# Duke Statistics Portfolio Report: Medicare Billing Fraud Detection and Analysis via Multiple Approaches

Yue Li (yl874)

March 2024

## 1 Introduction

Healthcare fraud is a type of white-collar crime wherein dishonest claims are filed to gain a profit. Fraud influences the healthcare system not only financially, but also places a significant burden on the perceived integrity and data value of the system. he Centers for Medicare & Medicaid Services, part of the Department of Health and Human Services, reported that the national health expenditure grew 4.6%, to 3.6 trillion dollars, in 2018. This figure translated to $11,172 per person, for billions of claims. Furthermore, the National Healthcare Anti-Fraud Association estimated that approximately tens of billions of dollars are lost due to healthcare fraud each year. This immense financial loss places the responsibility of recovery on insurance companies, but more importantly, on patients. Patients are cheated into compensating for the cost in primarily two ways: payment of fraudulent copays and higher insurance premiums. Thus, it is pertinent to determine the patterns in healthcare fraud and take preventative measures against such crimes.

Here are some common types of billing frauds in the healthcare system:

- Duplicated billing: involves deliberately charging twice for a service or product that was only performed or supplied once.
- Phantom billing: involves billing for a test or procedure or other medical service that was never actually performed.
- Upcoding: alters the codes assigned to specific billable services to reflect a higher-level service than what was actually performed, resulting in higher reimbursement cost.
- Unbundling: involves taking a comprehensive service and separating it into several specific services in order to bill for each one independently, resulting in higher reimbursement cost.
- Kickback: occurs when a provider accepts payment on behalf of a pharmaceutical company or medical device supplier, by exchanging of recommending or prescribing patients to use the product.

In this project, we would like to utilize the data from Center for Medicare & Medicaid Services in 2022 to analyze health insurance data at the level of healthcare providers, and uncover the methods used to commit fraud by detecting any patterns of inconsistencies within the data. The followings address my research questions:

- What are common patterns of billing fraud?
- What are some key characteristics of patients susceptible to fraud?
- What are some key characteristics of fraudulent healthcare providers?
- Are there any network effects between providers, doctors, and patients showing billing fraud?

## 2 EDA, Preprocessing, Feature Engineering

Our dataset comes from `CMS.org`, Center for Medicare & Medicaid Services. Table 2 is a summary of the 4 dataset. The dataset contains outpatient data (claim-level data for patients who didn't stay at hospital), inpatient data (claim-level data for patients who stayed at hospital), beneficiary data (information for patients who submitted claims), and provider data (indicate whether the provider is fraudulent).

Before getting into some details about the extensive data analysis done on the claims data, I would like to discuss how the data was preprocessed. First off, the missingness in the data was handled. The data included a lot of missing values, such as missing date of death if the patient is alive and missing operating physician if no surgical operation was performed. Missing information was imputed accordingly. Also, for uniform and efficient preprocessing, all categorical data was label encoded. We also decided to keep outliers in the data, as the outliers could provide key fraud indicator information. These outliers could very well be transactions where actual fraud is being committed. This was also the reason why the data was robust scaled before modeling.

**Q: Do the number of doctors and patients affect the probability of encountering potentially fraudulent providers?** According to the scatter plot above, we found that with a greater number of patients, doctors, or both, the probability of the provider being potentially fraudulent increased. As the number of patients and doctors decreased, there were less cases of providers being potentially fraudulent. This indicates that the larger providers (bigger hospitals with greater networks) might be more likely to be fraudulent.

**Q: Would patients with more chronic conditions have greater number of claims filed in contrast to patients with less chronic conditions?** Although we expected that the number of chronic conditions a patient has and the number of claims filed for the patient would share a positive correlation, we instead found that the highest number of claims were filed for patients with 4-6 chronic conditions; the graph shows a normal distribution.

**Q: How are deductible amounts and insurance reimbursed amounts distributed for inpatients and outpatients?** As can be seen in the top graph, the inpatient deductible amount paid is consistent at a value of about $1100, whereas the outpatient deductible amount paid is more varied with greater distribution between $0 to $200. However, the most frequent value is still $0. The bottom graph shows that the outpatient insurance claim amount reimbursed also tends to be near $0, though with a distribution between 0 and $20,000.

Conversely, the inpatient insurance claim amount reimbursed had a much wider and higher range of values with the maximum amount being reimbursed around $120,000. This indicates that inpatient services are significantly more expensive than outpatient services. The graph below summarizes the outpatient/inpatient costs based on averages.

**Q: Is there a difference between potentially fraudulent and non-fraudulent providers depending on the types of services they offer?** For non-fraudulent providers, we found that the number of providers offering solely outpatient services was significantly higher than those providers offering only inpatient services or those offering both inpatient and outpatient services. On the other hand, the number of fraudulent providers offering both inpatient and outpatient services is considerably higher than those offering either inpatient or outpatient services. This also indicates that, again, the larger providers are more likely to be fraudulent.

