HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2020 June 23.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2019 May ; 2019: . doi:10.1109/bhi.2019.8834638.

Feature Exploration and Causal Inference on Mortality of Epilepsy Patients Using Insurance Claims Data

Yuanda Zhu,

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Hang Wu,

Dept. of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

May D. Wang

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Abstract

Approximately 0.5–1% of the global population is afflicted with epilepsy, a neurological disorder characterized by repeated seizures. Sudden Unexpected Death in Epilepsy (SUDEP) is a poorly understood complication that claims the lives of nearly 1-in-1000 epilepsy patients every year. This paper aims to explore diagnosis codes, demographic and payment features on mortality of epilepsy patients. We design a mortality prediction model with diagnosis codes and non-diagnosis features extracted from US commercial insurance claims data. We present classification accuracy of 0.91 and 0.85 by using different feature vectors. After analyzing the aforementioned features in prediction model, we extend the work to causal inference between modified diagnosis codes and selected non-diagnosis features. The uplift test of causal inference using three algorithms indicates that a patient is more likely to survive if upgrading from a low-coverage healthcare plan into a high-coverage plan.

Keywords

epilepsy; insurance claims data; feature selection; causal inference

I. Introduction

Approximately 50 million people worldwide are afflicted with epilepsy [1], a neurological disorder characterized by repeated seizures. Sudden Unexpected Death in Epilepsy (SUDEP) is a poorly understood complication that claims the lives of nearly 1-in-1000 epilepsy patients every year [2]. Although the risk of SUDEP is shown to increase in those patients with refractory cases of epilepsy [3], the underlying cause of the condition remains elusive. To support understanding of epilepsy and SUDEP, we aim to predict mortality on

yzhu94@gatech.edu.

epilepsy patients using insurance claims data and explore the casual relationship between diagnosis codes and non-diagnosis information on epilepsy patients.

Few studies have associated epilepsy patients with insurance claims data. Shcherbakova et al [4] identified clinical, medication and demographic factors related to seizure recurrence with epilepsy patients receiving antiepileptic monotherapy by using US commercial insurance claims data from Jan. 2007 to Sept. 2010. They demonstrated that comorbid conditions and prior seizures are the likely predictors for epilepsy patients to require urgent care at one-year follow up. Cramer et al [5] explored the utilization and cost on epilepsy patients using an insurance claims database covering a study period from 1/1/2007 to 12/31/2009. The author confirmed that uncontrolled epilepsy patients have higher economic cost and require more non-epilepsy-related healthcare services than stable epilepsy patients.

How do we evaluate the relationship between diagnosis codes and non-diagnosis information, such as utilization/ cost? Causal inference is an effective tool. Identifying the causal effect of one variable on another is a key subject of causal inference, and has various applications in biomedical data analysis, for example, a recent study on the increase of emergence department usage due to Medicare insurance [6]. Specifically, we have a binary variable T and an outcome Y, and we aim to identify the average treatment effect (ATE) of administering T on Y, i.e., the *uplift* difference between two outcomes $Y(T = 1)$ and $Y(T = 0)$ [7].

In this paper, we work on the clinical (diagnosis codes) features and non-diagnosis features of epilepsy patients from a very large commercial insurance claims database. We have two main contributions:

- We design a mortality prediction model to figure out top features contributing to death on epilepsy patients and recommend a competitive feature vector that concatenates both diagnosis codes and non-diagnosis features.
- We explore the causal relationship between diagnosis code and non-diagnosis features. By utilizing uplift models, we demonstrate that a patient is more likely to survive if upgrading from a bad healthcare plan into a good healthcare plan, given the same conditions.

This paper is organized as follows: in section II, we introduce the U.S. commercial insurance claims database Truven and epilepsy patients in Truven. In section III, we explore the importance of features extracted from inpatient and outpatient health record by performing mortality prediction on epilepsy patients with different feature vectors and discuss the prediction results corresponding to top ranking features. We recommend a fair yet competitive feature vector in the end. In section IV, we validate the casual relationship that better healthcare plan gives rise to lower mortality rate on epilepsy patients by using three different uplift algorithms. In section V, we summarize our work and indicate possible future work.

II. Dataset

A. Data Tables in Truven Database

Insurance claims data in Truven database contain heath records for around 10 million patients across the U.S from 01/01/2011 to 09/30/2016. Truven has three major data tables: Inpatient Admission, Outpatient Claims and Prescription Claims.

