Rotman

Master of Management Analytics

Medical Insurance Fraud Investigation—Medicare

Criteria for Fraud Detection

October 22, 2020

Prepared by: Aichen Liu
Colleen Liu
Rita Lyu
William Chien

Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Overview

- Introduction
  - Objectives and background
  - Data
- Key Variables and Correlation
- Model Selection and Implication
- Recommendation
- Appendix
  - Appendix 1
  - Appendix 2
  - Appendix 3
  - Appendix 4
  - Appendix 5
  - Appendix 6

Here's where it changes.

Rotman

Introduction

Rotman School of Management
UNIVERSITY OF TORONTO

Introduction

Rotman

Objectives and Background

- Insurance fraud is usually an attempt to exploit an insurance contract for financial gain. Specialists found that physicians and providers associated with Medicare have involved in many cases of fraud.
- This exploratory research focuses on fraud investigations and recommendations on how to combat Medicare fraud.
- Medicare is the federal health insurance program for people who are 65 or older. Medicare has four parts, but in this research, we focus on Part A and Part B.
- Part A: Hospital insurance covers inpatient hospital stays, skilled nursing care, hospice care
- Part B: Doctor and outpatient services, covers doctor visits, lab tests, diagnostic screenings

Introduction

Data

- The data contains over fifty thousand observations and sixty variables, which involves 5410 providers and 506 of them are flagged as Fraud.

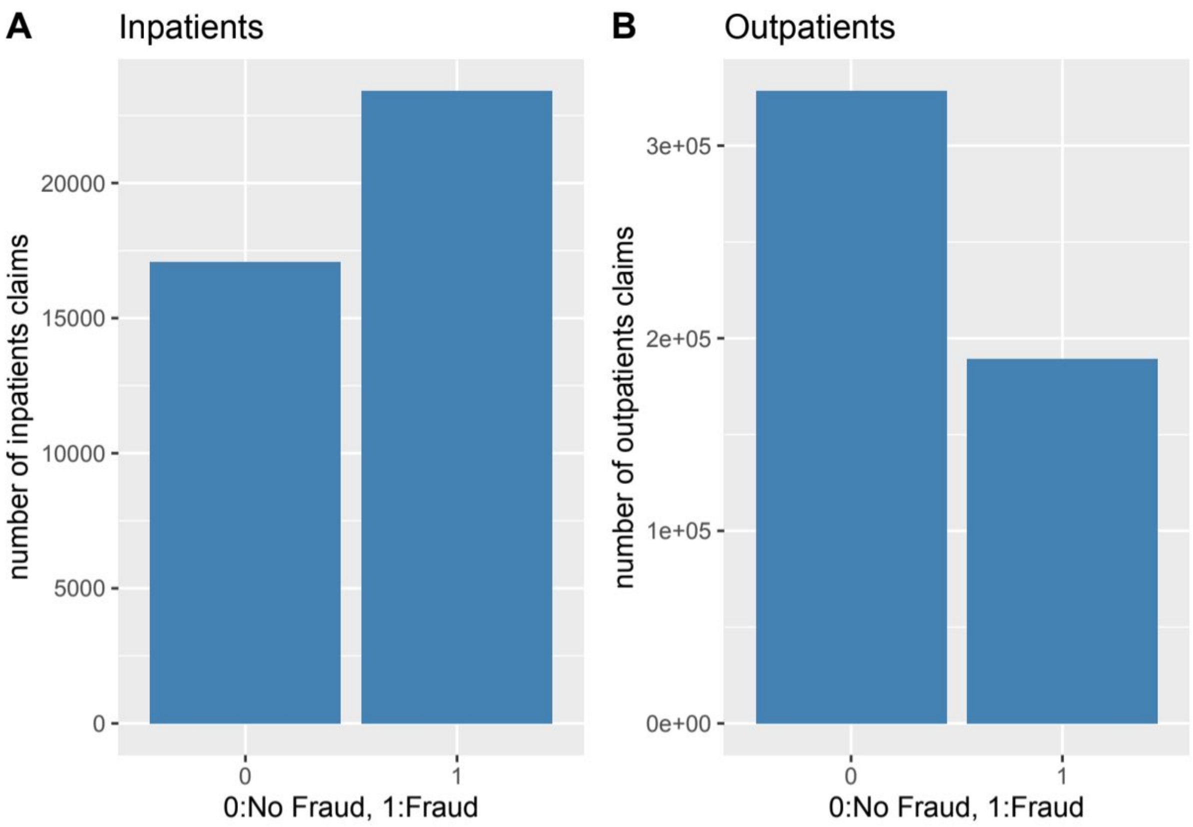
- Key variables considered:
    - ❖ Provider ID
    - ❖ Claims ID
    - ❖ Beneficiary ID
    - ❖ Reimbursed Amount each claim
    - ❖ Number of inpatients claims each provider made between Dec 2008 and Dec 2009
    - ❖ Number of outpatients claims each provider mad between Dec 2008 and Dec 2009
    - ❖ Diagnosis group code
    - ❖ Beneficiary’s gender
    - ❖ Beneficiary’s race
    - ❖ Types of chronic disease beneficiary have(such as Cancer, Alzheimer)
    - ❖ Whether beneficiary have Renal disease
    - ❖ Duration in hospital
    - ❖ Whether beneficiary is died
    - ❖ Whether beneficiary have surgery

