OJPHI

ISDS 2016 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Identification of Sufferers of Rare Diseases Using Medical Claims Data

Jieshi Chen* and Artur Dubrawski

Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA

Objective

To identify sufferers of a rare and hard to diagnose diseases by detecting sequential patterns in historical medical claims.

Introduction

Patients who suffer from rare diseases can be hard to diagnose for prolonged periods of time. In the process, they are often subjected to tentative treatments for ailments they do not have, risking an escalation of their actual condition and side effects from therapies they do not need. An early and accurate detection of these cases would enable follow-ups for precise diagnoses, mitigating the costs of unnecessary care and improving patients' outcomes.

Methods

A sequential rule learning algorithm[1] was applied to a medical claim dataset of about 1,700 patients, who are pre-selected to have medical histories indicative of Gaucher Disease (GD) but only 25 of these patients were confirmed positives. About 168,000 medical claims and 142,000 pharmaceutical claims were featurized into sequences of asynchronous events and regularly sampled time series as inputs for the model, such that an occurrence of a certain diagnosis code in a medical claim was counted as one event along the timeline of the patient's medical history. Similar method was applied to other key attributes of claims data including procedure codes, National Drug Codes, Diagnosis Related Groupers, etc. These types of events as well as their temporal statistics, e.g. moving frequencies, peaks, change points, etc., formed the input feature space for the algorithm which was trained to adjudicate each test case and estimate their likelihood of having GD. A random forest algorithm was also applied to the same feature set to comparatively evaluate the utility of sequential aspects of data. The models were evaluated with 10-fold cross-validation.

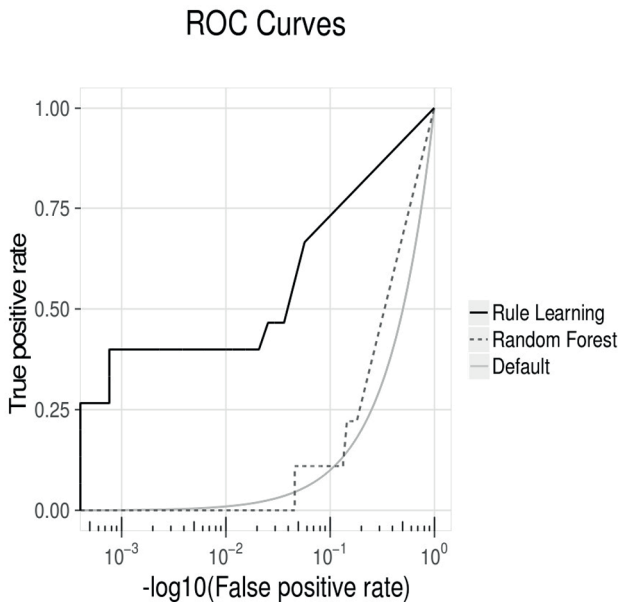
Results

Figure 1 shows the Receiver Operating Characteristic (ROC) curves of the temporal rule model with Area Under the Curve score exceeding 81% and significantly outperforming the random forest and default models. Considering the practical costs to perform follow-up genetic tests, we prefer a model achieving high positive recall at low risk of false detection. Our model correctly identifies more than 25% of known positive cases well within 0.1% of the false positive rate, while the performance of a more popular alternative is indistinguishable from random. This demonstrates the utility of sequential structure of medical claims in identifying patients who suffer from rare diseases.

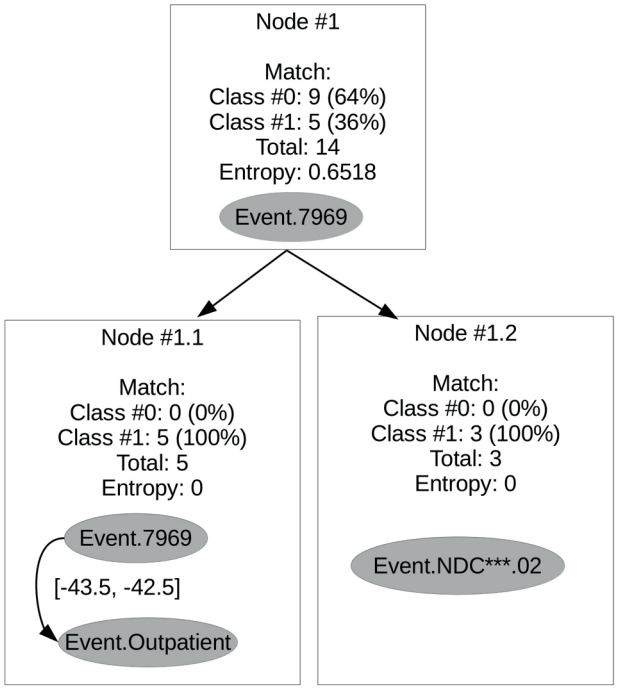
Our algorithm infers from data highly interpretable rules it uses in case adjudication. Figure 2 illustrates one of them. The root node of the case adjudication tree (Event.7969) reflects the ICD-9 diagnosis code of "Other nonspecific abnormal findings". Among the 14 patients that have this particular ICD-9 code present in their claim history, 36% are confirmed GD sufferers. Compared to default prevalence in our pre-selected data set of 1.47%, this rule lifts the estimated likelihood of GD 25 times. The rule further develops into two children nodes. The left child node adds the condition of having any outpatient claim observed within 43 claims recorded nearby the occurrence of the root node event. It isolates 5 patients all of whom are GD-positive. The right child shows that 3 patients without Event.7969 in their claim history but prescribed NDC 62756-0137-02 (Gabapentin by Sun Pharmaceutical Industries Ltd.) are all GD-positive. This is just one example of a simple and easy to implement business rule that is capable of identifying previously undiagnosed sufferers of rare diseases.

Conclusions

Our model successfully utilizes sequential relationships among events recorded in medical claims data and reveals interpretable patterns that can identify sufferers of rare diseases with high confidence. The algorithm scales well to large volumes of medical claims data and it remains sensitive in despite of a very low prevalence of target cases in data.



ROC diagrams of models trained to identify GD patients shown with decimal logarithmic scale of the false positive rate axis.



Example rule used to adjudicate GD cases.

Keywords
sequential patterns; medical history; rare diseases

Acknowledgments
This work has been partially supported by NSF (1320347) and CMU Disruptive Health Technology Institute.

References
1. Guillame-Bert M, Dubrawski A. Classification of Time Sequences using Graphs of Temporal Constraints. Journal of Machine Learning Research, 2016 (under review).

***Jieshi Chen**
E-mail: jieshic@andrew.cmu.edu