# Decoding and Analyzing Medical Data

Sirui Li(sl4653) & Xiaoli Sun(xs2338)
Medical Fraud Detection
February 10, 2020

## Introduction

According to a 2018 study by the Centers for Medicare & Medicaid Services (CMS), U.S health care cost $3.65 trillion in 2018, which suggests the average American household spent $11,121 per person on health care the last year[1]. With billions of insurance claims filled, The National Heath Care Anti-Fraud Association conservatively estimates that approximately 3% of the health care spending, $110 billion annually, are fraudulent. One the one hand, for patients, no matter they have employer-sponsored or self-afford health insurance, medical fraud undoubtedly increase their premiums and out-of-pocket cost. One the other hand, medical fraud also brought the unnecessary expense for insurance company's business. After all, medical fraud inevitably leads to a waste in the medical care system, which worse the cycle of health care system. From a business perspective, for example, needless medicine and tests translate higher operation costs for healthcare company, so company might raise the premium for consumers in order to balance the income and expenditure. Moreover, from a humanity perspective, unnecessary or unsafe medical treatment could slow the speed of intaking new patients, giving treatment, or even risking patients' life. Summarizing potential disadvantages, we could expect medical fraud not only harms the operations of insurance companies and providers, but also brings consumers greater financial burden and poorer health care service.

Taking a closer look at the past medical fraud cases, we can see most of the fraud is committed by health care providers. Unlike some other services, medical service makes consumers more vulnerable due to the requirement of professional knowledge. Considering making our work more functional, we come up with various development directions for our project. Firstly, there is no denying that the majority of patients would completely trust their doctors, and some dishonest providers are abusing their trust. Therefore, there is a need for a classification model to automate the detection of medical insurance anomalies. Secondly, successfully controlling disease at an early stage would undoubtedly benefit everyone in the medical system, whether you're the health provider, patients or insurance company. In this case, we hope to create a model that can retrieve patients with similar symptoms and treatment, in order to predict the probability of potential illness, furthermore, recommend providers prevention treatment.

[1]https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/ForecastSummary.pdf

# Related Work

**Visual Progression Analysis of Event Sequence Data (2019)**
This paper talks about the procedure of event sequence analysis and provides us with several algorithms to implement our data test. It explains how to find similar high-level structures in sequential event pattern analysis. It shows how to achieve event alignment, which solves a big challenge in our project. Finally, it gives a visual analysis to show the outcome of EDA.

**Visual analytics for temporal event sequence recommendation. (2016)**
This paper shows how to analyze event sequence data. The author analyzes the collected dataset of archived students' academic activities, and discuss how to augment the review manager's ability to make personalized recommendations for each student. Finally, it gives a student academic planning to help students achieve more. Our project also related to event sequence data and recommendation systems. We can learn a lot from the procedure of process data and given algorithms.

**Evaluating collaborative filtering recommender systems. (2004)**
Our project hopes to recommend/predict proper treating procedures for a patient and compare this recommendation/prediction with the actual treatment patient received from doctor to detect medical insurance fraud. Paper discusses what is recommender systems and introduced related mathematical theorems. It introduces a set of recommender tasks that categorize the user goals for a particular recommender system. It addresses the selection of appropriate datasets and evaluates a wide set of metrics.

**Finding similar people to guide life choices: Challenge, design, and evaluation(2017)**
To understand users' needs, the author reflected and built on experience accumulated from working with case study partners (medical researchers, doctors, marketing and transportation analysts, etc.) for more than a decade while developing tools and interfaces for the exploration of personal records. Searching for similar records was requested by many users. The long-term goal is to support prescriptive analytics interfaces that guide users as they make plans informed by the history of similar people. Searching for similar records is the focus of this paper, while our project is aiming to search for similar patients.

**Discovering clusters in motion time-series data(2003)**
In this paper, the author focuses on the problem of finding groups, or clusters of similar object motions within a database of motion sequences, and estimating motion time series models based on these groups. The author employs a probabilistic model-based approach, where the data is assumed to have been generated by a finite mixture model. In particular, he assumes that the observed sequences are generated by a finite mixture of hidden Markov models (HMMs), and estimate this mixture of HMMs via an Expectation-Maximization (EM)

formulation. Many probability theorems and equations are introduced in the paper, which helps us to understand the structure of our time-series data better.