Rotman

Key Variables and Correlation

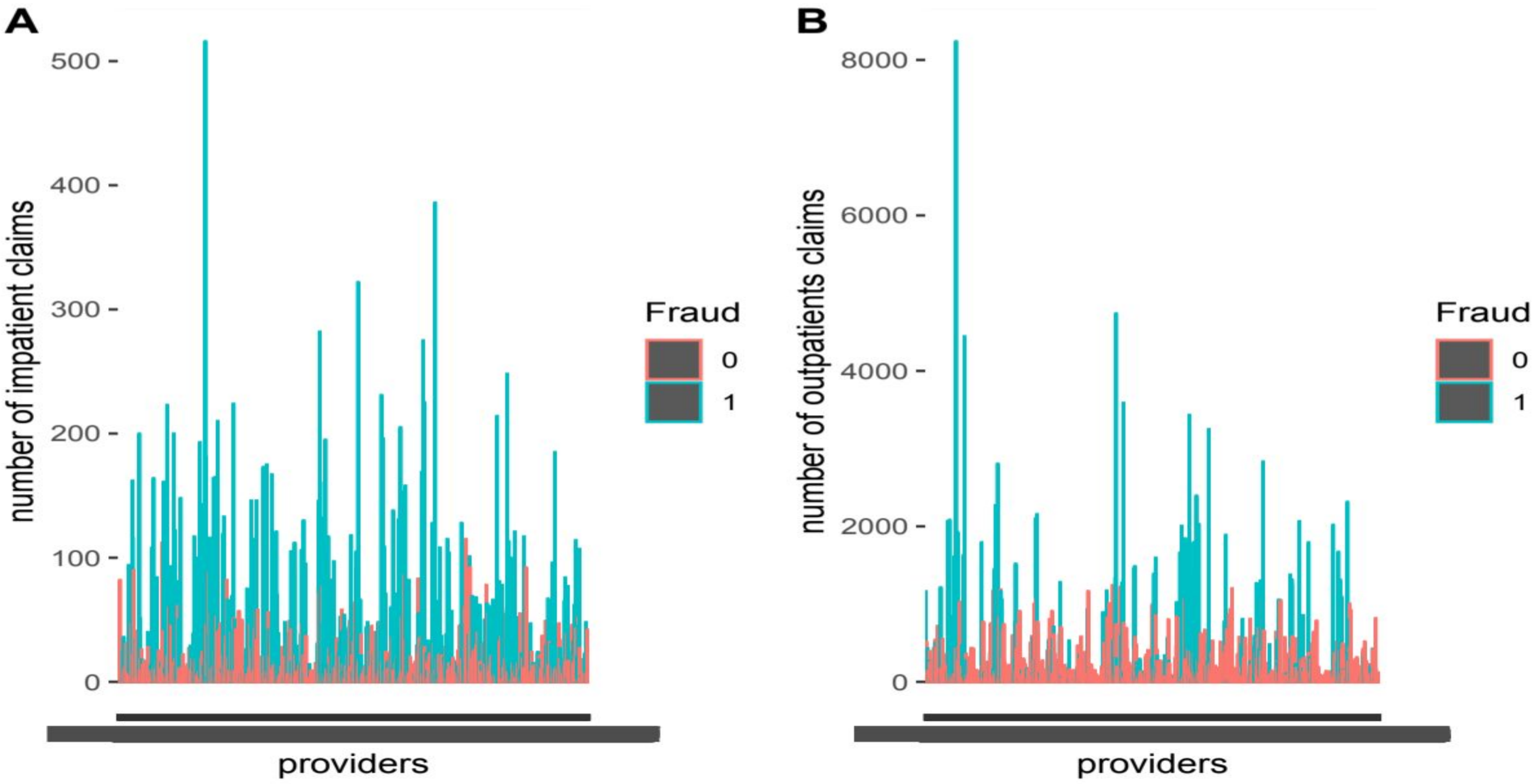
Rotman School of Management
UNIVERSITY OF TORONTO

Key Variables and Their Correlation



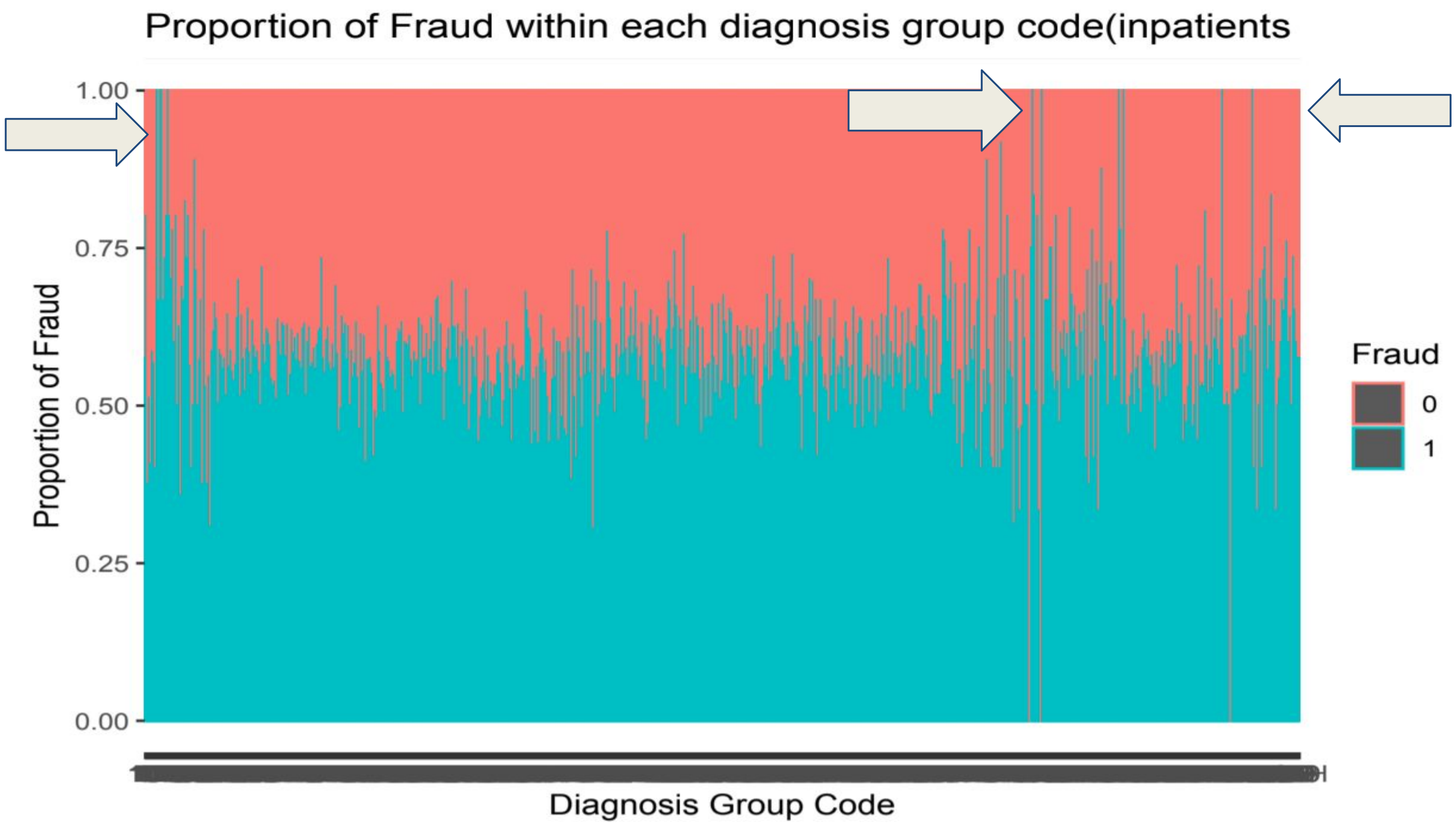
By looking at these bar charts, we can see that the proportion of fraud is higher for inpatients claims than for outpatients claims.

Key Variables and Their Correlation



These plots indicate the number of claims each providers made. Blue bars represent potential fraud providers and red bars represent normal providers. As indicated, providers who are potential fraud are more likely to make claims than non-flagged provides.

