Healthcare Provider Fraud Detection

This data dictionary represents the author's best understanding of the data sets provided on Kaggle for health care classification analysis.

Margaret Bowers, 12/2/2025

Data source

Dataset:
https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?select=Test_Beneficiarydata-1542969243754.csv

Data Summary

Approximately one year (2009) of inpatient, outpatient, beneficiary, and provider data.

Inpatient and Outpatient claims data includes 30 and 27 features respectively, with the 3 additional inpatient features specific to tracking hospital admissions. This data describes the physician IDs, procedures, providers, reimbursements and deductibles associated with each beneficiary claim.

Beneficiary data includes 25 features that describe the personal characteristics and health conditions of each unique beneficiary as well as their annual reimbursement and deductible totals.

Provider data consists of the labeled fraud status for providers. This is not a measure of known fraud, but a measure of suspected fraud.

Data files

Data provided was presented in Train_ and Test_ sets. The Test_ data is the unlabeled data used for the Kaggle competition.

- Train_Beneficiarydata-1542865627584.csv
- Train_Inpatientdata-1542865627584.csv
- Train_Outpatientdata-1542865627584.csv
- Train-1542865627584.csv

- Test_Beneficiarydata-1542969243754.csv

- Test_Inpatientdata-1542969243754.csv
- Test_Outpatientdata-1542969243754.csv
- Test-1542969243754.csv

Data dictionaries

Some terms:

*The **provider** (hospital) is who bills for the service*
*The **beneficiary** (patient) is who receives the service*

Beneficiary data

DESCRIPTIVE SUMMARY:
This data contains details of beneficiaries including their unique identifier, date of birth and death (if applicable), underlying health conditions, gender, race, region they belong to, reimbursement and deductible data.

SIZE:
The training set contains 138,556 observations and 25 features.
The test set contains 63,968 observations and 25 features.

SOURCES:
Train_Beneficiarydata-1542865627584.csv
Test_Beneficiarydata-1542969243754.csv

VARIABLE DESCRIPTIONS:
The data contains 4 object datatypes (2 of which are dates) and 21 numerical datatypes.

BeneID (object) : unique patient/beneficiary identifier
BENExxxxx
DOB (object) : date of birth of beneficiary
Y-M-D
DOD (object) : date of death of beneficiary. NaN if not deceased
Y-M-D

Gender (int64) : gender code for beneficiary

1

2

Race (int64) : race code for beneficiary

1

2

3

4 *no instances in either train/test*

5

ESRD (end-stage renal disease) patients have higher health care costs and utilization. They are billed differently for different services (ex: dialysis, EPO meds). These patients likely have dramatically higher reimbursement amounts that can explain legitimate high-cost claims.

RenalDiseaseIndicator (object) : Flag indicating whether the beneficiary has a condition related to kidney failure or not.

0	No renal disease
Y	Has renal disease

State (int64) : state code (52 categories) for beneficiary

train: 1-54, *no values for '40' or '48'*

test: 1-54, *no values for '40' or '48'*

County (int64) : county code for beneficiary

NoOfMonths_PartA/BCov fields help identify partial-year enrollees vs full-year enrollees, affecting annual reimbursement totals.

NoOfMonths_PartACov (int64) : 13 categories

Number of months the beneficiary had Part A coverage (Hospital Insurance). Part A covers **inpatient** hospital stays, skilled nursing facility care, hospice care, and some home health care.

0-12

NoOfMonths_PartBCov (int64) : 13 categories

Number of months the beneficiary had Part B coverage (Medical Insurance). Part B covers doctor visits, **outpatient** care, preventive services, and medical equipment.

0-12

Chronic Condition columns indicate whether a beneficiary has a specific disease condition. Columns identify whether the beneficiary has chronic conditions related to

Alzheimer's, heart failure, kidney disease, cancer, obstructive pulmonary, depression, diabetes, ischemic heart disease, osteoporosis, rheumatoid arthritis, and or stroke. Chronic condition columns each contain 2 categories. The meaning of 1 and 2 is assigned based on EDA.

