**Towards Interpretable Precision HEOR**

Austin Bell (alb2307)

Healthy Life: Patient Health Development Analysis and Prediction for Precision Health

February 3, 2020

**Abstract:** Withthe continual improvement in Big Data technologies, precision healthcare is beginning to generate significant improvements within our healthcare system. A logical next step is to leverage techniques and methodologies developed for precision healthcare and utilize them towards improving health economics and outcomes research (“HEOR”). This analysis looks to take state of the art topological / graph analysis to combine heterogeneous clinical patient data, such that I can more effectively identify granular patient subpopulations. Survival models are then developed for each of the patient subgroups. Through this methodology, I seek to maintain high-levels of interpretability, which can be utilized to improve patient-physician shared decision making. Additionally, I hope that through combining non-traditional and heterogenous data sources, I will be able to improve the effectiveness of simple, easy to interpret survival models.

1. **Background (Review of Related Literature):**

With recent advances in big data technologies, precision healthcare and personalized medicine looks set to significantly improve our healthcare system. This will be accomplished through identifying granular disease sub-types and developing treatments that might benefit particular subgroups of the patient population. In the end, depending on a patients’ genetic makeup they may receive a different treatment. This project looks to expand in a different, but equally important, direction towards precision health economic and outcomes research (“Precision HEOR”). Precision HEOR looks to measure potential outcomes and provide an economic analysis of varying treatments for granular patient populations.

The benefits for precision HEOR are significant, including improvements in intervention development, clinical guidelines, targeted reimbursement, and shared decision making.[[1]](#endnote-1) Within intervention development, precision HEOR can utilize real-world evidence to demonstrate the value of the product, identify new populations that can benefit from existing treatments, and generate an overall more nuanced value story. It can improve clinical guidelines by helping to account for patient variability post clinical trial. Payers could implement targeted reimbursement strategies by understanding how one treatment might affect certain patients differently and the corresponding economic ramifications. Finally, improved shared decision making can occur through better understanding the patient safety and health outcomes that an individual patient might face.

This project is focused on identifying patient subgroups leveraging patient similarities across multiple clinical variables and heterogenous data sources. A significant number of studies leveraging patient similarities have been performed. While most studies utilized cross-sectional data, there were a couple that also used longitudinal or cross-sectional and longitudinal data.

Typically cross-sectional is considered the easiest data type to handle. Bobowski et al. (2006)[[2]](#endnote-2) utilizes a series of linear transformations to better separate patient groups and guide medical decisions. Park et al (2006)[[3]](#endnote-3) developed a k-nn algorithm that was dynamic to the target patient and demonstrated improvements across a variety of health datasets. Lee et al (2015)[[4]](#endnote-4) utilized the cosine similarity metric on a summary of the patients’ ICU stays before predicting the 30 day mortality rate for each subgroup and found that identifying patient subgroups improved accuracy at the expense of increased computational burden.

Longitudinal and longitudinal plus cross-sectional data were also explored. Saeed (2006)[[5]](#endnote-5) developed a novel time series similarity metric for the detection of impending hemodynamic deterioration. Sun et al (2010, 2010)[[6]](#endnote-6),[[7]](#endnote-7) utilized a localized supervised algorithm for identifying patient similarity to predict event outcome within the next hour. Finally, Gottlieb (2010)[[8]](#endnote-8) leverages multiple patient similarity metrics for cross-sectional plus longitudinal data to better predict ICD9 disease diagnosis.

Additionally, this project seeks to combine heterogenous data sources to achieve this goal. While most of the literature looks at how to combine omics with non-omics data[[9]](#endnote-9), I will primarily focus on clinical data generated during a patient’s hospital stay.

One such methodology that combines both omics and non-omics is referred to as patient similarity networks, which refers to a graph network where each node is a patient and each edge represents an estimate of similarity of a variable between patients. Topological data analysis is then applied to the network to identify patient subgroups. Wang et al. (2013) [[10]](#endnote-10) first developed similarity networks by generating similarity matrices across patients using heterogenous data types. Their team was able to incorporate mRNA expression, DNA methylation, and microRNA expression data together to identify additional cancer subgroups, which they found helpful in better predicting survival. Li et al (2015)[[11]](#endnote-11) take this further and include patient EHR data plus genomics data into their similarity network. Further topological data analysis enabled them to identify previously unknown subgroups in Type 2 Diabetes Patients. Finally, additional classification algorithms have been proposed to run on top of this patient network[[12]](#endnote-12).

My plan is to leverage what has been researched for patient similarity networks in a omics plus non-omics setting and focus on just the clinical setting. Rather than identifying unknown disease types, I will focusing on identifying granular patient populations. I will expand the current literature in multiple ways. First, I will look into new a disease area that has not been explored leveraging patient similarity networks. Next, I will seek to incorporate temporal, rather than just cross-sectional clinical data. It should also be noted that my framework can very easily include additional heterogeneous data sources including omics data and PRO data. Finally, I will evaluate whether there is improvement for traditional health outcomes analyses leveraging the results of the patient network.

