MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231046

Predicting Medical Event Occurrence Using Medical Insurance Claims Big Data

Hiromasa YOSHIMOTO[a,1], Naohiro MITSUTAKE[b], and Kazuo GODA[a]
[a] *Institute of Industrial Science, The University of Tokyo, Japan*
[b] *Institute for Health Economics and Policy, Japan*
ORCiD ID: Hiromasa Yoshimoto https://orcid.org/0000-0001-9702-2879

Abstract. Medical events are often infrequent, thus becomes hard to predict. In this paper, we focus on predictor that forecasts whether a medical event would occur in the next year, and analyzes the impact of event's frequency and data size via predictor's performance. In the experiment, we made 1572 predictors for medical events using Medical Insurance Claims (MICs) data from 800,000 participants and 205.8 mil- lion claims over 8 years. The result revealed that (a) forecasting error will be increased when predicting low-frequency events, and (b) increasing the number of training dataset reduces errors. This result suggests that increasing data size is a key to solve low frequency problems. However, we still need additional methods to cope with sparse and imbalanced data.

Keywords. Insurance claims, perdition model, machine learning, sparse data, low-frequency event

1. Introduction

Medical and healthcare institutions yield vast amounts of digital data every day. One of the goals of the informatics field is to utilize these big data to optimize future; big data analytics can be used to predict medical events in the future and facilitate decision-making ahead of them [1-5]. One of the datasets for this purpose is medical insurance claims (MICs), that are bills submitted by medical service providers to patient's insurance providers. These invoices record details of medical services, including medical procedures (e.g., consultations, surgeries, dental care, etc.), dosages and frequencies of medications, and medical equipment used. In Japan, these data are recorded electronically in a standardized format. When medical services are provided, a 9-digit number corresponding to the service provided is recorded with the participant ID, age, gender, date, and amount billed. An advantage of using MIC is the full coverage of patients; as the Japanese government has achieved a universal healthcare service system since 1961, all citizens are required by the law, to have health insurance, where most MIC records (93% or more) are recorded electrically. Similar systems have been adopted in South Korea, Taiwan, and other countries [6].

Another data source is electronic health record (EHR). EHR contains personal statistics like body weight, blood pressure, smoking, income, and X-ray images, etc.

[1] Corresponding Author: Hiromasa YOSHIMOTO; he currently works at Institute for Health Economics and Policy, Japan, email: yoshimoto@tkl.iis.u-tokyo.ac.jp.

These personal data were known as good explainer for medical outcomes and widely used in medical research [3,7].

However, these personal statistics are highly confidential pieces of data, and it is not easy to collect vast amounts of data from multiple medical institutions [6]. Collecting vast amounts of data may leads new findings in the medical field. For example, genome research reveals that analyzing low-frequency variation in gene coding sequence can explain sclerosis risk [8]. In general, low-frequency variation is difficult to analyze because the datasets tend to be huge, sparse, and imbalanced. However, these problems may be solved or mitigated by increasing both amounts of data and computer's resources. In this study, we focus on prediction task of medical events occurrence using MIC big data. The dataset we used consists of 205.8 million records about 800,000 participants. Each record includes diagnostic names (e.g., diabetes, hypertension), medical practices (e.g., nursing, surgery, dental care), the amount and frequency of medications, and medical equipment (e.g., gauze, syringe). Based on these real data, this paper analyzes the performance of the predictors in terms of frequencies of events.

2. Methods

2.1. Pre-processing of MIC records

The longitudinal analysis of insurance claims is achieved following the steps outlined below. First, an anonymized personal identifier is used to trace each participant [9], thus creating a time sequence of medical events for the individuals. Each event in the sequence is assigned a nine digits integer code that identifies the type of medical service provided to the participant. The coding system is managed by Japan governments and is updated rapidity. For example, every commercial medicine has a unique code, and once it is discontinued, the code becomes obsoleted. Similarly, when a new disease such as COVID-19 appeared, it is assigned a new code. As the result, the possible code space spreads to 109 with many reserved spaces. In other words, code table becomes very sparse one.

To deal with this sparsity of coding system, we use typical methods in topic modeling, as known as bag-of-words (BoW) representation and its related methods. First, we aggregate all medicine's codes by its active ingredient name. Next, we make a dictionary consisting of pairs of nine digits codes and their number of occurrences in the entire MIC. Then, we filter out codes based on frequency. This is because codes that are too rare or too common are not informative. In the experiment, we remove codes with a frequency of occurrence of less than 1/10,000 or more than 50%. Note that although this code filtering is a common technique in BoW representation, it is also known to cause undesirable artifacts. The artifacts must be investigated in our future work.