**Q: Are the total counts of claims for different claim admit diagnosis codes greater for potentially fraudulent or non-fraudulent providers?** As can be seen in the graph above, non-fraudulent providers surprisingly had the larger counts of claim admit diagnosis codes with exceptions to two codes: 486 and 78650. Thus, further research and analysis should be done on these two codes.

We are presented with three dataset of features at the level of patients and claims, whereas the target is the at the provider level flagged as potentially fraudulent or not. We need to aggregate and transform the inpatient, outpatient, and beneficiary data to create a new dataset based on the provider. The merged dataset contains 79 feature. Table 3 displays the categorical breakdown of the features we engineered for further representing the data, and there are some new findings.

**`TotalClaimAmount` and `DailyTotalCharge`:** After combining the features, we see a stronger distinction between fraudulent and non-fraudulent features within finances. In the left plot, the total per claim median is approximately $340 for non-fraudulent providers, and $2700 for fraudulent. On average, the total claim amount for fraudulent providers is $2500 more than the total claim amount for non-fraudulent providers. In the right plot, distributions for daily total charge are shown, and we found on average, fraudulent providers charge $470 more per day than non-fraudulent providers.

The results concerning the financial data are insightful as well as understandable. It is logical to *hide the fraud within total claim* to distribute the fraudulent activity and remain inconspicuous versus *overcharging within one area* where there is a fixed pattern that is more detectable.

**`numDiffDiagnosisCode` and `numDiffGroupDiagCode`:** The median number of unique group diagnosis codes for fraudulent providers is 24, whereas for non-fraudulent, it is 0, and on average fraud providers have used 38 more codes. The median number of unique claim admit diagnosis codes for fraudulent providers is 57 whereas for non- fraudulent, it is 7, and on average, fraud providers have used 65 more codes. Referencing back to the EDA, we found that the total number of claims per code was not higher for fraudulent providers. Instead, here we find that it is the count of the number of unique codes used that is an important signifier. This also ties into networks and types of services offered. Providers mostly flagged as fraudulent are operating at higher levels with greater networks in bigger hospitals and are operating within both inpatient and outpatient. Thus, it definitely holds that the number of unique codes used will be greater for fraudulent providers. This was a very interesting find, and the unique group diagnosis codes count will be further explored in association rule mining.

## 3 Predictive Modeling

I am using five models for predicting `PotentialFraud` for providers: logistic regression, AdaBoost, Random Forest, Extreme Gradient Boosting (XGBoost), and LightGBM. In short, one linear classifier, one bagging classifier, and three boosting classifiers. The reason for choosing these classifiers are no less than for predicting fraudulent providers, but more inclined to choose the most important features for fraudulent providers and patients susceptible to fraud. The data is highly imbalanced like most insurance data. We need to upsample the minority class (fraudulent target) using SMOTE by synthesizing new examples from the minority.

If we assume no model is perfect, we acknowledge that misclassification occurs and that each type of misclassification carries a cost. The two misclassifications are: failing to detect a fraudulent provider (FN) and falsely accusing an innocent provider of fraudulent (FP). The cost of not identifying a fraudulent provider is to let the theft of reimbursements continue and serve as an enticement for others to commit fraud. The cost of FP's is reputatioanal damage of the provider, but also extra investigative costs and legal costs. We attempt to maximize the number of claims and ratio of the amount of money identified to the investigation expenses. In machine learning language, we penalize both FPs and FNs, since the former represent extra cost and the latter reduce the number of claims identified for the recovery. So our main evaluation metric will be F1-score, the harmonic mean of precision of sensitivity.

The machine learning pipeline is the following:

- Train-test split
- Upsample the minority class via SMOTE
- Standardize the features
- Gridsearch for the best parameters for each model
- Model evaluation via F1-score
- Model interpretation

Table 1 summarizes the model performances. Tree-based models outperform linear model (logistic regression). However, logistic regression has a low FNR. Boosting (XGBoost & LightGBM) outperforms bagging (Random Forest). LightGBM slightly outperforms XGBoost. Then, let's see what features has the highest ranking by feature importance for logistic regression, XGBoost, and LightGBM: `NumUniqGroupDiagCode`, `NumUniqClmAdmitDiagCode`, `Total Claim Amount`, `Service Type`.

# 4 Associate Rule Mining

From the results of predictive modeling, `NumUniqGroupDiagCode` is selected by both LightGBM and XGBoost, the two models yielding the best result. Also from EDA, we see that, on average, fraudulent providers use 38 more DRGs than non-fraudulent ones. We would like to know why Group Diagnosis Code (DRGs) is important for provider billing frauds. DRG is a way of classifying patients under a particular group, within which similar levels of medical resources are consumed. Each DRG has a payment weight assigned to it. For example, if a physician simply records the diagnosis as "X", X being a disease type, the lowest or neutral DRG category will be applied. Recording the diagnosis as "acute X" means a higher DRG category will be applied. A difference between these categories could mean double or triple the medical cost. This explains how `numDiffGroupCode` affects the strength of `DailyTotalCharge` and other financial features. Upcoding and unbundling are two types of billing frauds when DRGs are assigned inconsistently by doctors during diagnosis.