1. **Introduction to the Project:**

The overall goal of the project is to leverage precision medicine techniques and big data technologies to better understand the future health outcomes surrounding individual patients. Applying HEOR analyses towards highly fine-grained patient populations will lead to significant improvements within our healthcare system. Once these granular patient populations have been identified, I will develop separate survival models for each population. Given the domain, I plan to focus on ensuring interpretability of my results and models throughout the entire process.

The end result will be a series of simpler predictive models which, at the same time, will hopefully improve upon or equal more complicated, “black-box” algorithms. I expect interpretability to increase through two modes: 1) by leveraging similarity matrices and identifying subgroups, we can evaluate each subgroups’ intra-group similarities and inter-group differences and 2) by identifying granular similar patient subgroups, for which I can run individual survival models on, I am reducing the amount of noise per sample set while also controlling for previous unobserved confounders[[13]](#footnote-1). I expect that this second point will also lead to improved prediction accuracy over previous approaches.

I expect this research and general foray into precision HEOR to improve our understanding of how different patient subgroups should be treated differently, in addition to improving patient-physician shared decision making. At no point would the algorithm make a decision for a physician, the analysis is developed to guide a physician towards relevant subpopulations that they can then explore and determine the best course of action based on their individual patient’s characteristics. This “guiding” process is made to be extremely interpretable (e.g., the patient population exists because the patients here exhibit similarities in X, Y, and Z variables). Additionally, future tools can be built upon this to enable physicians to toggle clinically relevant variables they would like to include while identifying patient subpopulations.

In regards to implementation, my first step is to generate patient similarity networks that fuses heterogenous data from clinical EHRs (e.g., longitudinal, cross-sectional, data with constraints, etc.). For each clinical variable(s), I will develop a similarity matrix between all patients using relevant similarity metrics[[14]](#footnote-2). These similarity matrices are combined to form a single graph representing the entire patient-patient network. I will explore two methods for generating the network:

1. A nearest neighbors approach where an edge between two patients for each variable(s) is only generated if they are within the top k most similar values
2. Leveraging dimensionality reduction techniques, I will reduce the number of patient-patient similarity outputs, but maintain all patient-patient edges (e.g., via SVD, PCA, or L-Infinity Centrality)

Once a single network is developed, I will utilize topological data analysis to identify relevant subgroups across patients.

However, this project takes this one step further towards estimating health outcomes within these subgroups. For each clinical subgroup, I will run survival analyses utilizing the Cox Hazard Regression methodology to determine whether different patient subgroups should expect different outcomes. Subject to time availability, I will explore the possibility of running a network meta-analysis utilizing real world data to understand how varying treatments will affect patient subgroups differently.

One of the primary reasons that I chose this methodology is that the patient network is made up of individual similarity matrices that are highly interpretable. Therefore, the final stage of the project will focus on interpretation of the results through topological and graph visualizations. These visualizations will be made interactive through the R Shiny platform to enable the users to easily explore the various patient subgroups and corresponding results.

1. **Introduction to the Dataset:**

This project will utilize the MIMIC-III Dataset[[15]](#endnote-13), which is a free database made up of de-identified health information for over forty-thousand patients who stayed in critical care units at Beth Israel Deaconess between 2001 and 2012. The database includes a variety of information such as demographics, vital signs, chart events, lab events, microbiology events, procedures, imaging, mortality data, and clinical notes. Most notably, this database is renowned for its very granular temporal data. Consisting of 25+ separate tables, this database provides a highly detailed view of patient hospital admissions.

While the database has a large focus on patients who ended up in critical care units, I will largely be focusing on their general hospital stay rather than the patients’ ICU stays.

I foresee two major difficulties in utilizing this data for my analysis:

* This is a somewhat complex RDMBS, which I need to familiarize myself with quite quickly.
* Beth Israel changed their data collection platform in 2008, therefore, I will likely need to manually clean a significant portion of this data to account for these changes (for example, “heart rate” measurements have different item ID’s across the two systems, so these need to manually be combined).

I want to note that there are a couple things that I do not foresee to be a problem because of the approach that I am taking such as:

* Heterogenous data sources: by standardizing each clinical variable to a similarity measure, I can easily combine and compare different types of data
* Missing data: patient to patient similarities will only be considered if both patients have the data. For example, I will compare outcomes post a specific procedure if both patients have experienced the procedure.[[16]](#footnote-3)

1. **Plan:**

In this section, I outline what I plan to have accomplished by each of the milestones.