After the filtering, we acquire total N codes, where N is the number of codes in the dictionary. With N codes, we encode each record as a BoW-style feature vector of length N.

2.2. Classifiers

Using BoW representation, the classifier can be formulized as a binary classifier that predicts a code occurrence from a feature vector containing N past event occurrences. To predict whole N event occurrences, we just combine N binary classifiers. We make each binary classifier using extreme gradient boosting (XGBoost) algorithm [10]. For XGBoost, the dataset is randomly divided into two portions for training and testing. The training dataset is used for supervised learning of binary classifies. First, all records in the training dataset are first aggregated by participants ID, and then extracted by using two time-windows defined at year t. One window is past five years, and the other is for next one year. With these two time-windows, corresponding events are merged into two BoW vectors: v_{past} and $v_{feature}$. Sliding these time windows, we generate three pairs at $t = \{2018, 2019, 2020\}$. Additionally, we append participant's age and gender to v_{past}. With these procedures, we finally obtain labeled training data as a set of (age, gender, v_{past}) and $v_{feature}$ so that we can train N binary classifies, where i-th binary classifier, $i = \{1, \dots, N\}$ forecasts i-th element of $v_{feature}$.

In our experiment, tuning of hyper parameters for XGBoost were achieved by using grid search approach used with K-fold cross validation, where we used K=3 in our experiment.

2.3. Performance evaluation

Model performance is evaluated using two metrics: area under the ROC curve (AUC) and balanced accuracy (BA). Here, AUC summarizes the trade-off between the true and false positive rates for the predictive model; BA is widely used metrics for imbalanced data. We also analyzed feature ranks calculated by permutation importance method. We used python 3.10 and related modules: XGBoost 1.6.2 for making binary classifiers, SciPy 1.9.3 for spare matrix manipulations, scikit-learn 1.1.2 for the scripting implementation framework, and shap 0.41.0 for explanation analysis [11].

3. Results

We used the MICs dataset of public insurance providers in Gifu prefecture area in Japan. This dataset contains approximately 800,000 participants aged 0 to over 100 years, and their 205.8 million claims recorded from April 2014 to March 2022. From this dataset, our procedure extracted 1572 different events automatically (N =1572), thus converted them into BoW feature vectors by using SciPy's sparse matrix data container. Final data file size became about 4.2GB total when using .npy file format.

Next, we split the participants into 80% for train data and 20% evaluation data randomly. The Training data is further divided into K(=3) parts in K-fold cross-validation in learning phase. We used remaining 20% of the participants' data for evaluation purpose only for fair performance evaluation.

Figure 1 shows an overview of the performance evaluation. Each graph consists of N points that correspond to the binary classifiers; its position represents the performance in predicting the occurrence of the event. The x position represents the frequency of occurrence of the i-th event. The frequency of occurrence of events ranges from 0 to 35,000 per 100,000 population in a year. The y position of each point represents its performance by using AUC or BA metrics. The distribution of the points indicates that when the frequency of occurrence of an event is low, in other words rare events, the prediction performance may become low. On the other hand, when the frequency is higher than 3,000 per 100,000 population in a year, the AUC is about greater than 0.75. This figure also shows the difference between AUC and BA; if BA is greater than AUC;

it means that there are more false positives or false negatives. This means that when predicting infrequent events, false positives and false negatives are more likely to occur.