Inpatient Admission table has the most critical information, such as up to 15 diagnosis codes, up to 15 procedure codes and discharge status per record. Discharge status is the only information that indicates whether a patient is dead or alive by the end of the hospital visit. In addition, Inpatient Admission includes demographic information, such as gender, age, Metropolitan Statistical Area (MSA, which is equivalent to city, town or county) and state. Healthcare plan information, including coordination of benefits (COB), copay, co-insurance (COIN) and deductible, and net payment as well as total payment, are also included.

Outpatient Claims table has fewer diagnosis codes and procedure codes, while preserving the same amount of demographic, healthcare plan and payment information. Prescription Claims table has prescription information, such as National Drug Code and refill limits.

B. Epilepsy in Truven

After consulting physicians from Union Chimique Belge (UCB) Pharma., we adopt the following criteria to identify epilepsy patients: 1) at least one epilepsy diagnosis code (345 in ICD9, G40 in ICD10) or 2) at least two convulsion codes (7803 in ICD9, R56 in ICD10). As shown in Fig. 1, we identify a total of 972,008 epilepsy patients. Among them, only 14,025 patients are indicated dead from their inpatient admission records, and only 365,049 live epilepsy patients have inpatient visits.

III. Feature vectors and mortality prediction

In this section, we design a mortality prediction model to find out top features contributing to death on epilepsy patients and recommend optimal feature vector based on our definition. Our initial definition towards mortality prediction is: given all diagnosis codes and non-diagnosis features, predict whether an epilepsy patient will die by the end of the last visit.

A. Feature vector: Diagnosis codes + non-diagnosis features

We first extract all inpatient and outpatient visits for each epilepsy patient, filling the diagnosis codes into a sparse matrix: each row corresponds to one visit to the physician/hospital, and the columns are all diagnosis codes in the database; then we condense the matrix into one row by summing up the occurrence of each diagnosis code. This vector of diagnosis codes is highly sparse: among all 45,547 diagnosis codes, in average less than 1% entries are non-zero.

When defining non-diagnosis entries, we include demographic information, health plan information and payment information from the last day of visit for each epilepsy patient in the database. To be more specific, each epilepsy patient shall have a non-diagnosis feature vector that consists of 11 entries: age, gender, state, MSA, Plan type, COB, co-insurance,

copay, deductible, net payment and total payment. The non-diagnosis feature vector is shown in Table 1.

B. Classification

We choose gradient boosting, the ensemble method implemented by scikit-learn [8] for classification. To search for the optimal classification results, we loop the learning rate from 0.01 to 0.5. The learning rate of 0.1 or 0.2 provides the highest accuracy. We select all 14,025 dead patients and randomly sample the same number of live patients from the data set. In addition, 5-fold cross validation is implemented to average the performance of each fold. Three performance metrics are considered: accuracy, f1 score and Matthews correlation coefficient (MCC).

C. Results and Discussion

As indicated in Table 2, the feature vector with all 45,547 diagnosis codes and 11 non-diagnosis features achieves 0.912 accuracy. We realize that, nevertheless, *Net Payment* and *Total Payment* from the last visit are too dominating when ranking the 11 non-diagnosis features alone using Spearman correlation score (shown in Table 3) and analyzing the top 10 features from the entire feature vector in classification (shown in Table 4). Besides, since in practice *Net Payment* and *Total Payment* wouldn't be obtained shortly after the patients' visit to the hospital, we decide to remove them from death prediction feature vector.

After modifying our feature vector as all diagnosis codes plus the nine non-diagnosis features, the predication accuracy drops to 0.887 (shown in Table 2). We further notice that, certain top-ranking diagnosis codes in Table 3, such as *acute respiratory failure* and *cardiac arrest*, tend to appear during the last visit before the patient dies. These factors are too close to death, which makes prediction meaningless. For fairness consideration, we eliminate all diagnosis codes from the last visit.

Now the new definition towards mortality prediction is: given the diagnosis codes of *(N-1)* visits and the nine non-diagnosis features, predict whether the epilepsy patients will die during the last (*Nth*) visit. We assume that demographic information and health plan information remain the same for each epilepsy patient. Consequently, our recommended feature vector consists of all diagnosis codes without the last visit and the nine non-diagnosis features. This feature vector leads to prediction accuracy 0.859.