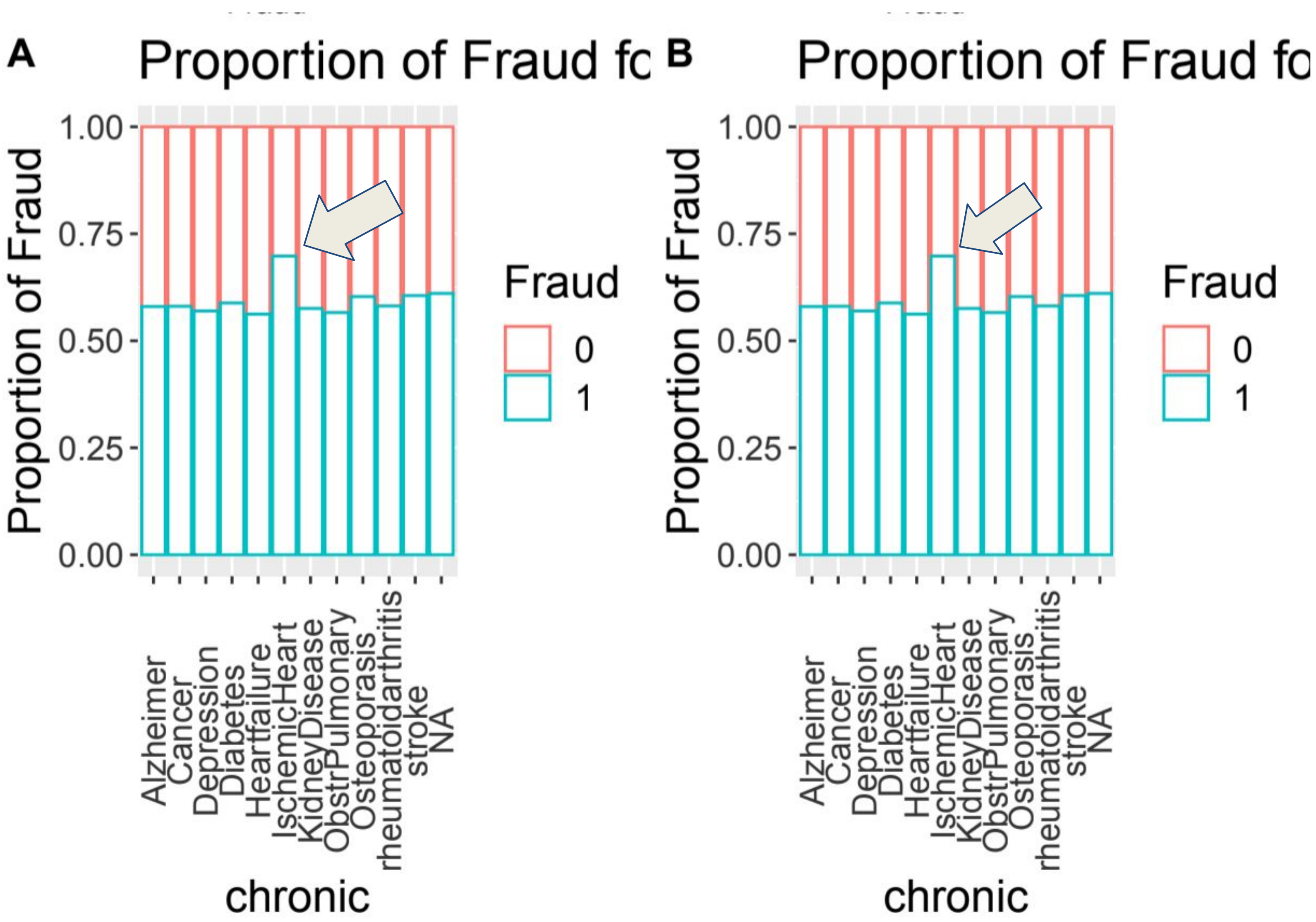
Key Variables and Their Correlation



[1] 113 115 116 124 768 776 843 846 927 955

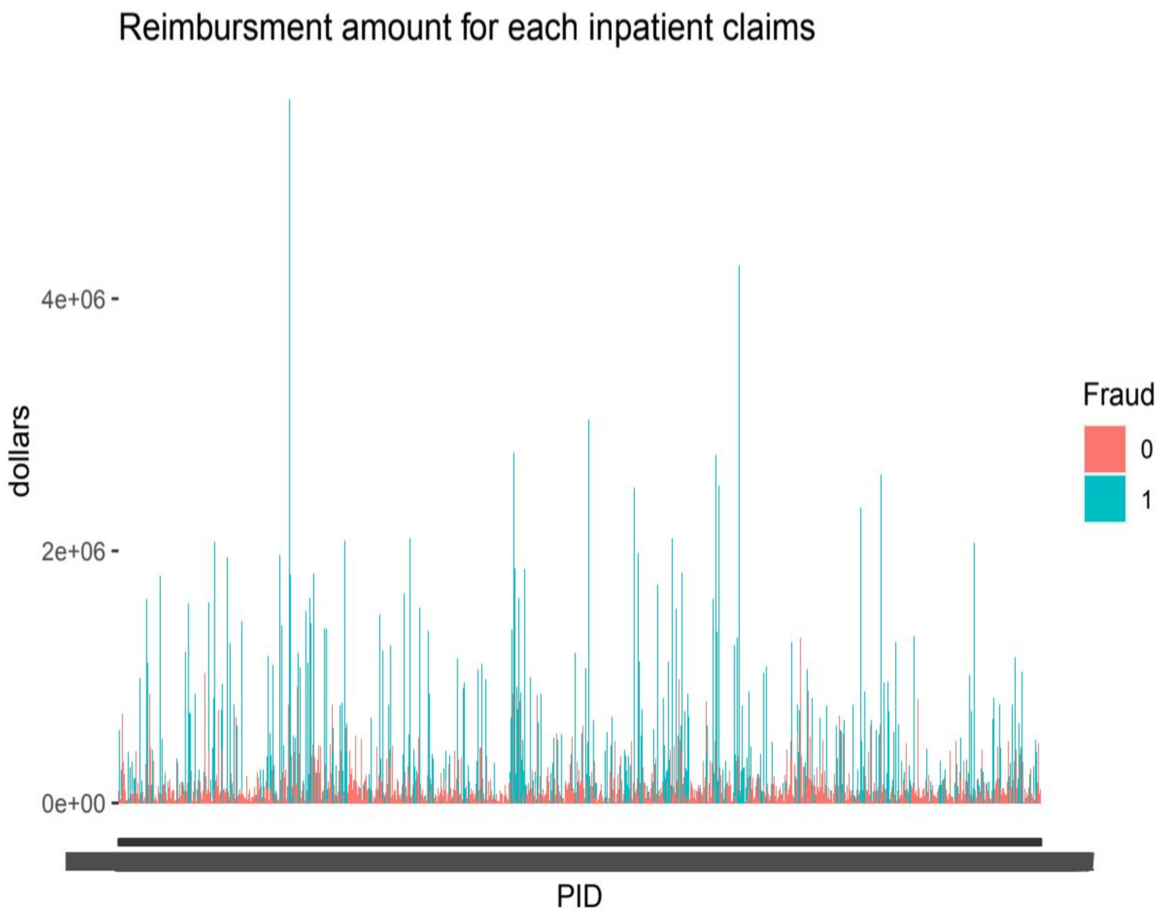
This plot groups claims by diagnosis groups. The x-axis represents each diagnosis group. The length of the blue and red bars represent the proportion of fraud claims and the proportion of the normal claims respectively. There are some diagnosis groups which their bars are all blue. Those are the groups contain only fraud claims(Arrows). They are: 113, 115, 116, 124, 768, 776, 843, 846, 927, 955. These groups should be paid more attention.