To study this, we need to determine what common patterns of chronic conditions might be associated with fraudulent providers. **Association rule mining** is an approach to extract common patterns and relations between items using the Apriori algorithm. First, let us look at some concepts of association rule mining:

- $I$ is the item set, $I = \{i_1, i_2, \ldots, i_m\}$. An itemset of cardinality $k$ ($1 \leq k \leq m$) is called $k$-itemset.
- $T$ is the transaction set, $T = \{t_1, t_2, \ldots, t_n\}$, $t_i \subseteq I$ for all $i$ and each $t_i$ is identified by a unique ID

In our example, $I$ is the 11 chronic conditions patients have when visiting hospitals, and $T$ contains chronic profiles for each patient.

An association rule is a directed rule $X \to Y$ between two itemsets $X$ (antecedent) and $Y$ (consequent), $X, Y \subseteq I$ and $X \cap Y = \emptyset$. In our example, if such an association rule $\{C_1, C_2\} \to C_3$ exists, it means that patients with chronic conditions $C_1$ and $C_2$ are likely to have condition $C_3$.

There are two important measures in association rule mining. The support of an item $X$, $\text{Supp}(X) = \frac{\#X}{|T|}$, is the occurrence of it within the transaction set. For an association rule $X \to Y$, the confidence of the rule, $\text{Conf}(X \to Y) = \frac{P(X \cap Y)}{P(X)}$ indicates the probability that given items in itemset $X$, items in itemset $Y$ will occur. In the Apriori algorithm, we need to set minimum support for extracting frequent itemsets, and minimum confidence to find the strong rules between frequent itemsets.

There are two steps in association rule mining:

- Step 1: Find all itemsets satisfying the minimum support. These itemsets are frequent.
- Step 2: Extract rules with high confidence between the itemsets mined in step 1. These rules are strong rules.

We can complete the above steps using Apriori algorithm, which is based on Apriori Principle: if an itemset is frequent, then all of its subsets are frequent; if an itemset is not frequent, then all of its supsets are not frequent. Following this principle, the algorithm yields 1-itemsets $F_1$ by scanning through the whole data for the support for each item. Starting from $k = 2$, the algorithm generates $k$-itemset $C_k$ through frequent $k-1$-itemsets $F_{k-1}$ for pruning. The algorithm still needs to scan through the data to generate frequent $k$-itemsets $F_k$ from $C_k$. It is an iterative algorithm via increasing $k$ until no more frequent itemsets are generated. The process of generating frequent $k$-itemsets can be summarized by joining and pruning: the join step generates all possible $k$-itemsets, and the prune step removes the infrequent candidates according to the minimum support. Then we generate the association rules. For every frequent itemset $X$, generate all non-empty proper subset of $X$. For every such subset $A \subsetneq X$, let $B = X \setminus A$. If $\text{Conf}(X \to Y) \geq \text{Min-Conf}$, output the rule $A \to B$.

In our case, we choose the minimum support to be 0.2, and select the top 30 rules by descending confidence level. Figure 6 is a network concept map showing the filtered association rules between frequent chronic conditions. Diabetes and ischemic heart are the common consequent of all the selected association rules. They are strong associated, which mirrors reality where diabetes is recognized as a strong risk factor for coronary artery disease/ischemic heart. Among fraudulent providers, 76% of the patients have diabetes, and 71% of the patients have ischemic heart, while only 59% of the patients have heart failure, the third most common chronic condition shown by 5. It is worth noticing that we extract frequent chronic conditions of patients using the whole dataset, regardless of whether the provider is fraudulent. The patterns of chronic profiles for patients are very similar in both cases. Patients who have the most common chronic diseases in the healthcare system are susceptible to billing fraud.

Here are some reasons for mining associate rules of patients' chronic conditions. If we simply count the number of chronic conditions in fraudulent/non-fraudulent providers, the result tells us the number of each condition alone without giving out a complete profile of patients. As shown by EDA, a single patient will have multiple chronic conditions in the data. So counting will give us inconclusive information. Instead, if we represent the patient profile by common chronic patterns (association rules with high likelihoods), it will give us a detailed description of patient profile. Moreover, mining association rules greatly reduces the complexity of data wrangling to find out the relation between fraudulent providers and chronic illnesses. Since diabetes and ischemic heart are two consequents for the association rules with the highest confidence levels, we can look only look at the proportion of patients who visited the fraudulent providers that have diabetes and ischemic heart.

## 5 Network Analysis

We can extract fraud networks using network analysis between providers, doctors (`AttendingPhysician`), and patients. First, we must gauge the size and extent of the networks. In the EDA part, we have examined the number of providers operating for a given number of states. Most of providers are local. They operate in one state or might service one or two more, that are located on the border of states. However, as the given number of states a provider operates in increase, the relative proportion in the number of providers between non-fraudulent and fraudulent decreases: fraudulent providers are overrepresented in large provider operations that span services over several states.