1 - has condition
2 - does not have condition

ChronicCond_Alzheimer (int64) : 2 categories
ChronicCond_Heartfailure (int64) : 2 categories
ChronicCond_KidneyDisease (int64) : 2 categories
ChronicCond_Cancer (int64) : 2 categories
ChronicCond_ObstrPulmonary (int64) : 2 categories
ChronicCond_Depression (int64) : 2 categories
ChronicCond_Diabetes (int64) : 2 categories
ChronicCond_IschemicHeart (int64) : 2 categories
ChronicCond_Osteoporasis (int64) : 2 categories
ChronicCond_rheumatoidarthritis (int64) : 2 categories
ChronicCond_stroke (int64) : 2 categories

Reimbursement: what Medicare/insurer (?) paid to providers
Deductible: What the patient paid out-of-pocket before coverage kicked in

IPAnnualReimbursementAmt (int64) : continuous amount

Total amount reimbursed for inpatient hospital services during the year

IPAnnualDeductibleAmt (int64) : continuous amount

Total deductibles the beneficiary paid for inpatient services during the year

OPAnnualReimbursementAmt (int64) : continuous amount

Total amount reimbursed for outpatient services (doctor visits, tests, procedures done without hospital admission) during the year

OPAnnualDeductibleAmt (int64) : continuous amount

Total deductibles the beneficiary paid for outpatient services during the year

Outpatient data

DESCRIPTIVE SUMMARY:

This data provides details about the claims filed for those patients who visit hospitals but are not admitted.

SIZE:
The training set contains 517,737 observations and 27 features.
The test set contains 125,841 observations and 27 features.

SOURCES:
Train_Outpatientdata-1542865627584.csv
Test_Outpatientdata-1542969243754.csv

VARIABLE DESCRIPTIONS:
The data contains 19 object datatypes (2 of which are dates) and 8 numerical datatypes.

BeneID (object) : unique patient/beneficiary identifier
BENExxxxx
ClaimID (object) : unique claim identifier

Claim start and end dates reflect the service period, not when the claim was filed or processed with the span between these dates determining the episode of care. They can be used to calculate the length of stay (for inpatient) or treatment duration (for outpatient). Multiple claims can have overlapping date ranges if a patient received different types of services during the same period
ClaimStartDt (object) : date when claim service began.
Y-M-D
ClaimEndDt (object) : date when claim service ended.
Y-M-D

Provider (object) : unique provider (hospital/institution or individual physician) identifier.
PRVxxxxx

AttendingPhysician (object) : unique identifier for attending physician who has primary responsibility for the patient's overall care during the encounter. In the inpatient setting, this is typically the doctor who admits the patient and manages their case throughout the hospitalization.
PHYxxxxxx

OperatingPhysician (object) : unique identifier for operating physician who performs the primary surgical procedure. This field is presumed only populated when there's a surgical procedure involved.

PHYxxxxxx

OtherPhysician (object) : unique identifier for physician providing significant services during the encounter but who aren't the attending or operating physician. This could include consulting physicians, specialists brought in for specific issues, assistant surgeons, or physicians who perform secondary procedures. Multiple "other physicians" can be involved in a single case.

PHYxxxxxx

Diagnosis codes identify the medical conditions, diseases, injuries, or symptoms associated with the patient's claim. ClmDiagnosisCode_1 represents the principal diagnosis and drives the payment logic (-Claude) for the encounter. Other codes represent secondary diagnosis.

ClmDiagnosisCode_1 (object) : primary condition diagnosis code. This is the main condition, *determined after the encounter/study*, which is determined to be the reason for treatment.

ClmDiagnosisCode_2 (object) :

ClmDiagnosisCode_3 (object) :

ClmDiagnosisCode_4 (object) :

ClmDiagnosisCode_5 (object) :

ClmDiagnosisCode_6 (object) :

ClmDiagnosisCode_7 (object) :

ClmDiagnosisCode_8 (object) :

ClmDiagnosisCode_9 (object) :

ClmDiagnosisCode_10 (object) :

Claim Procedure codes identify treatments, surgeries, tests, or services performed

ClmProcedureCode_1 (float64) :

ClmProcedureCode_2 (float64) :

ClmProcedureCode_3 (float64) :

ClmProcedureCode_4 (float64) :

ClmProcedureCode_5 (float64) :

ClmProcedureCode_6 (float64) :

ClmAdmitDiagnosisCode (object) : diagnosis code that reflects the patient's condition or reason for admission *at the time they entered the hospital*. This field is primarily used in inpatient claims and may not always be populated in outpatient datasets.

