Data-driven hypothesis generation for coexisting medical conditions

Mainak Chowdhury and Frank DeVilbiss

September 1, 2016

Introduction

There is a wealth of information available publicly about different medical conditions, that has been especially enabled with improved data collection and recording practices [1]. Sometimes the coexisting medical conditions may be unrelated. In a significant portion of other cases, however, the conditions may **not** be mere coincidences. They may, in fact, be side-effects or otherwise harmful interactions resulting from the standard treatments for one or more of the conditions. Rigorous identification of the latter, however critical, is very time consuming and expensive, requiring elaborate drug sensitivity testing or clinical trials.

There is thus a pressing and ever-present need to enable a more accurate identification of these harmful interactions through transparent methodologies. New research suggests potential in data mining/information analysis methods [2, 3]. Such methods are not complete solutions alone [4] and cannot replace sound medical experiments and proper statistical analysis. However, as evidenced by the works above, they can be powerful drivers of hypotheses and focused investigations, especially when combined with sound medical knowledge. Our project aims to streamline information analysis based on publicly available information about coexisting conditions to make subsequent hypothesis generation and validation easier.

In the first year of our project, we have started with data from the publicly available **Multiple Causes of Death (MCD)** dataset [5] and developed tools to enable effective and flexible visualizations of the data and support data-driven hypothesis generation and validation.

Progress made

Our work has mainly focused on the following components:

- web tools that visualize networks of related medical conditions,
- web tools that enable user assisted exploration and discovery of the effects of demographic features and diseases.

Past and ongoing challenges in implementing the tools above have been mainly algorithmic. For the scale and variety of queries that we would like to support we continue to research and test various data structures and algorithms for representing and querying the multiple causes of death dataset.

Output

We have prepared a web interface that anyone can use at `http://www.datatherapeutics.org`. A CSoI brown bag talk describing our methods and motivation was also presented; with a recording available at `https://www.youtube.com/watch?v=q_7YEiN-Zqc`. Once we have released a stable version of our platform, we aim to publicize our work through a software methods paper in a bioinformatics venue highlighting the algorithms and data structures that enable us to handle the scale of queries.

Month	Number of meetings
October 2015	1
November 2015	4
December 2015	12
January 2016	3
February 2016	3
March 2016	1
April 2016	1
August 2016	1

Table 1: Interactions grouped by month

Projected expenses till the end of the year

In prior discussions with the center, we have confirmed the Center's support of our use of funds towards the purchase of cloud server space that hosts the outcomes of this work. We have been hesitant to finalize our purchase of these computational resources due to the evolving nature of our data analysis tools. After spending more time developing these tools, we are ready to commit to a particular server configuration.

We have yet to spend any of our seed grant funds on this project. Our initial grant was for the amount of $5000. We would like to secure these funds for the next year to further our project.

- $1000-$1500 - to purchase cloud server space,
- $3500-$4000 - to fund travel for further meetings and conference presentations.

Details of meetings (mainly online)

Throughout the course of the past year, we can find record of at least 25 meetings to discuss the project and make progress toward our goals stated in the introduction section. These meetings have mostly taken place through electronic means and are itemized by month in Table 1. Three days worth of in-person interaction took place during the annual site visit.

Plans for the next year

While this team will have graduated by the end of the next academic year, we do not want this effort to end. We are strongly committed to continuing and maintaining this project even after graduation. Towards this end, we would like to open up a discussion with the Center on how best to proceed.

References

[1] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[2] N. P. Tatonetti, P. Y. Patrick, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science translational medicine*, vol. 4, no. 125, pp. 125ra31–125ra31, 2012.

[3] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi, "Assessing Google flu trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic," *PloS one*, vol. 6, no. 8, p. e23610, 2011.

[4] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[5] M. D. Redelings, F. Sorvillo, and P. Simon, "A comparison of underlying cause and multiple causes of death: US vital statistics, 2000–2001," *Epidemiology*, vol. 17, no. 1, pp. 100–103, 2006.