I use the `igraph` and `ggraph` packages in R to plot the bipartite graphs. We look at each state's provider-patient-physician networks on county levels. These are **undirected** graphs. The red squares are fraudulent providers. The green squares are non-fraudulent providers. The triangles represent doctors. The cirlces represent patients. The edges between each actor and a provider is weighted and sized by the total reimbursements from all claims (`InscClaimAmtReimbursed`). The sizes of the vertices are proportional to strength of nodes, which is calculated by the sum of the edges connecting the vertices. Below are the provider-patient and provider-doctor networks for County # 200 in NY, one of the states with the most fraudulent providers.

For provider-patient network, large providers (likely to be general hospitals) are not fraudulent and connects with a lot of patient. While this seems appropriate, many fraudulent providers, which are smaller (likely to be clinics), share patients with those large providers. Some patients have multiple connections with both fraudulent and non-fraudulent providers. For provider-doctor network, we see fraudulent providers tend to have more connections with doctors. One of the them have very high reimbursement amounts from the claims shared with doctors, which is very suspicious. It seems that providers share patients more than doctors.

We can perform bipartite projection to showcase this. A bipartite projection takes a network with two types of actors (provider-doctor/patient) and projects into a single type, establishing links through shared relations from the full graph. Suppose the provider set is $X$ and the patient set is $Y$. The projection of the bipartite graph onto the provider set $X$ is a graph that is constructed such that the nodes are providers only, and edges join the providers because they share the same patient. The providers in the CMS data, across states and counties, are way more likely to be linked to one another through shared patients than through shared physicians. There are claims filed with the same patient for two different providers, but there is no claim fuled with the same doctor between providers.

We can also detect duplicated claims through provider-provider and doctor-doctor network across state. Through previous EDA, we see signs of duplicated claims, both from inpatient and outpatient providers, via duplicating medical records and reusing IDs. To identify these claims, we can cast a wide net by identifying claims with the same patient having the same three diagnosis codes (`ClmDiagnosisCode1-3`) as duplicated. We need to filter out many null entries which likely represented claims associated with patient visits with no diagnosis or procedure performed. We find three duplicated patterns (Figure 12, Figure 13, Figure 14) and interpretations in provider-provider networks:

- Unary: Provider $X$ duplicates claims internally
- Binary: Provider $Y$ receives claim from Provider $X$

- Ternary: Multiple arrows between providers indicate multiple claims being traded

We also find largest duplication network in the data, corresponding to NY-NJ provider network (Figure 15). The network is a combination of binary and ternary relations. We see providers with greater node strength tend to be both sources and targets of duplicated claims, and providers with smaller strength tend to targets of duplicated claims.

We can apply a similar approach in plotting doctor-doctor networks. A doctor who duplicates claims will tend to have a significant number of claims with the doctors, hence representing a node strongly linked to the others. Figure 11 is a part of the network in NY doctor network:

We can see two types of doctor networks. For the first type, the weight of each doctor are almost uniform across the network. For the second type, one doctor has the largest weight among all doctors, being the central doctor for duplicating claims in the network.

## 6 Conclusion and Future Work

Based on the three methods we used, we can have a comprehensive framework on characteristics of fraud providers and patients susceptible to fraud:

**Patients:**

- Aged between 65 and 82, with high number of claims filed.
- With ischemic heart failure and diabetes in their chronic profile, with high number of chronic conditions.
- Received high reimbursement amounts and paid high deductibles, thus with high claim amount.

**Providers:**

- With high total reimbursements filed in provider-patient network.
- With large networks of duplicated claims between providers, and between doctors.
- With both inpatient and outpatient services.
- With a higher number of DRGs and Claim Admit Diagnosis Code listed on claims.

**Recommendations:**

- Stricter regulations of assigning diagnosis codes to mitigate risks of upbundling and upcoding.
- Investigation into fraudulent networks, especially on big provider networks and shared patients to mitigate risks of duplicated billing.
- Pay attention to old patients with ischemic heart failure and diabetes

**Future directions:**

- We need more knowledge on how DRGs are assigned across different states, counties, or geographical units. In this case, we can further study upcoding and unbundling: we can judge whether the severity of the patient's chronic conditions are consistent with the diagnosis code assigned by the primary doctor; we can also determine whether it is legitimate to assign various DRGs to a single patient.
- Based on network analysis, we can construct graph features and combine them with traditional ML algorithms to boost in prediction accuracy. There are two ways to complete this, according to [1]. The first way is to construct graph features, as centrality measures of providers in the place of doctors and patients in the network, and add them to traditional machine learning techniques. The goal is still predicting whether the provider is fraudulent or not. The second way is to construct a graph with edges connecting providers and patients/doctors and train separate Graph Neural Networks. The goal, however, is to label the fraudulent providers within the provider-patient/doctor network.

All figures and tables are in the appendix.