InscClaimAmtReimbursed (int64) : represents the actual amount that the insurer paid to the provider for the services rendered on that claim. The beneficiary may still owe additional amounts (deductible, coinsurance, or copayments) on top of what the insurer reimbursed. So this field specifically captures the insurer's portion of the payment, not the total cost of care or what the patient might owe.
DeductibleAmtPaid (int64) : Represents the portion of the claim that was applied toward the beneficiary's deductible. This is the amount from this specific claim that counted toward the patient's deductible. This is money the patient owes, not what the insurer paid. Useful for understanding beneficiary cost burden. The patient is billed for this amount directly.

Inpatient data

DESCRIPTIVE SUMMARY:
This data provides insights about the claims filed for those patients who are admitted in the hospitals. It also provides additional details like their admission and discharge dates and admit diagnosis codes.

SIZE:
The training set contains 40,474 observations and 30 features.
The test set contains 9551 observations and 30 features.

SOURCES:
Train_Inpatientdata-1542865627584.csv
Test_Inpatientdata-1542969243754.csv

VARIABLE DESCRIPTIONS:
The data contains 22 object datatypes (4 of which are dates) and 8 numerical datatypes. It is structured similarly to the Outpatient datasets, with the exception of 3 additional columns: **AdmissionDt, DischargeDt,** and **DiagnosisGroupCode**

AdmissionDt (object) : Admission date. Date an inpatient was admitted to the hospital or facility

Y-M-D

DischargeDt (object) : Discharge date. Date an inpatient was discharged from the hospital or facility

Y-M-D

DiagnosisGroupCode (object) : A classification system that groups inpatient stays into categories for payment purposes. Each diagnosis related group (DRG) represents a clinically similar group of patients with similar resource consumption. Examples: DRG 470 = "Major hip and knee joint replacement", DRG 291 = "Heart failure and shock". The insurer calculates a fixed payment amount for each DRG with payment is based on the principal diagnosis, procedures performed, complications, and patient characteristics. Hospitals receive the DRG payment regardless of actual costs (incentivizing efficiency). More complex/severe cases are assigned DRGs with higher payment rates. Primary driver of Medicare payment amounts. *Is this a good measure of hospital efficiency then?*

Features assumed identical to those in Outpatient dataset:

BeneID
ClaimID
ClaimStartDt
ClaimEndDt
Provider
InscClaimAmtReimbursed
AttendingPhysician
OperatingPhysician
OtherPhysician

ClmAdmitDiagnosisCode
DeductibleAmtPaid

ClmDiagnosisCode_1
ClmDiagnosisCode_2
ClmDiagnosisCode_3
ClmDiagnosisCode_4
ClmDiagnosisCode_5
ClmDiagnosisCode_6
ClmDiagnosisCode_7

ClmDiagnosisCode_8
ClmDiagnosisCode_9
ClmDiagnosisCode_10

ClmProcedureCode_1
ClmProcedureCode_2
ClmProcedureCode_3
ClmProcedureCode_4
ClmProcedureCode_5
ClmProcedureCode_6

Provider data

DESCRIPTIVE SUMMARY:
This data contains “the mapping of Provider’s unique id and the class label signifying whether the Provider is fraud or not. For the Test data, only the Provider’s unique id is given and the class label has to be found.” 1.

SIZE:
The training set contains 5410 observations and 2 features.
The test set contains 1353 observations and 1 feature.

SOURCES:
Train-1542865627584.csv
Test-1542969243754.csv

VARIABLE DESCRIPTIONS:
The training set contains 2 object datatypes. The test set contains 1 object datatype.

Provider (object) : unique identifier for Provider (*Hospital/Institution, physician?)*
PotentialFraud (object) : class label identifying Provider as fraud or not fraud. Training set only.
Yes
No