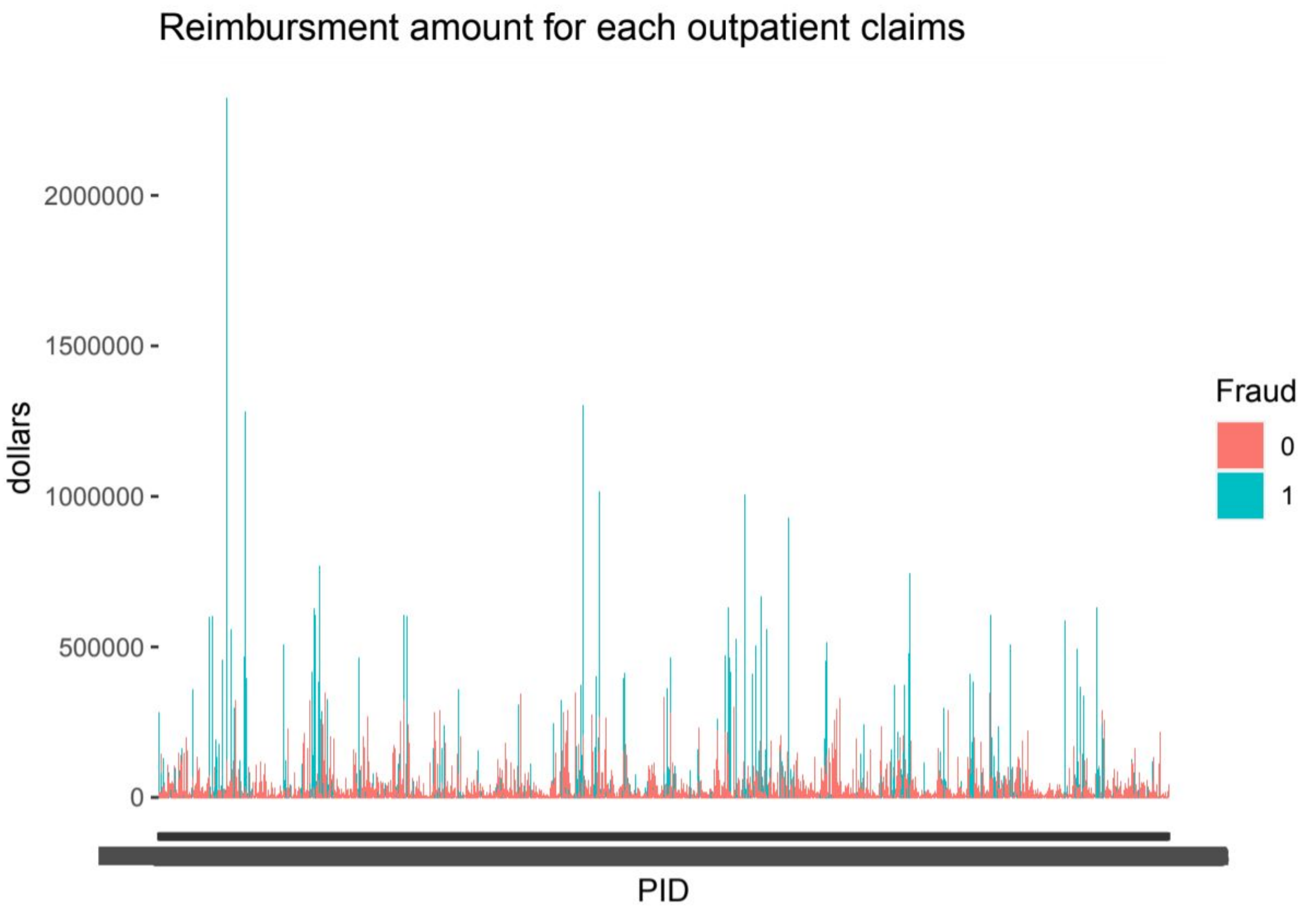
Key Variables and Their Correlation



Similar as above, the x-axis of this plot represent the types of chronic disease. It is obvious that ‘ischemicheart’ has the highest proportion of fraud claims. Therefore, claims for ischemic heart should be paid more attention.