## References

[1] Yeeun Yoo, Jinho Shin, and Sunghyon Kyeong. Medicare fraud detection using graph analysis: A comparative study of machine learning and graph neural networks. *IEEE Access*, 2023.

| Model | Precision | Recall | F1 Score | AUC Score | FPR | FNR |
|---|---|---|---|---|---|---|
| Logistic | 0.825 | 0.478 | 0.606 | 0.689 | 25.79% | 5.39% |
| AdaBoost | 0.692 | 0.664 | 0.678 | 0.684 | 3.25% | 38.50% |
| RandomForest | 0.805 | 0.718 | 0.759 | 0.772 | 13.59% | 9.22% |
| XGBoost | 0.895 | 0.821 | 0.856 | 0.862 | 8.95% | 4.83% |
| LightGBM | 0.906 | 0.828 | 0.866 | 0.871 | 8.58% | 4.28% |

Table 1: Model evaluations

# 7 Appendix

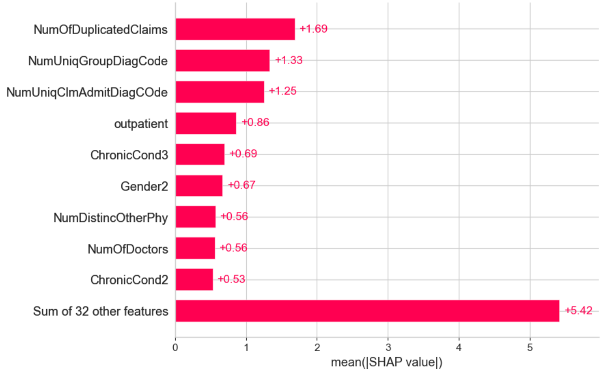

Figure 1: Logistic Regresion SHAP values

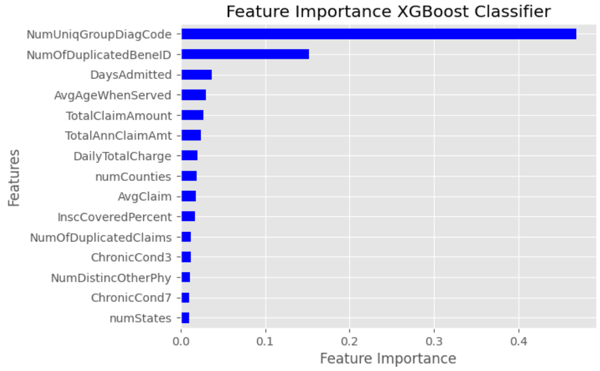

Figure 2: XGBoost Feature Importance

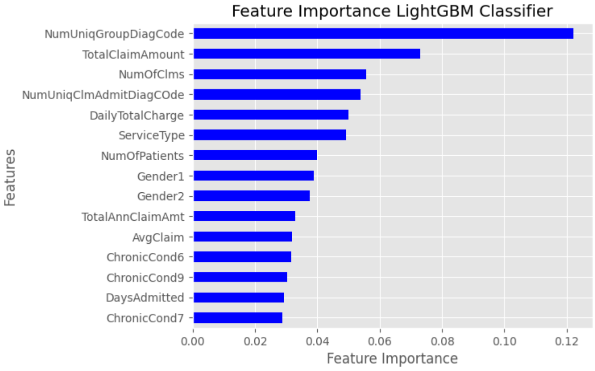

Figure 3: LightGBM Feature Importance

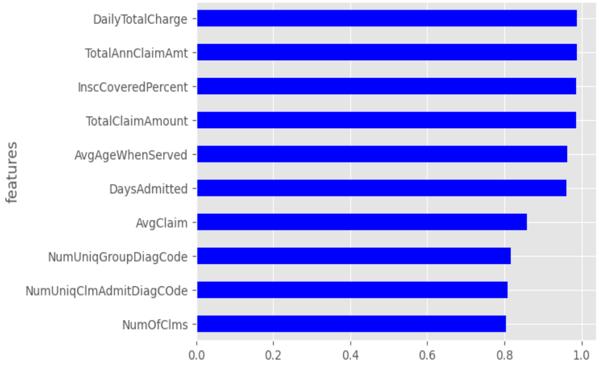

Figure 4: Unary classification check for accuracy of individual features

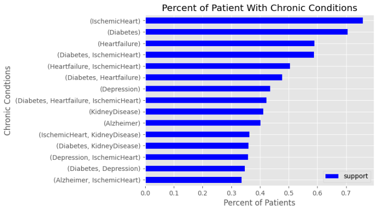

Figure 5: Generated frequent chronic condition(s) ranked by support, with minimum support 0.2.

