection criteria. For example: A time stage is a period of time that a chronic disease is present. For example, a time stage in which a CKD patient is diagnosed with diabetes 3 years before and then gets hypertension after 2 years. These subjects will represent different group as compared to those who got only diabetes and then CKD but did not hypertension throughout cohort.
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.

1. We have explained clearly and revised the methodology section in our manuscript.

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.

1. Thanks for reviewer comments and we agree with reviewer that this tool is only visualization tool. However, we have added more information about data pre-processing before the visualization process in our revised version. This visualization tool could help users to understand diseases’ evolution and their associations over time. In the future work, we are also planning a more flexible and interactive system for fitting different users research needs.