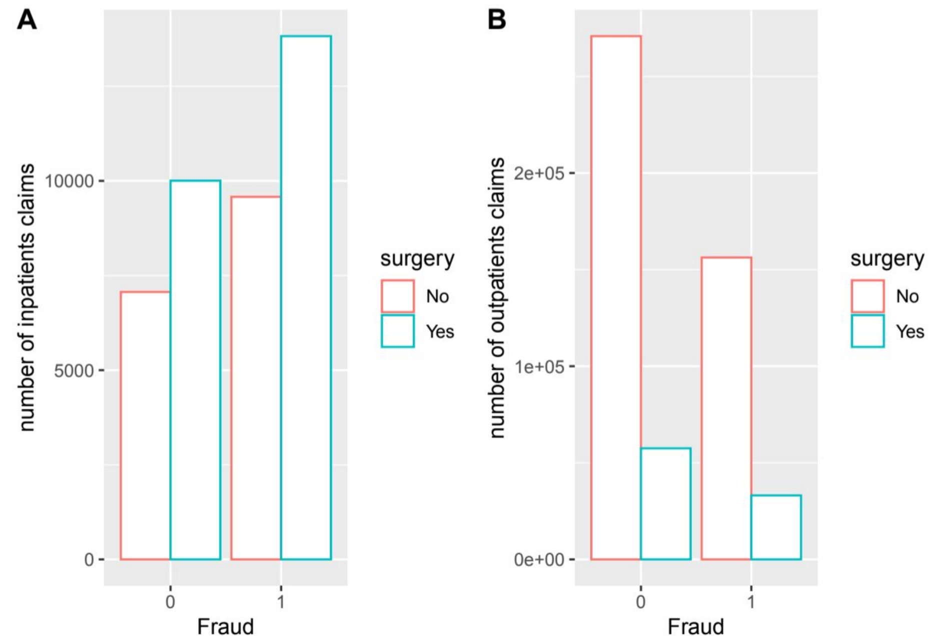
Key Variables and Their Correlation

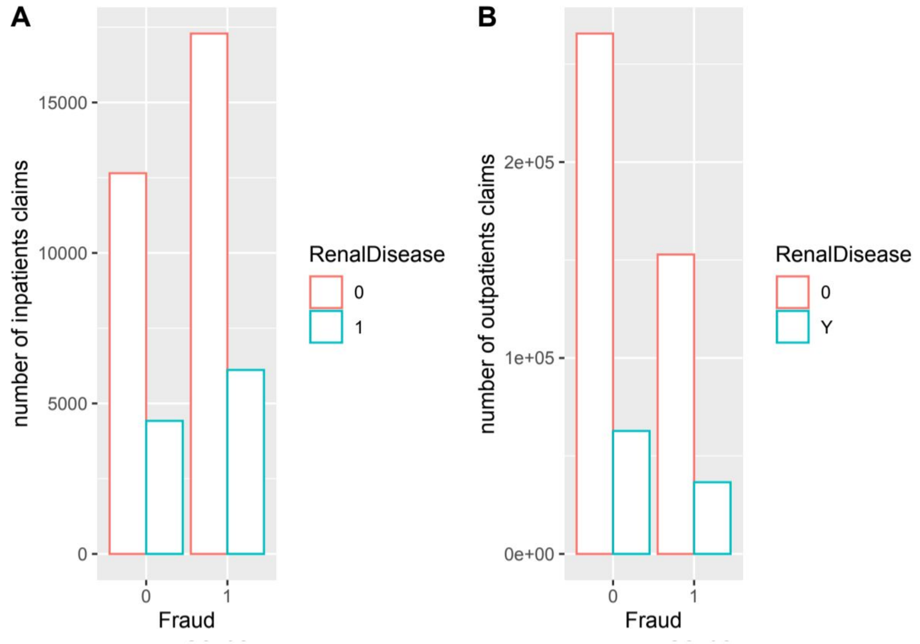




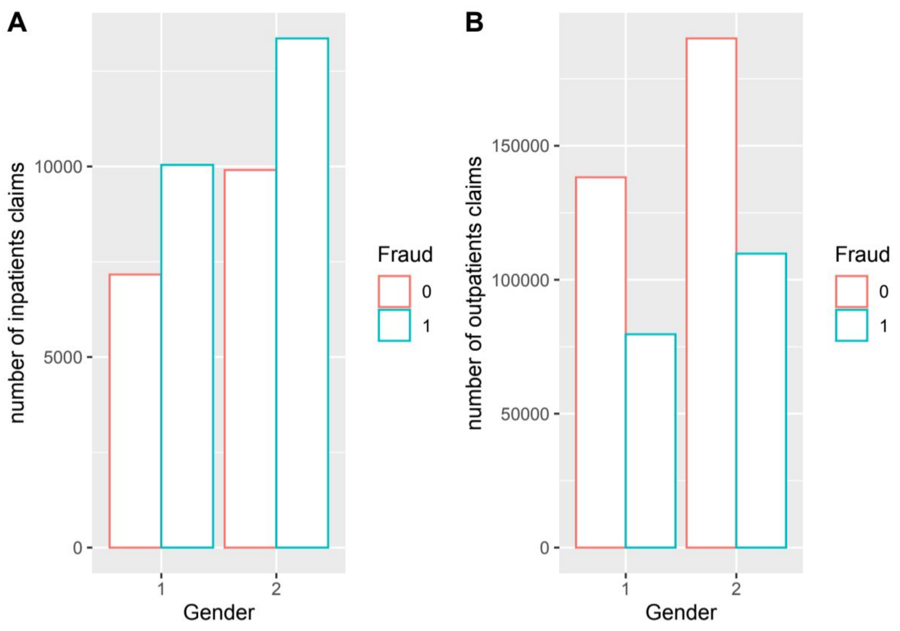
These plots indicate the reimbursed amount for one selected claim for each provider. Blue bars represent potential fraud providers and red bars represent normal providers. As indicated, providers who are potential fraud tend to make larger amount claims than normal providers.

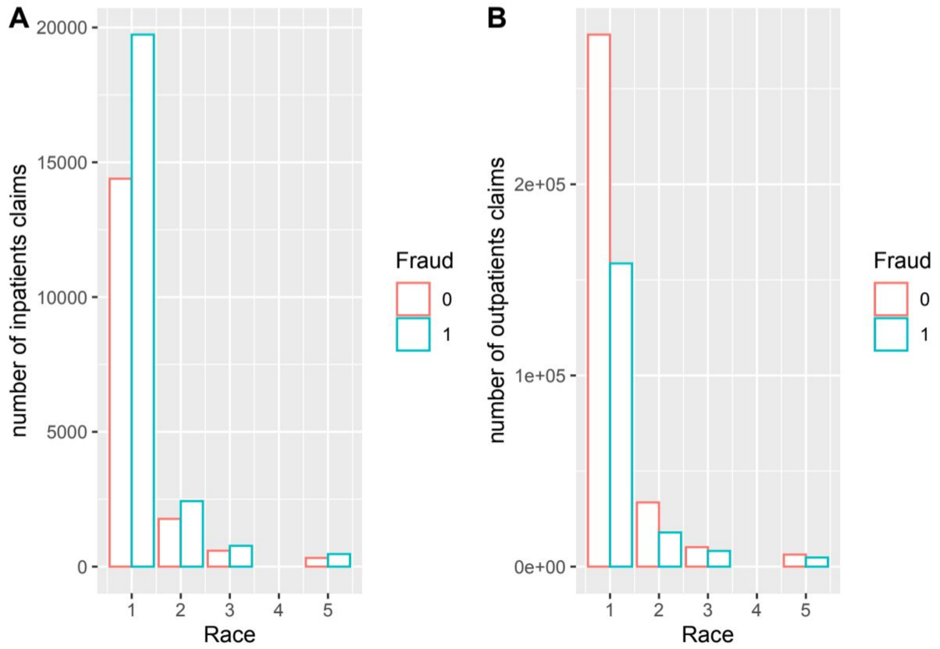
Key Variables and Their Correlation





The relationship between fraud and gender, race, whether have renal disease, or whether have surgery are subtle. It is not very evident from the graphs.