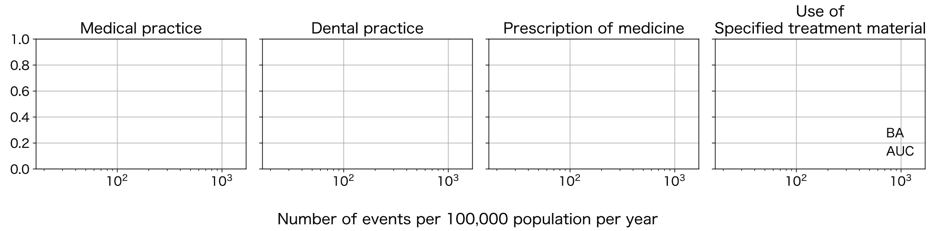

Figure 1. Distribution of 1572 predictors performance

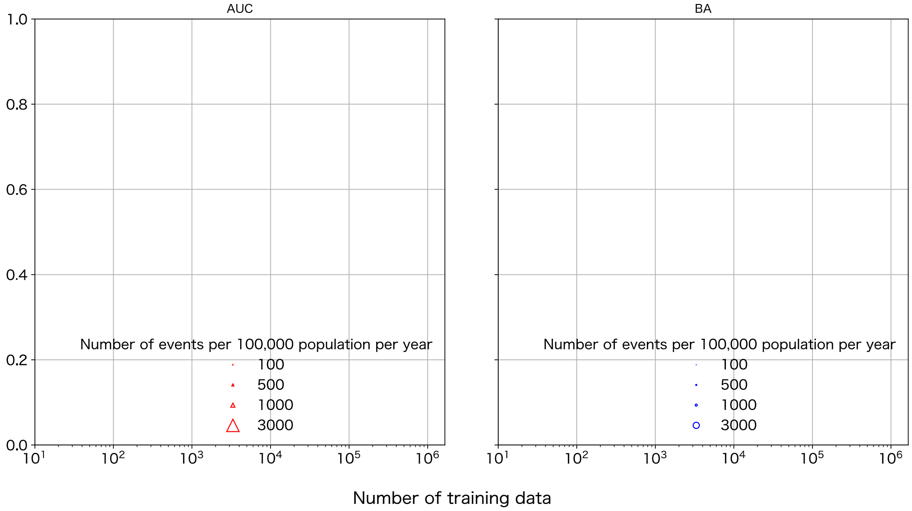

Figure 2. The impact of training dataset size.

Figure 2 shows the impact of the size of training dataset. Its horizontal axis represents size of training dataset, and the vertical axis represents performance metrics. Those two graphs shows that XGBoost-based classifiers require at least 10,000 or 100,000 participant claims data in our configuration.

4. Discussion

The main contribution of this paper is to present both state-of-the-art performance and its limitation in medical event predictors by using actual medical big data. The distribution of the performance visualized in Figures 1 and 2 present profiles of the performance, thus helps those who start a new study by using MICs or related big data. However, several limitations of this study deserve mention. First, the influence of the length N of the BoW feature vector should be analyzed more deeply. The two thresholds to determine N could be optimized by using grid search method, however, this paper did not cover this point. Second, Figure 2's BA score seems not to be saturated even if 500,000 or more participant claims are used. As described above, BA tends to ignore the false positive and negative, thus may arises an over fitting. To investigate this issue, we still need additional methods to cope with sparse and imbalance data.

5. Conclusions

In this study, we have proposed an algorithm for predicting future medical events from past medical events. The experimental result shows the structure of the difficulties behind the medical event's prediction; the frequency of each medical event's occurrences influences the prediction performance. In our configuration, XGBoost-based classifiers require data from at least 10,000 or 100,000 participants to forecast the occurrences of medical practices (e.g., nursing, surgery, dental care). The next possible challenge is to focus on imbalanced data, especially ultra low-frequency events.

References

[1] Sato J, Mitsutake N, Kitsuregawa M, Ishikawa T, Goda K. Predicting demand for long-term care using Japanese healthcare insurance claims data. Environ Health Prev Med. 2022 Oct;27:42.

[2] Barbieri S, Kemp J, Perez-Concha O, Kotwal S, Gallagher M, Ritchie A, Jorm L. Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk. Sci Rep. 2020 Jan;10(1):1111.

[3] Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, Haga A, Shinzawa M, Yamakawa M, Fukui S, Moriyama T. Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Sci Rep. 2022 Oct;12(1):15889.

[4] Kaushik S, Choudhury A, Sheron PK, Dasgupta N, Natarajan S, Pickett LA, Dutt V. AI in healthcare: time-series forecasting using statistical, neural, and ensemble architectures. Front Big Data. 2020 Mar;3:4.

[5] Dall TM, Storm MV, Chakrabarti R, Drogan O, Keran CM, Donofrio PD, Henderson VW, Kaminski HJ, Stevens JC, Vidic TR. Supply and demand analysis of the current and future US neurology workforce. Neurology. 2013 Jul;81(5):470-8.

[6] Umemoto K, Goda K, Mitsutake N, Kitsuregawa M. A prescription trend analysis using medical insurance claim big data. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE); 2019 Apr 8; p. 1928-39.

[7] Wang Y, Chen R, Ghosh J, Denny JC, Kho A, Chen Y, Malin BA, Sun J. Rubik: knowledge guided tensor factorization and completion for health data analytics. KDD. 2015 Aug;2015:1265-74.

[8] International Multiple Sclerosis Genetics Consortium. Low-frequency and rare-coding variation contributes to multiple sclerosis risk. Cell. 2018 Nov;175(6):1679-87.e7.

[9] Sato J, Yamada H, Goda K, Kitsuregawa M, Mitsutake N. Enabling patient traceability using anonymized personal identifiers in japanese universal health insurance claims database. AMIA Jt Summits Transl Sci Proc. 2019 May;2019:345-52.

[10] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug; p. 785-94.

[11] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec; p. 4768-77.