**VMSP: Efficient vertical mining of maximal sequential patterns. (2014)**
This paper proposed a novel algorithm for maximal sequential pattern mining that we name VMSP (Vertical mining of Maximal Sequential Patterns). The algorithm incorporates three efficient strategies named EFN (Efficient Filtering of Non-maximal patterns), FME (Forward Maximal Extension checking) and CPC (Candidate Pruning by Co-occurrence map) to effectively identify maximal patterns and prune the search space. VMSP is developed for the general case of a sequence database rather than strings and it can capture the complete set of maximal sequential patterns with a single database scan. We may need this algorithm to find patterns in the patient's clinical record and analyze our data.

**Exploring point and interval event patterns: Display methods and interactive visual query. (2012)**
To assist in the exploration of medical records, the author initially developed a powerful visualization tool, Lifelines2, which allows these records to be searched for meaningful sequences using the Align, Rank, Filter, and Summary concepts. Medical researchers and practitioners used this tool to support hypothesis generation, and find cause-and-effect relationships in a population. This work evolved further into the LifeFlow aggregation tool, which consolidates records with similar event patterns into a summarizing display. LifeFlow allows for an entire set of patient records to be represented, manipulated, and evaluated on a single screen. Both LifeLines2 and LifeFlow were designed to support point-based events, which are discrete events that occur at a single point in time. The final step of our project is to visualize the clinical record and show our result with graphs. This paper may give us some hints and useful techniques.

## Dataset

Our dataset is from MIMIC-III Critical Care Database[2]. MIMIC-III (Medical Information Mart for Intensive Care III) is a freely-accessible database, and it includes health data between 2001 and 2012 from over 40,000 ICU patients of the Beth Israel Deaconess Medical Center. Submitting a request to access the dataset is required.

With this comprehensive database, we acquired a large dataset(60G), which consisting of 40 tables, 534 columns, and 728,556,685 rows. Therefore, we have to import our data to Google Cloud Platform in order to utilize analysis tools, such as SQL, more easily. We will conduct a preliminary categorizing by diseases ICD-9 codes and focus on one category before milestone 3.

[2] http://www.nature.com/articles/sdata201635

## Approach

For data cleaning and medical terminology implementation part, we use the data schema[3] provided by MIMIC to connect tables. ICD-9 code and the Wikipedia to determine how diseases are related. From the description of the disease crawled from wikipedia, we will implement NLP skills to related symptoms and proper treatment for specific diseases, so that we can cluster diseases into different groups.

Additionally, we will try to utilize EventThread, which attempted to tackle the stage identification problem by algorithmically segmenting groups of event sequences into fixed-width time intervals, then grouping similar event sequence segments into clusters (i.e. soft patterns) within each interval. Each cluster can then be summarized statistically to characterize a series of latent stages of event sequence progression. We apply EventThread Method to the clinical records of patients and try to get every patient's event sequence progression. Synthesized all these event sequence progressions from different patients, we have an analyzable dataset.

Furhtermore, we will summarize a clinical pathway from the dataset which regularly records all kinds of patient therapy and treatment activities in clinical workflow by various hospital information systems. The proposed approach formally defines the clinical pathway summarization problem as an optimization problem that can be solved in polynomial time by using a dynamic-programming algorithm. More specifically, given an input event log, the presented approach summarizes the pathway by segmenting the observed time period of the pathway into continuous and overlapping time intervals, and discovering frequent medical behavior patterns in each specific time interval from the log.

A deep learning model with two recurrent neural networks (RNNs) to predict the likelihood of occurrence for a set of potential diseases based on a patient's historical medical record. RNN-A calculates weights that indicate the accumulated influence on the prediction results at each time point. RNN-B calculates weights that indicate the accumulated influence on the prediction results at each time point. Synthesize the outcome of the two RNNs, we can make some predictions. Then we use skip-gram algorithm and procedures introduced in related work to analyze this event sequence progression dataset. For this part, we hope to detect the anomaly, predict the possibility of diseases, and recommend treatment given a cluster of clinical event sequence progressions. Beyond this, we will build a Web-Front-End visualization demo for the results of our model.

## Challenge

Firstly, this project requires us to intensively study new knowledge, including medical knowledge to create related terminology matching graph, natural language processing skills to encoding data from text to numbers, and machine learning strategies for model training.