Overlapping inpatients and outpatients period

```sql
# BID who has overlapping inpatient and outpatient period
SELECT DISTINCT
    a.BID, a.CID, a.StartDt AS StartDt_inpa, a.EndDt AS EndDt_inpa, a.PID as PID_inpa,
    b.PID as PID_outpa, b.BID, b.CID, b.StartDt AS StartDt_outpa, b.EndDt AS EndDt_outpa
FROM inpatients a JOIN outpatients b
ON a.BID = b.BID
AND b.StartDt < a.EndDt
AND a.StartDt < b.EndDt
AND a.StartDt < b.StartDt
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

BID	CID	StartDt_inpa	EndDt_inpa	PID_inpa	PID_outpa	BID	CID	StartDt_outpa	EndDt_outpa
BENE100075	CLM31820	2008-12-31	2009-01-07	PRV52723	PRV52648	BENE100075	CLM176569	2009-01-02	2009-01-02
BENE100075	CLM31820	2008-12-31	2009-01-07	PRV52723	PRV52572	BENE100075	CLM211583	2009-01-02	2009-01-03
BENE100075	CLM31820	2008-12-31	2009-01-07	PRV52723	PRV52723	BENE100075	CLM381108	2009-01-05	2009-01-05
BENE100075	CLM31820	2008-12-31	2009-01-07	PRV52723	PRV52723	BENE100075	CLM402857	2009-01-06	2009-01-06
BENE100075	CLM31820	2008-12-31	2009-01-07	PRV52723	PRV52723	BENE100075	CLM432258	2009-01-06	2009-01-06
BENE100075	CLM31820	2008-12-31	2009-01-07	PRV52723	PRV57530	BENE100075	CLM691510	2009-01-01	2009-01-01
BENE100422	CLM31613	2008-12-30	2009-01-08	PRV52065	PRV54888	BENE100422	CLM236434	2009-01-03	2009-01-03
BENE100422	CLM31613	2008-12-30	2009-01-08	PRV52065	PRV52159	BENE100422	CLM400040	2009-01-06	2009-01-06
BENE100590	CLM31743	2008-12-31	2009-01-03	PRV53266	PRV53242	BENE100590	CLM217961	2009-01-02	2009-01-02
BENE100590	CLM31743	2008-12-31	2009-01-03	PRV53266	PRV53295	BENE100590	CLM717844	2009-01-02	2009-01-02
BENE100614	CLM31969	2009-01-01	2009-01-18	PRV56461	PRV56442	BENE100614	CLM174744	2009-01-02	2009-01-02
BENE100614	CLM31969	2009-01-01	2009-01-18	PRV56461	PRV56442	BENE100614	CLM257659	2009-01-03	2009-01-03
BENE100614	CLM31969	2009-01-01	2009-01-18	PRV56461	PRV56425	BENE100614	CLM276164	2009-01-03	2009-01-03
BENE100881	CLM32715	2009-01-06	2009-01-17	PRV56416	PRV56448	BENE100881	CLM605410	2009-01-09	2009-01-09

BENE100075 was in hospital from 2008-12-31 to 2009-01-07, but in the same period, BID was used by 6 different providers to claim outpatient reimbursement.

Similar potential fraud claims can be found for other BID.

Model Selection and Implication

Since Fraud can take two value 1(fraud) and 0(no fraud), we use logistic model in this case.

The first equation gives that Yi follows Binomial distribution where Yi is the number of fraud claims, Ni is the number of students in group i. μi is the estimated probability that the ith claim is a fraud.

In the second equation,logit(p) represents the log of odds of claim being a fraud.

$$Y_i \sim Binomial(N_i, \mu_i),$$

$$logit(p) = \beta_0 + \beta_1 I_{IschemicHeart} + \beta_2 X_{AmtReimbursed} + \beta_3 X_{numofclaims} + \beta_4 I_{RenalDisease} + \beta_5 I_{surgery}$$

Two models are fitted for inpatients and outpatients claims

Inpatient claim

	est	0.5 %	99.5 %
(Intercept)	0.143	0.125	0.165
Chronic_IschemicHeart	1.020	0.925	1.125
AmtReimbursed	1.000	1.000	1.000
num_claims	1.040	1.038	1.041
RenalDisease1	0.998	0.922	1.079
surgeryYes	1.003	0.933	1.078

Outpatient claim

	est	0.5 %	99.5 %
(Intercept)	0.083	0.080	0.086
Chronic_IschemicHeart	1.004	0.980	1.028
AmtReimbursed	1.000	1.000	1.000
num_claims	1.003	1.003	1.003
RenalDiseaseY	1.021	0.995	1.048
surgeryYes	0.988	0.962	1.015

The first column is the estimated value of the odds of being fraud. The second and last columns give the confidence interval for each subgroup. (On a natural scale)

As indicated, since confidence interval for reimbursement amount and number of claims don't contain 1, Therefore, they are significantly affecting whether the provider is fraud.

Rotman

Recommendation

Recommendations

Rotman

Based on the analysis, our team proposes to audit the providers who submitted claims include …

Category	Criteria	Remark
Diagnosis Group Code	**113, 115, 116**	• Research suggests that these codes does not exist • Most likely these claims are fake
	124, 768, 776, 843, 846, 927 and 955	• All providers who submit claims in these codes are all flagged
Beneficiary's chronic disease	**ischemic heart**	• Contains the highest proportion of fraud claims among all the chronic disease
Number of claims submitted	**over 200 impatient claims**	• From the model, this factor significantly drives the outcome of the fraud test • From the graphs, fraud providers exhibit the features that submit higher volume of claims
	over 2000 outpatient claims	
Reimbursement of a claim	**over $500,000 for a outpatient claim**	• From the model, this factor significantly drives the outcome of the fraud test • From the graphs, the reimbursement amount of claims submitted by fraud providers is significantly higher
	over $2,000,000 for an inpatient claim	

https://www.aapc.com/codes/icd9-codes/955

Recommendations Cont.

Rotman

Based on the analysis, our team proposes to audit the providers who submitted claims include …

Category	Criteria	Remark
Overlapping inpatient and outpatient periods	**over the same period, a BID is associate with one inpatient claim and multiple outpatient claims**	• Outpatient claims made by other provides are likely to be the fraud claims.
Frequency of claiming same diagnosis code	**over 7 times of the common diagnosis code**	• Fraud claims usually happen with high reimbursement in a short period. • If a provider claims the same diagnosis code within a given period, it is considered a potential fraud. • Common diagnosis code include 486, v5789, 0389, 41401, 49121, etc., which are frequent among all providers. Therefore can set a higher limit for fraud.
	over 3 times of the un-common diagnosis code	

Rotman

Master of Management Analytics

Thank you for watching

October 22, 2020

Prepared by: Aichen Liu
Colleen Liu
Rita Lyu
William Chien

Rotman School of Management
UNIVERSITY OF TORONTO

Rotman

Appendix

Appendix 1

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(cowplot)
inpatients = read.csv("merged_data_in.csv")
outpatients = read.csv("merged_data_out.csv")
inpatients$Fraud = as.factor(inpatients$Fraud)
inpatients = inpatients %>% mutate(death = case_when(
  is.na(inpatients$DOD) ~ "No"))
inpatients$death[is.na(inpatients$death)] <- "Yes"
outpatients = outpatients %>% mutate(death = case_when(
  is.na(outpatients$DOD) ~ "No"))
outpatients$death[is.na(outpatients$death)] <- "Yes"
outpatients$Fraud[outpatients$Fraud=="No"] <- 0
outpatients$Fraud[outpatients$Fraud=="Yes"] <- 1
#outpatients$RenalDisease[outpatients$RenalDisease=="Y"] <- "1"
inpatients = inpatients %>% mutate(chronic = case_when(
   Chronic_Alzheimer ==2 ~ "Alzheimer",
   Chronic_Heartfailure ==2~"Heartfailure",
   Chronic_KidneyDisease==2~"KidneyDisease",
   Chronic_Cancer==2~"Cancer",
   Chronic_ObstrPulmonary==2~"ObstrPulmonary",
   Chronic_Depression==2~"Depression",
   Chronic_Diabetes==2~"Diabetes",
   Chronic_IschemicHeart==2~"IschemicHeart",
   Chronic_Osteoporasis==2~"Osteoporasis",
   Chronic_rheumatoidarthritis==2~"rheumatoidarthritis",
   Chronic_stroke==2~"stroke"

inpatients = inpatients %>% mutate(chronic = case_when(
   Chronic_Alzheimer ==2 ~ "Alzheimer",
   Chronic_Heartfailure ==2~"Heartfailure",
   Chronic_KidneyDisease==2~"KidneyDisease",
   Chronic_Cancer==2~"Cancer",
   Chronic_ObstrPulmonary==2~"ObstrPulmonary",
   Chronic_Depression==2~"Depression",
   Chronic_Diabetes==2~"Diabetes",
   Chronic_IschemicHeart==2~"IschemicHeart",
   Chronic_Osteoporasis==2~"Osteoporasis",
   Chronic_rheumatoidarthritis==2~"rheumatoidarthritis",
   Chronic_stroke==2~"stroke"
    ))

outpatients = outpatients %>% mutate(chronic = case_when(
   Chronic_Alzheimer ==2 ~ "Alzheimer",
   Chronic_Heartfailure ==2~"Heartfailure",
   Chronic_KidneyDisease==2~"KidneyDisease",
   Chronic_Cancer==2~"Cancer",
   Chronic_ObstrPulmonary==2~"ObstrPulmonary",
   Chronic_Depression==2~"Depression",
   Chronic_Diabetes==2~"Diabetes",
   Chronic_IschemicHeart==2~"IschemicHeart",
   Chronic_Osteoporasis==2~"Osteoporasis",
   Chronic_rheumatoidarthritis==2~"rheumatoidarthritis",
   Chronic_stroke==2~"stroke"
    ))
```
```

Appendix 2