In addition, we also test feature vector of 11 non-diagnosis features, nine non-diagnosis features and diagnosis codes without the last visit. It's interesting to notice that feature vector of 11 non-diagnosis features alone outperforms all feature vectors without *Net Payment* and *Total Payment*. Besides, the concatenation of all diagnosis codes and the nine non-diagnosis codes outperforms either one of them. Thus, our recommended feature vector is meaningful and competitive.

IV. Uplift Causal Inference on healthcare plan

From previous section, we figure out that both diagnosis codes and non-diagnosis features contribute to mortality prediction. In this section, we would like to show the causal

relationship between them. We define the problem as whether the same patient with feature vector X will have a lower mortality rate (denoted as Y) by upgrading a bad healthcare plan ($T = 0$) into a good healthcare plan ($T = 1$), i.e., the difference of mortality probability with/out a good healthcare plan $ATE = P(Y = 1;\ T = 1|X) - P(Y = 1;\ T = 0|X)$. Since in reality we cannot change the real healthcare plan for the patients, we can only evaluate the average treatment effect (ATE) on the virtual manipulation in causal inference model.

We categorize the patients' plans based on the ratio between net payment and total payment from the last visit. We eliminate "inadmissible" records, including zero total payment, negative net payment and net payment larger than total payments from our data set. In Fig. 2 and Fig. 3, we observe that the majority of patients have healthcare plan that covers more than 50% of total cost. Thus, we define a good healthcare plan to have the ratio between net payment and total payment smaller than 50%. Vice versa, a bad plan indicates the ratio over 50%.

We test three uplift algorithms, transformed outcome method implemented by pylift [9], as well as ordinary least square (OLS) and matching methods from Causal Inference Package [10]. We select all "admissible" 13,823 dead patients and randomly sample the same number of live patients from the data set five times, to test the aforementioned algorithms.

As shown in Table 5, the mean value of Average Treatment Effect (ATE) from all three algorithms fall into the interval of [−0.191, −0.181] and the standard deviation is small (< 0.005). The results indicate that a patient tend to have a lower mortality rate if upgrading from a bad healthcare plan to a good healthcare plan given the same diagnosis code condition.

V. Conclusion and Future Work

In this paper, we design a mortality prediction model to figure out top features leading to death on epilepsy patients and recommend an ideal feature vector that concatenates diagnosis codes and non-diagnosis features. We have shown that both diagnosis codes and non-diagnosis features contribute to the prediction of death on epilepsy patients. All diagnosis codes with 11 non-diagnosis features lead to prediction accuracy 0.912; our recommended feature vector, diagnosis codes without the last visit concatenated with 9 non-diagnosis features lead to prediction accuracy of 0.859. The latter feature vector is a safer definition towards prediction settings: given the *(N-1)* hospital records, predict death on the *Nth* visit.

In addition, we explore the causal relationship between diagnosis codes and non-diagnosis features. By using three different algorithms of uplift, we successfully demonstrate that an epilepsy patient is more likely to survive if upgrading into a better coverage healthcare plan given the same diagnosis condition.

Further work can include hierarchical features on diagnosis codes and non-diagnosis features. Besides, Long Short-Term Memory (LSTM) is capable of extracting time series features from health records. Statistical methods, such as Hawkes Processes, can be implemented to record the time interval between hospital visits.

Besides, causal inference can be extended to explore clinical variables. For example, with professional clinical background knowledge, people can test whether a certain procedure code can replace another procedure code and consequently lead to a higher survival probability using the insurance claims data on one specific cohort of patients.

Acknowledgment

This work was supported by the funding from UCB Pharma. The author would thank Edward J.K. Han-Burgess, Jonathon Williams, Robertson Joseph, Eun Jung Choi and Sunil Ajagekar from UCB Pharma. for providing guidance on understanding the epilepsy patients in Truven database.

The author would also thank Johnny Chen from Georgia Institute of Technology for his exploration on phenotyping of the epilepsy patients.

This research project is supported by BME UCB subproject 12567GV.

References

[1]. Megiddo I, Colson A, Chisholm D, Dua T, Nandi A, and Laxminarayan R, “Health and economic benefits of public financing of epilepsy treatment in India: An agent-based simulation model,” Journal of the International League Against Epilepsy, 2016.