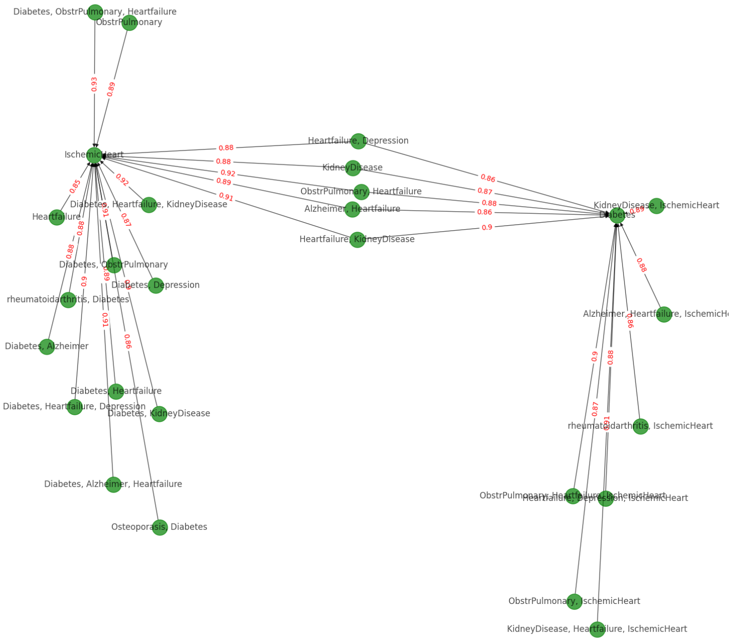
Diabetes, ObstrPulmonary, Heartfailure
ObstrPulmonary
0.93
0.89
IschemicHeart
Heartfailure, Depression
KidneyDisease
ObstrPulmonary, Heartfailure
Alzheimer, Heartfailure
Heartfailure, KidneyDisease
0.88
0.88
0.92
0.89
0.91
0.85
0.92
Diabetes, Heartfailure, KidneyDisease
Heartfailure
0.91
0.87
0.88
0.88
Diabetes, ObstrPulmonary
Diabetes, Depression
0.9
0.89
rheumatoidarthritis, Diabetes
0.91
0.86
Diabetes, Alzheimer
Diabetes, Heartfailure
Diabetes, Heartfailure, Depression
Diabetes, KidneyDisease
Diabetes, Alzheimer, Heartfailure
Osteoporasis, Diabetes
0.86
0.87
0.88
0.86
0.9
KidneyDisease, IschemicHeart
Diabetes
0.89
0.88
Alzheimer, Heartfailure, IschemicHe
0.86
0.9
0.88
0.87
0.91
rheumatoidarthritis, IschemicHeart
ObstrPulmonary, Heartfailure, IschemicHeart
Heartfailure, Depression, IschemicHeart
ObstrPulmonary, IschemicHeart
KidneyDisease, Heartfailure, IschemicHeart

Figure 6: Top 30 association rules between frequent condition(s), filtered by confidence level descend.

| Data | Column | Description |
|---|---|---|
| Outpatient & Inpatient | BeneID | Unique id of each patient |
| | ClaimID | Unique id of the claim submitted by the provider |
| | ClaimStartDt | Date when the claim started |
| | ClaimEndDt | Date when the claim ended |
| | Provider | Unique id of the provider |
| | InscClaimAmtReimbursed | Amount reimbursed for that particular claim |
| | AttendingPhysician | Id of the Physician who attended the patient |
| | OperatingPhysician | Id of the Physician who operated on the patient |
| | OtherPhysician | Id of the Physician other than AttendingPhysician and OperatingPhysician who treated the patient |
| | ClmDiagnosisCode 1-10 | Codes of the diagnosis performed by the provider on the patient for that claim |
| | ClmProcedureCode 1-6 | Codes of the procedures of the patient for treatment for that claim |
| | DeductibleAmtPaid | Amount by the patient before insurance plan (=Total claim amount — Reimbursed amount) |
| | ClmAdmitDiagnosisCode | Numerical Diagnosis Code on the claim indicating patient's initial diagnosis at admission |
| | DiagnosisGroupCode | Numerical Diagnosis Code assigned by the primary doctor according to diagnosis results |
| Inpatient | AdmissionDt | Date on which the patient was admitted into the hospital |
| | DischargeDt | Date on which the patient was discharged from the hospital |
| Beneficiary | BeneID | Unique id of the patient |
| | DOB/DOD | Date of Birth/Death of the patient |
| | Gender, Race, State, County | Demographics of the patient |
| | RenalDiseaseIndicator | Indicates if the patient has existing kidney disease |
| | IPAnnualReimbursementAmt | Maximum reimbursement amount for hospitalization annually |
| | IPAnnualDeductibleAmt | Premium paid by the patient for hospitalization annually |
| | OPAnnualReimbursementAmt | Maximum reimbursement amount for outpatient visits annually |
| | OPAnnualDeductibleAmt | Premium paid by the patient for outpatient visits annually |
| Provider | Provider | Unique id of the provider |
| | PotentialFraud | **Target**: whether the provider is fraudulent |