Milestone 1:

* Step-by-step research plan (subject to change as project progresses)
* Extract EMR data from MIMIC III and create GCP Big Query Database
* Identify patients with Liver Disease
* Identify clinically relevant variables within lab events, microbiology events, chart events, procedures, medical history, and demographics
  + Relevant variables are identified through three methods: 1) Research discussing well understood tests and risks associated with liver disease, 2) survival analyses and early predictors of liver analysis diagnosis / mortality, and 3) frequency of events associated with relevant patients
* Clean data and prepare for analysis
  + Data is stored in a variety of formats (longitudinal, cross-sectional, with constraints, etc.)
  + Multiple files will be saved as the basis for the analysis

Milestone 2:

* Generate Similarity Matrices
  + Includes various associative analyses across each of the data sources
* Generate patient-patient network
* Run topological data analysis to identify patient subgroups
* Compute Survival Analysis across each patient subgroup

Milestone 3:

* Identification of underlying differences between each patient subgroup
* Visualization of patient-patient network
* Dependent on time and data availability, explore the possibility of additional HEOR analyses (e.g., network meta-analysis)

Finalization:

* Interactivity to enable different views of the network based on different subsets of the data highlighting different variables
* Create report and final presentation

1. Yixi Chen, Gregory F. Guzauskas, Chengming Gu, Bruce C. M. Wang, Wesley E. Furnback, Guotong Xie, Peng Dong and Louis P. Garrison, “Precision Health Economics and Outcomes Research to Support Precision Medicine: Big Data Meets Patient Heterogeneity on the Road to Value”, 2016, Journal of Personalized Medicine. [↑](#endnote-ref-1)
2. Bobrowski L. Induction of similarity measures and medical diagnosis support rules through separable, linear data transformations [↑](#endnote-ref-2)
3. Park Y, Kim B, Chun S. New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis. Expert Systems. 2006 Feb;23(1):2–20. [↑](#endnote-ref-3)
4. Lee J, Maslove DM, Dubin JA. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. PLoS One. 2015 May [↑](#endnote-ref-4)
5. Saeed M, Mark R. A novel method for the efficient retrieval of similar multiparameter physiologic time series using wavelet-based symbolic representations. AMIA Annu Symp Proc. 2006:679–683. [↑](#endnote-ref-5)
6. Sun J, Sow D, Hu J, Ebadollahi S. Localized supervised metric learning on temporal physiological data. 20th International Conference on Pattern Recognition (ICPR); Aug 23-26, 2010; Istanbul, Turkey. 2010. pp. 23–26. [↑](#endnote-ref-6)
7. Sun J, Sow D, Hu J, Ebadollahi S. A system for mining temporal physiological data streams for advanced prognostic decision support. 2010 IEEE International Conference on Data Mining; Dec 13-17, 2010; Sydney, Australia. 2010. [↑](#endnote-ref-7)
8. Gottlieb A, Stein GY, Ruppin E, Altman RB, Sharan R. A method for inferring medical diagnoses from patient similarities. BMC Med. 2013 Sep;11:194. doi: 10.1186/1741-7015-11-194. [↑](#endnote-ref-8)
9. Evangelina López de Maturana, Lola Alonso, Pablo Alarcón, Isabel Adoración Martín-Antoniano, Silvia Pineda, Lucas Piorno, M. Luz Calle, and Núria Malats, “Challenges in the Integration of Omics and Non-Omics Data”, 2019, Genes. [↑](#endnote-ref-9)
10. Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains & Anna Goldenberg, “Similarity network fusion for aggregating data types on a genomic scale”, 2014, Nature. [↑](#endnote-ref-10)
11. Li Li, Wei-Yi Cheng, Benjamin S. Glicksberg, Omri Gottesman, Ronald Tamler, Rong Chen, Erwin P. Bottinger, and Joel T. Dudley, “Identification of type 2 diabetes subgroups through topological analysis of patient similarity”, 2015, Science Translational Medicine [↑](#endnote-ref-11)
12. S. Pai, et al. “netDx: interpretable patient classification using integrated patient similarity networks”, 2016, bioRXiv preprint [↑](#endnote-ref-12)
13. A significant amount of Judea Pearl’s work on Causality demonstrates the negative impacts of over-controlling for potential confounders, which he terms “collider bias”. For example, if I am measuring the causal effects of a treatment for liver disease mortality, then by controlling for other outcomes that are directly caused by the treatment (e.g., reduction in Bilirubin) then this may have negative impact on my ability to measure the causal effect of the treatment. If only looking at patients which had a reduction in Bilirubin then we may find that the treatment had no significant effect, but, in reality, the treatment may significantly increase the likelihood of reducing a patient’s Bilirubin level. In summary, I need to perform more research to ensure that this is taken into account if I make any causal statements. [↑](#footnote-ref-1)
14. For cross-sectional data, I will utilize cosine similarity. For longitudinal data, I will utilize either dynamic time warping or time warped edit distance. [↑](#footnote-ref-2)
15. <https://mimic.physionet.org/about/mimic/> [↑](#endnote-ref-13)
16. I will explore the possibility of adding a penalty for patients who are missing some data because missing data may actually reflect information (i.e., a procedure did not occur) [↑](#footnote-ref-3)