```
inpatients = inpatients %>% mutate(death = case_when(
  DOD !=NA ~ 1,
   ))
outpatients = outpatients %>% mutate(death = case_when(
  DOD !=NA ~ 1,
   ))
mma1 =inpatients %>% ggplot(aes(x=Fraud)) +
geom_bar(fill="steelblue")+xlab("0:No Fraud, 1:Fraud") +
ylab("number of inpatients claims")+ggtitle("Inpatients")
mma2=outpatients %>% ggplot(aes(x=Fraud)) +
geom_bar(fill="steelblue")+ggtitle("Outpatients")+xlab("0:No
Fraud, 1:Fraud") + ylab("number of outpatients claims")
plot_grid(mma1, mma2, labels = "AUTO")

mma3=ggplot(inpatients, aes(x=PID,color=Fraud)) +
geom_bar()+ xlab("providers") + ylab("number of impatient
claims")

mma4=ggplot(outpatients, aes(x=PID,color=Fraud)) +
geom_bar()+ xlab("providers") + ylab("number of outpatients
claims")
plot_grid(mma3, mma4, labels = "AUTO")
c = data.frame(inpatients %>% select(PID) %>%
table(useNA="always"))
names(c)[1] <- "PID"
inpatients = merge(x = inpatients, y = c, by = "PID", all.x =
TRUE)
names(inpatients)[63] <- "num_claims"
d = data.frame(outpatients %>% select(PID) %>%
table(useNA="always"))
names(d)[1] <- "PID"
outpatients = merge(x = outpatients, y = d, by = "PID", all.x
= TRUE)
names(outpatients)[56] <- "num_claims"
ggplot(inpatients,
aes(x=DiagnosisGroupCode,color=Fraud)) +
geom_bar(position = "Fill")+xlab("Diagnosis Group Code") +
ylab("Proportion of Fraud")+ggtitle("Proportion of Fraud
within each diagnosis group code(inpatients")
a = inpatients%>% select(DiagnosisGroupCode) %>%
table(useNA="always")
a = data.frame(a)
sub_yes = filter(inpatients, Fraud == 1)
b = data.frame(sub_yes%>% select(DiagnosisGroupCode)
%>% table(useNA="always"))
Diagn_GroupCode = merge(x = a, y = b, by = ".", all.x =
TRUE)
Diagn_GroupCode = Diagn_GroupCode %>%
mutate(yes_p = Freq.y/Freq.x)
Diagn_GroupCode_F = filter(Diagn_GroupCode, yes_p ==
1)
Diagn_GroupCode_F$.
inpatients$Gender = as.factor(inpatients$Gender)
mma5=ggplot(inpatients, aes(x=Gender,color=Fraud)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ ylab("number of inpatients claims")
```

Appendix 3

```
inpatients$Gender = as.factor(inpatients$Gender)
mma5=ggplot(inpatients, aes(x=Gender,color=Fraud)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ ylab("number of inpatients claims")
outpatients$Gender = as.factor(outpatients$Gender)
mma6=ggplot(outpatients, aes(x=Gender,color=Fraud)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ ylab("number of outpatients claims")
plot_grid(mma5, mma6, labels = "AUTO")
mma7=ggplot(inpatients, aes(x=Race,color=Fraud)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ylab("number of inpatients claims")
mma8=ggplot(outpatients, aes(x=Race,color=Fraud)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ylab("number of outpatients claims")
plot_grid(mma7, mma8, labels = "AUTO")
inpatients$RenalDisease = as.factor(inpatients$RenalDisease)
mma9 = ggplot(inpatients, aes(x=Fraud,color=RenalDisease)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ylab("number of inpatients claims")
outpatients$RenalDisease =
as.factor(outpatients$RenalDisease)
mma10=ggplot(outpatients, aes(x=Fraud,color=RenalDisease))
+ geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ylab("number of outpatients claims")
plot_grid(mma9, mma10, labels = "AUTO")
```