Table 2: Data Description

| Category | Column | Description |
|---|---|---|
| Financial | DeductibleAmtPaid | Average duration of patients stay in hospital by provider |
| | InscClaimAmtReimbursed | Average insurance claim amount reimbursed for patient per provider |
| | InscCoveredPercent | Average insurance coverage per provider |
| | TotalClaimAmount | InscClaimAmtReimbursed+DeductibleAmtPaid |
| | DailyTotalCharge | TotalClaimAmount/DaysAdmitted |
| Provider | DaysAdmitted | DischargeDt-AdmissionDt |
| | ServiceType | Type of service provided by provider (Inpatient/Outpatient/Both) |
| | numState | Number of states the provider operates in |
| | numCounties | Number of counties the provider operates in |
| Patient | NumOfPatients | Number of patients per provider |
| | NumChronicCond | Average number of chronic conditions per provider |
| | NumOfDuplicatedBeneID | Count of duplicated patient id per provider |
| | AvgAgeWhenServed | Average patient age per provider |
| Physician | NumOfDoctors | Number of doctors per provider |
| | NumOfDuplicatedAttendingPhysician | Count of duplicated attending physician per provider |
| | NumDistincOpPhy | Count of distinct operating physicians per provider |
| | NumDistincOtherPhy | Count of distinct other physicians per provider |
| Claim | NumOfClms | Number of claims filed per provider |
| | WeeklyClaims | Number of claims filed weekly per provider |
| | MonthlyClaims | Number of claims filed monthly per provider |
| | NumOfDuplicatedClaims | Count of duplicated claims filed per provider |
| | NumDistincClmProCode1-5 | Count of distinct procedure diagnosis code 1-5 per provider |
| | NumDistincClmDiagCode1-10 | Count of distinct claim admit diagnosis code 1-10 per provider |
| Code | numDiffDiagnosisCode | Number of claim admit diagnosis code per provider |
| | numDiffGroupDiagCode | Number of group diagnosis code per provider |

