RAND Corporation, Reverse Engineering High-Quality Care:

Hsuan Chiu, PhD Student, University of California Los Angeles
Daniella Meeker, Information Scientist, RAND Corporation

Goal

The goal of this project is to discover **discriminative sequential treatment patterns** between patients with **better than expected** and **worse than expected outcomes** at given health status at incident diagnosis with a chronic condition. That is, we wish to reverse-engineer potentially novel measures of high-quality care pathways that may involve multiple treatments, procedures, or other health-care system interactions.

Methods

Identifying patients with better and worse than expected outcomes. We will identify patients with incident disease for common disease categories such as cancer, diabetes, congestive heart failure, and high blood pressure. At first step, we will use a logistic regression against a binary variable for "favorable outcome" and "unfavorable outcome" (where outcome might be death, hospitalization, heart attack, recurrence, etc.). This regression equation will only include "immutable covariates" that cannot be controlled during the course of medical care, such as age, sex, race, comorbidity, and living location. Logistic regression is useful for predicting expected outcomes; through this model, we are able to discover outliers among these two classes based on patients' nature conditions. This step creates labels we will use for discriminative pattern identification: "better than expected", "worse than expected", and perhaps, "as expected".

Identifying patterns of care.

After obtaining two patient datasets, labeled as "better than expected" and "worse than expected", we are going to construct frequent sequential treatment patterns among them. Several sequential pattern algorithms such as GSP, SPADE, PREFIXSPAN are considered and time constraints will be incorporated into the mining algorithm. Initially, we will use drugs and procedure sequences as the treatment pattern. This will generate the most frequently observed sequential treatment patterns.

Sequential pattern discrimination

The last step is to discover the sequential treatment patterns that are most closely assoicated with better and worse than expected outcomes. After generating two sets of frequent sequential treatment pattern, it is important to inspect hidden implications carried by those patterns because not all frequent patterns related to outcome can be explained by human. We will take some naïve methods like entropy to discover which sequential patterns are most significant to discriminate between better than expected and worse than expected outcomes. Those discriminative sequential treatment patterns can be further applied on classification or prediction. They may also be potential for generating unexpected clinical discovery.

Alternative strategies.

We may also explore alternative strategies. For example, we might use an unsupervised approach to identify the most frequently observed patterns in the entire database of patients with incident disease as describe above,and add these as binary indicator variables in a logistic regression that also includes the immutable covariates (demographics, comorbidities,etc.) that predict an outcome of interest. Any patterns that show significant explanatory power in such a regression would be strong candidates for quality measures. This would enable us to exploit all of the observations in the original data set, but may require a model selection process and approaches to ensure that explanatory power associated with immutable covariates (e.g. repeated treatments for a comorbid condition).

Impact

Two major contributions are expected to be completed through this project, one is for unexpected medical facts discovery and the other is for clinical path improvement. For the first purpose, we hope through our methodology, we can find out unexpected but meaningful treatment patterns which are hard to be discovered through conventional approach. We expected that our method could expedite medical research process and become a useful tool for this domain. Another goal is for enhancing the treatment quality. We expect that our mined sequential treatments could be a useful prediction model for increasing the expected outcome probability for patients at early stage.

Timeline

August 15-August 20: Data management

August 20-September 1: Logistic Regression and outlier detection

September 1-October 1: Implementation of graphical methods for temporal pattern recognition

October 1-December 1: Iterative review of dominant patterns for face validity with domain experts