[2]. Hirsch LJ, Donner EJ, So EL, “Abbreviated Report of the NIH/NINDS Workshop on Sudden Unexpected Death In Epilepsy,” Neurology, 76(22): 1932–1938, 2011. [PubMed: 21543734]

[3]. Spencer DC, “SUDEP: Sudden Unexpected Death in Epilepsy on Placebo?,” Epilepsy Currents, 2012.

[4]. Shcherbakova N, Rascati K, Brown C, Lawson K, Novak S, Richards KM and Yoder L, “Factors associated with seizure recurrence in epilepsy patients treated with antiepileptic monotherapy: A retrospective observational cohort study using US administrative insurance claims,” CNS Drugs, vol. 28, no. 1047, 2014.

[5]. Cramer JA, Wang ZJ, Chang E et al., “Healthcare utilization and costs in adults with stable and uncontrolled epilepsy,” Epilepsy & Behavior, vol. 31, pp. 356–362, 2014. [PubMed: 24239435]

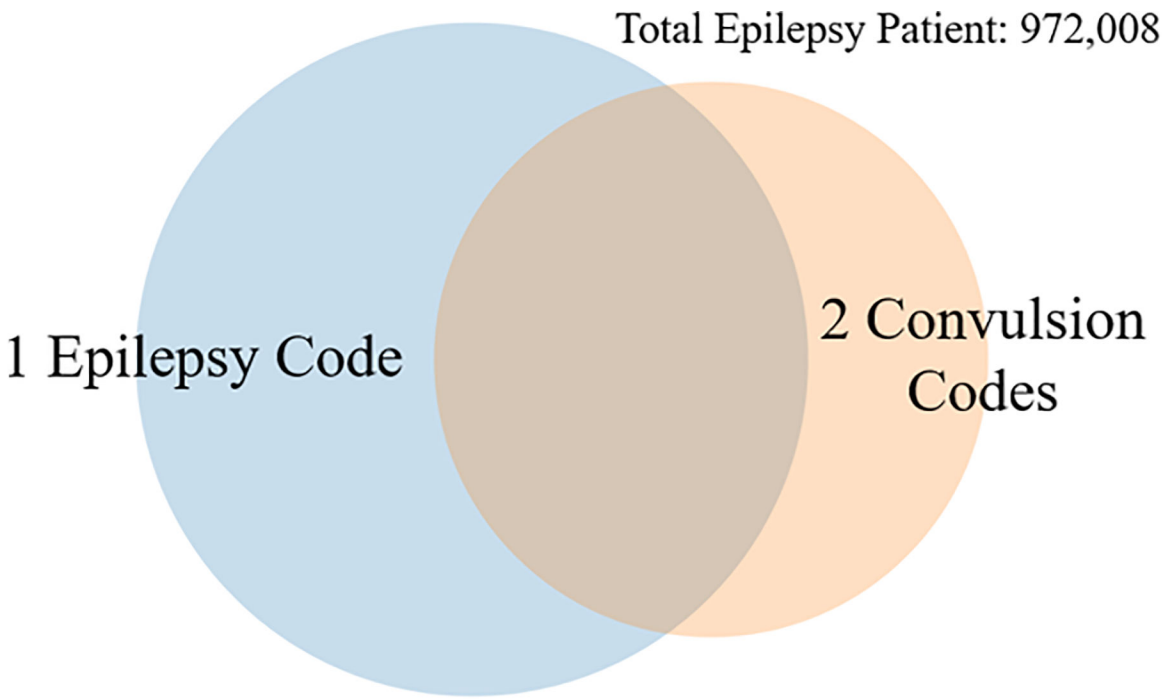
[6]. Taubman SL, Allen HL, Wright BJ, Baicker K and Finkelstein AN, “Medicaid Increases Emergency-Department Use: Evidence from Oregon’s Health Insurance Experiment,” Science, vol. 343, no. 6168, 2014.

[7]. Gutierrez P and Gérardy J, “Causal Inference and Uplift Modelling: A Review of the Literature,” Proceedings of The 3rd International Conference on Predictive Applications and APIs, vol. 67, 2017.

[8]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B and Grisel O et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825—2830, 2011.

[9]. “Transformation-based uplift modeling package,” [Online]. Available: https://github.com/wayfair/pylift.

[10]. Wong L, “Causal Inference in Python,” [Online]. Available: http://causalinferenceinpython.org/.



	1 Epi Code only	Intersection	2 Convolution Codes only
# Patients	445,702	328,103	198,203

Figure 1.
Venn map for components of epilepsy patients in Truven.