Table 3: Feature engineering

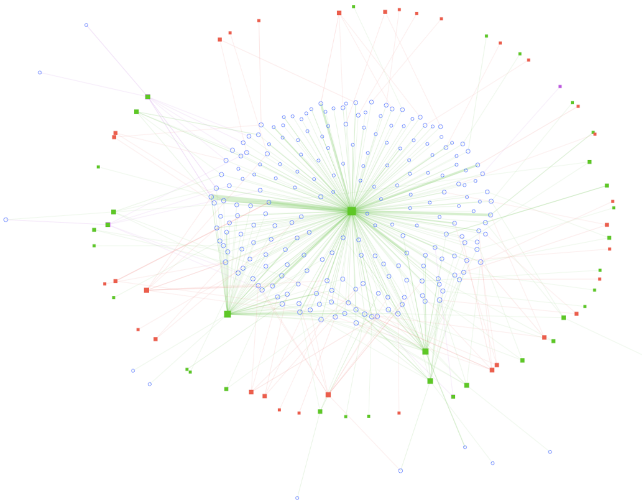

Figure 7: Provider-Patient Network for NY, County # 200

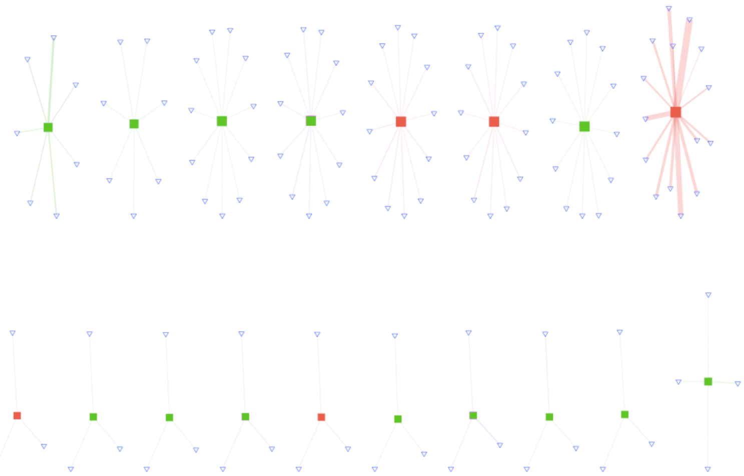

Figure 8: Provider-Doctor Network for NY, County # 200

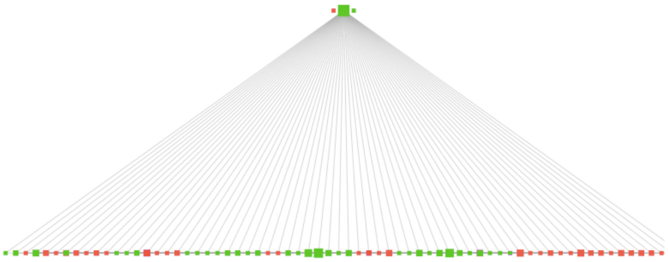

Figure 9: Projected Provider-Patient Network for NY, County # 200

Figure 10: Projected Provider-Doctor Network for NY, County # 200

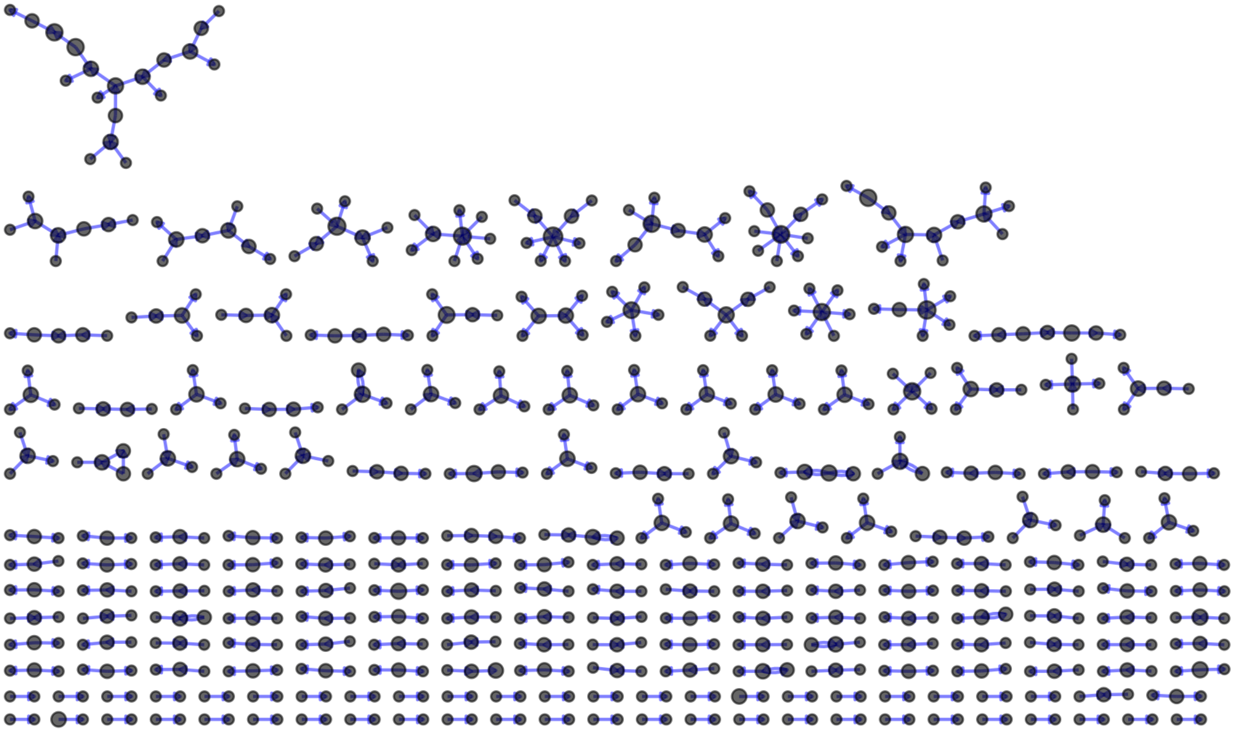

Figure 11: Portion of NY doctor network

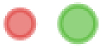

Figure 12: Unary relation

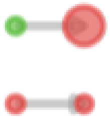

Figure 13: Binary relation

Figure 14: Ternary relation

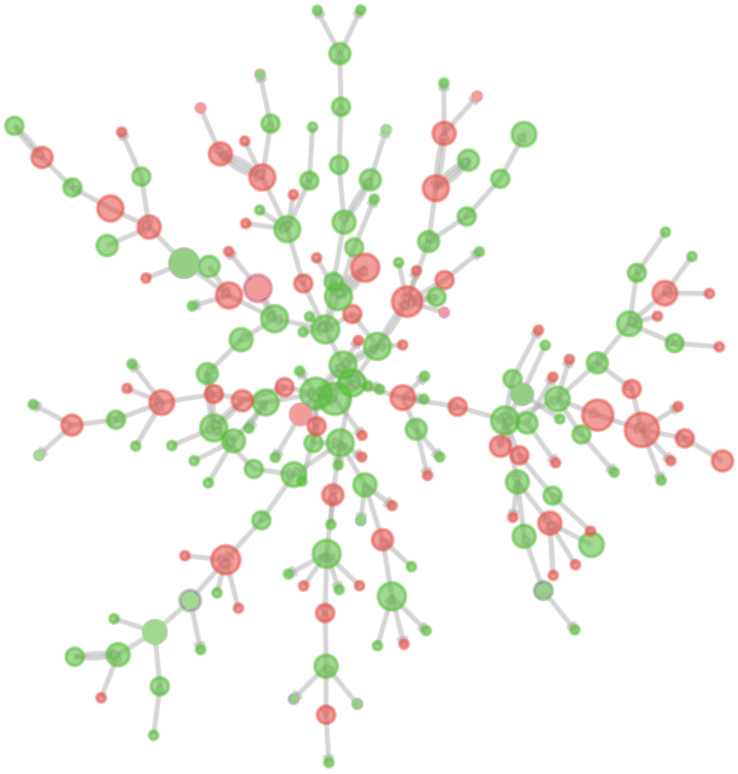

Figure 15: NY-NJ provider network