[3] https://mit-lcp.github.io/mimic-schema-spy/

Moreover, we are concerned about the feasibility of anomaly detection. Discussed with a friend who works in the health insurance industry. Due to the lack of fraud tags, we found out it might be hard to define medical fraud with our limited data. For example, a doctor could give a patient test A, B and C since the symptoms might result from various diseases. Given this situation, we probably cannot determine if this doctor is giving unnecessary tests or not. However, we would try to implement some real-world fraud data to help our detection model building.

## Plan

Through this project, we aim to use machine learning and natural language processing strategies and models to create a deception detection, disease prediction and treatment recommendation system, starting within the scope of disease included in one chapter of ICD-9[4], and eventually expand to a more general scope. This project could be integrated into a visualization platform for various functions display and use.

For our milestone 1, we would start from studying related work and writing proposals. After meeting with instructor, we've already acquired the proper dataset, and we're setting up Google Cloud Platform for loading dataset and future programming use. With data loaded, we would get our hands on cleaning data and creating the medical knowledge graphs for all chapters of ICD-9 diseases.

Milestone 2 would include model building and training. Trying to build algorithms for medical fraud detection, probability of potential disease prediction, and early-stage prevention treatment recommendation, however, we will focus on the scope of the first chapter of ICD-9 diseases (001–139: infectious and parasitic diseases).

In milestone 3, we would broaden our scope and implement our models for the disease of the rest of chapters of ICD-9 code. We are also considering conduct accuracy tests for all functions, with the implementation of real-world medical claims and fraud case data to evaluate the credibility of our work. Moreover, we are intending to design and create a Web-Front-End visualization demo for our work.

For final project, we would discuss with instructor to find out potential methods of improving the accuracy. To keep a record of our project, we will write a comprehensive final report, and create a video for project presentation.

[4] https://en.wikipedia.org/wiki/List_of_ICD-9_codes

## Reference


[1]National Health Expenditure Projections 2018-2027. (n.d.). Centers for Medicare & Medicaid Services. Retrieved February 14, 2020, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/ForecastSummary.pdf

[2]Shrank, W. H., Rogstad, T. L., & Parekh, N. (2019). Waste in the US Health Care System, Estimated Costs and Potential for Savings. *JAMA*, *322*(15). doi: 10.1001/jama.2019.13978

[3]MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: *10.1038/sdata.2016.35*. Available from: http://www.nature.com/articles/sdata201635

[4]Guo, S., Jin, Z., Gotz, D., Du, F., Zha, H., & Cao, N. (2019). Visual Progression Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 417–426. doi: 10.1109/tvcg.2018.2864885

[5]Du, F., Plaisant, C., Spring, N., & Shneiderman, B. (2016). EventAction: Visual analytics for temporal event sequence recommendation. *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. doi: 10.1109/vast.2016.7883512

[6]Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. doi: 10.1145/963770.963772

[7]Huang, Z., Lu, X., Duan, H., & Fan, W. (2013). Summarizing clinical pathways from event logs. *Journal of Biomedical Informatics*, *46*(1), 111–127. doi: 10.1016/j.jbi.2012.10.001

[8]Du, F., Plaisant, C., Spring, N., & Shneiderman, B. (2017). Finding Similar People to Guide Life Choices. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI 17*. doi: 10.1145/3025453.3025777

[9]Alon, J., Sclaroff, S., Kollios, G., & Pavlovic, V. (n.d.). Discovering clusters in motion time-series data. *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* doi: 10.1109/cvpr.2003.1211378

[10]Chen, Y., Xu, P., & Ren, L. (2018). Sequence Synopsis: Optimize Visual Summary of Temporal Event Data. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 45–55. doi: 10.1109/tvcg.2017.2745083

[11]Fournier-Viger, P., Wu, C.-W., Gomariz, A., & Tseng, V. S. (2014). VMSP: Efficient Vertical Mining of Maximal Sequential Patterns. *Advances in Artificial Intelligence Lecture Notes in Computer Science*, 83–94. doi: 10.1007/978-3-319-06483-3_8

[12]Monroe, M., Wongsuphasawat, K., Plaisant, C., Shneiderman, B., Millstein, J., & Gold, S. (n.d.). Exploring point and interval event patterns: Display methods and interactive visual query. Retrieved from http://www.cs.umd.edu/hcil/trs/2012-06/2012-06.pdf