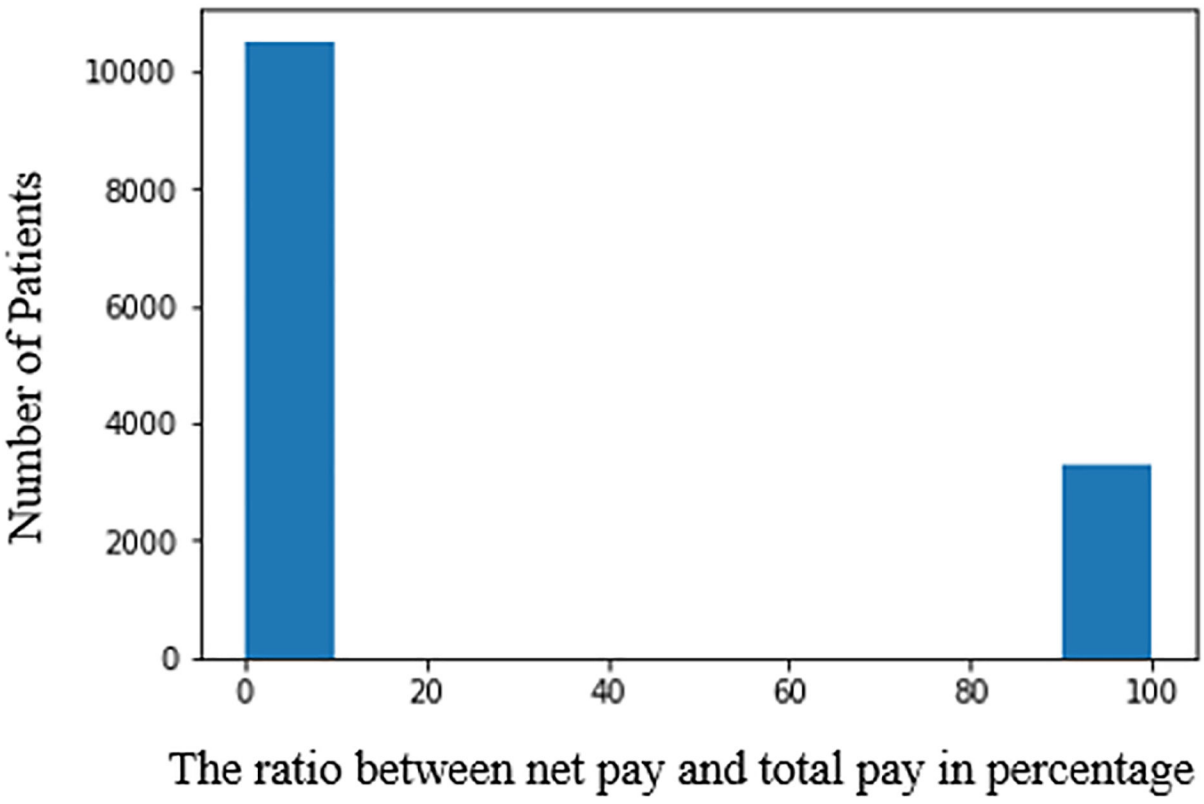


Figure 2.
Healthcare plan percentage coverage for dead patients.