```
mma11 = ggplot(inpatients, aes(x=chronic,color=Fraud)) +
geom_bar(position = "fill",fill="white", )+theme(axis.text.x =
element_text(colour = "grey20", size = 12, angle = 90, hjust = 0.5, vjust =
0.5),
                    axis.text.y = element_text(colour = "grey20", size = 12),
        text = element_text(size = 16))+ylab("Proportion of
Fraud")+ggtitle("Proportion of Fraud for each chronic disease(inpatients")

```

```{r}
mma12=ggplot(inpatients, aes(x=chronic,color=Fraud)) +
geom_bar(position = "fill",fill="white", )+theme(axis.text.x =
element_text(colour = "grey20", size = 12, angle = 90, hjust = 0.5, vjust =
0.5),
                    axis.text.y = element_text(colour = "grey20", size = 12),
        text = element_text(size = 16))+ylab("Proportion of
Fraud")+ggtitle("Proportion of Fraud for each chronic
disease(outpatients")

plot_grid(mma11, mma12, labels = "AUTO")
ggplot(inpatients, aes(x=Fraud, y=Duration_hospital )) +
  geom_boxplot()+ ylab("number of days")+ggtitle("Number of days of
In-hospital care ")
ggplot(inpatients, aes(x = PID, y = AmtReimbursed,fill = Fraud)) +
       geom_bar(stat = "identity")+ ylab("dollars")+ggtitle("Reimbursment
amount for each inpatient claims")
ggplot(outpatients, aes(x = PID, y = AmtReimbursed,fill = Fraud)) +
       geom_bar(stat = "identity")+ ylab("dollars")+ggtitle("Reimbursment
amount for each outpatient claims")
```

Rotman

Appendix 4

```
inpatients = inpatients %>% mutate(surgery = case_when(
  is.na(inpatients$OperatingPhysician) ~ "No"))
inpatients$surgery[is.na(inpatients$surgery)] <- "Yes"

mma13=ggplot(inpatients, aes(x=Fraud,color=surgery)) +
geom_bar(position = position_dodge(preserve = "single"),fill="white") +
ylab("number of inpatients claims")

outpatients = outpatients %>% mutate(surgery = case_when(
  is.na(outpatients$OperatingPhysician) ~ "No"))
outpatients$surgery[is.na(outpatients$surgery)] <- "Yes"

mma14=ggplot(outpatients, aes(x=Fraud,color=surgery)) +
geom_bar(position = position_dodge(preserve =
"single"),fill="white")+ylab("number of outpatients claims")
plot_grid(mma13, mma14, labels = "AUTO")
```
```{r}
inpatients$death = as.factor(inpatients$death)
ggplot(inpatients, aes(x=death,color=Fraud)) + geom_bar(position =
"fill",fill="white") + ylab("proportion of Fraud") + ggtitle("Proportion of
Fraud VS. Death")
logi_inpatients = glm(Fraud ~
Chronic_IschemicHeart+AmtReimbursed+num_claims +
RenalDisease+surgery,
data=inpatients, family='binomial')
aaa = cbind(est=logi_inpatients$coef, confint(logi_inpatients,
level=0.99))
oddsMat = exp(aaa)
knitr::kable(oddsMat, digits=3)
outpatients$Fraud = as.factor(outpatients$Fraud)
logi_outpatients = glm(Fraud ~
Chronic_IschemicHeart+AmtReimbursed+num_claims +
RenalDisease+surgery,
data=outpatients, family='binomial')
#knitr::kable(logi_outpatients, digits=3,caption = "Figure 2")
#summary(logi_outpatients)
bbb = cbind(est=logi_outpatients$coef,
confint(logi_outpatients, level=0.99))
oddsMat = exp(bbb)
knitr::kable(oddsMat, digits=3)
$$ Y_i \sim {Binomial}(N_{i},\mu_{i}),$$
$$logit(p)=\beta_0
+\beta_1I_{IschemicHeart}+\beta_2X_{AmtReimbursed}+\b
eta_3X_{numof
claims}+\beta_4I_{RenalDisease}+\beta_5I_{surgery}$$
```

Rotman

Appendix 5. Sanity check on dates

	Inpatients	Outpatients
StartDt > EndDt (eg. StartDt = 2009-01-09 EndDt = 2009-01-01)	0	1,986
StartDt = EndDt	605	495,189
StartDt < EndDt	39,869	20,562

```
# Sanity check if EndDt is before StartDt
SELECT BID, CID, StartDt, EndDt
FROM outpatients
WHERE StartDt > EndDt
```

Result Grid Filter Rows: Edit:

BID	CID	StartDt	EndDt
BENE100036	CLM607135	2009-01-09	2009-01-01
BENE100075	CLM578054	2009-01-09	2009-01-01
BENE100209	CLM602028	2009-01-09	2009-01-01
BENE100235	CLM573690	2009-01-09	2009-01-01
BENE100292	CLM605404	2009-01-09	2009-01-01
BENE100305	CLM593516	2009-01-09	2009-01-01
BENE100305	CLM603685	2009-01-09	2009-01-01
BENE100398	CLM591000	2009-01-09	2009-01-01
BENE100564	CLM574601	2009-01-09	2009-01-01

- Majority of outpatients records are on the same day.
- Outpatients whose StartDt > EndDt, are all starting and ending on the same date.

Appendix 6. Frequent claimed diagnosis/procedure code by PID

```sql
# frequent diagnosis/procedure code
# Inpatients/outpatients Diagnosis 1-10
SELECT PID, DiagnosisCode_1, COUNT(DiagnosisCode_1)
FROM inpatients
GROUP BY PID, DiagnosisCode_1
ORDER BY COUNT(DiagnosisCode_1) DESC
;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch r

PID	DiagnosisCode_1	COUNT(DiagnosisCode_1)
PRV55462	486	22
PRV52019	486	20
PRV52019	V5789	19
PRV52019	49121	15
PRV52019	0389	14
PRV52019	41401	14
PRV52019	42731	14
PRV55462	V5789	13
PRV55462	41071	13
PRV55209	486	13
PRV55209	0389	13
PRV54367	486	12
PRV57191	41401	12
PRV54367	V5789	11
PRV52120	42731	11
PRV51501	486	11