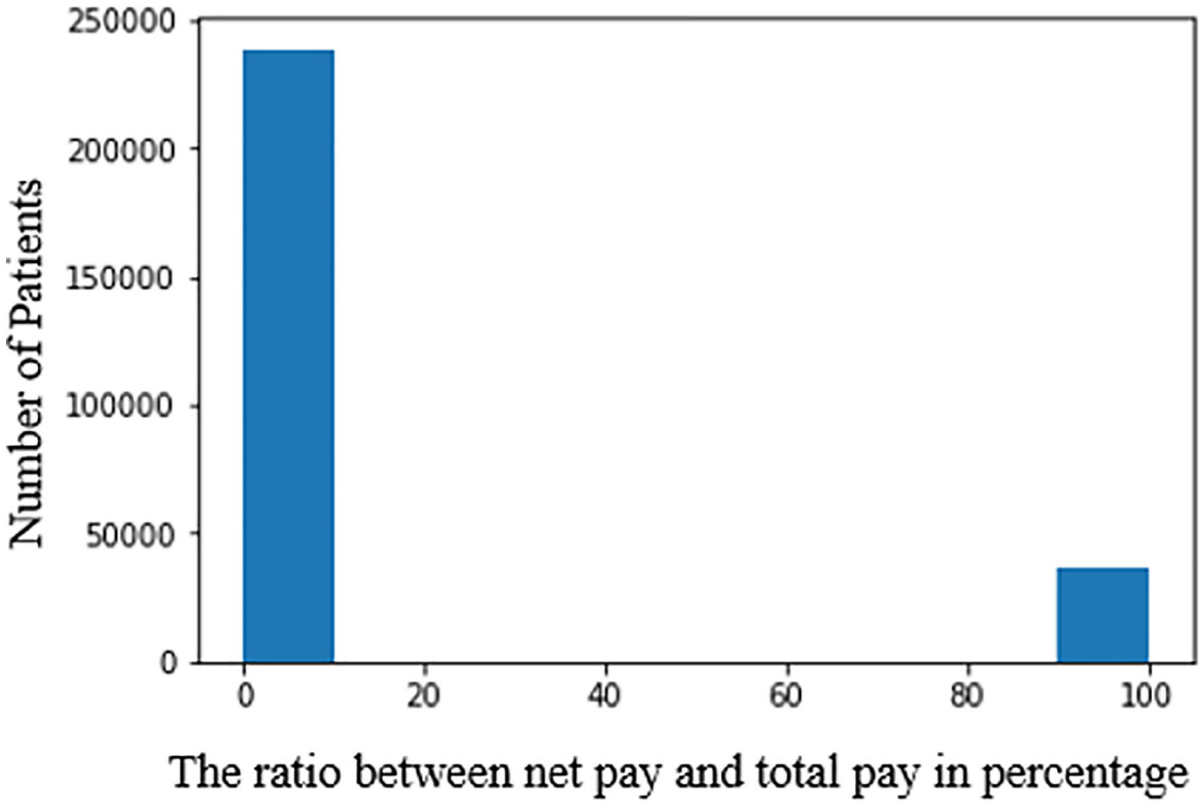


Figure 3.
Healthcare plan percentage coverage for live patients.

Table 1.

An example of feature vector of 11 non-diagnosis entries.

Age	Gender	State	MSA	Plan Type	COB	COIN	COPAY	Deductible	Net Pay	Total Pay
75	1	28	35840	8	0	0	300	0	15754	16054

Table 2.

Prediction results by using different feature vectors described in section III. The performance is ranked from high to low accuracy. For each feature vector, the best result is shown out of all classifiers.

Feature Vector	Accuracy	F1 Score	MCC
Diagnosis codes (including last visit) + 11 non-diagnosis	0.91240	0.91207	0.82481
11 non-diagnosis only	0.89954	0.89556	0.80198
Diagnosis codes (including last visit) + 9 non-diagnosis	0.88741	0.88759	0.77489
Diagnosis codes (without last visit) + 9 non-diagnosis	0.85929	0.85746	0.71907
9 non-diagnosis only	0.80848	0.79990	0.62171
Diagnosis codes only (without last visit)	0.78563	0.79433	0.57361

Table 3.

Spearman correlation score between 11 non-diagnosis features and mortality. The score is between −1 and 1. A score close to 1 indicates strong positive correlation; a score close to 0 indicates low correlation; a score close to −1 indicates strong negative correlation.

Non-diagnosis Feature	Spearman Correlation Score
Total Payment	0.670
Net Payment	0.663
Age	0.277
Copay	0.242
COB	0.230
Deductible	0.197
MSA (City)	0.016
Coinsurance	−0.046
State	−0.058
Gender	−0.078
Plan Type	−0.083

Table 4.

Top 10 features from feature vector of 45,547 diagnosis codes and 11 non-diagnosis entries. The third column is short description to the entries on the second column.

Ranking	All Features	
1	Total Pay	
2	Net Pay	
3	ICD9: 51881	Acute respiratory failure
4	ICD9: 4275	Cardiac arrest
5	AGE	
6	ICD9: 0389	Unspecified septicemia
7	STATE	
8	ICD9: V667	Encounter for palliative care
9	ICD9: V4986	Do not resuscitate status
10	ICD9: 78039	Unspecified convulsions

Table 5.

Average Treatment Effect with different methods across five different sampling of dead and live patients from the entire data set.

Method	Average	Std.
Transformed Outcome	−0.19169	0.00234
OLS	−0.19075	0.00426
Matching	−0.18160	0.00273