# HW3: Insurance Claims Data & Analytics

*Team 1 - Zhiqi Chen, Nkosingiphile Shongwe, Salil Redkar,*
*Debarati Mazumdar, Lakshmi Malavika Andavilli*

## Introduction

The increasing costs of healthcare are not affordable for Vermonters. Many struggle to access preventive primary care services, and health care outcomes for substance abuse, mental health, and chronic disease need to improve. We studied the VERMONT ALL-PAYER ACCOUNTABLE CARE ORGANIZATION MODEL AGREEMENT. With a goal to limit health care cost growth to no more than 3.5% in aggregate across all payers, Vermont is taking major steps to achieve this. Three goals, linking the health care delivery system to population health improvement and public health

1) Increase access to primary care
2) Reduce deaths due to suicide and drug overdose
3) Reduce the prevalence and morbidity of chronic disease

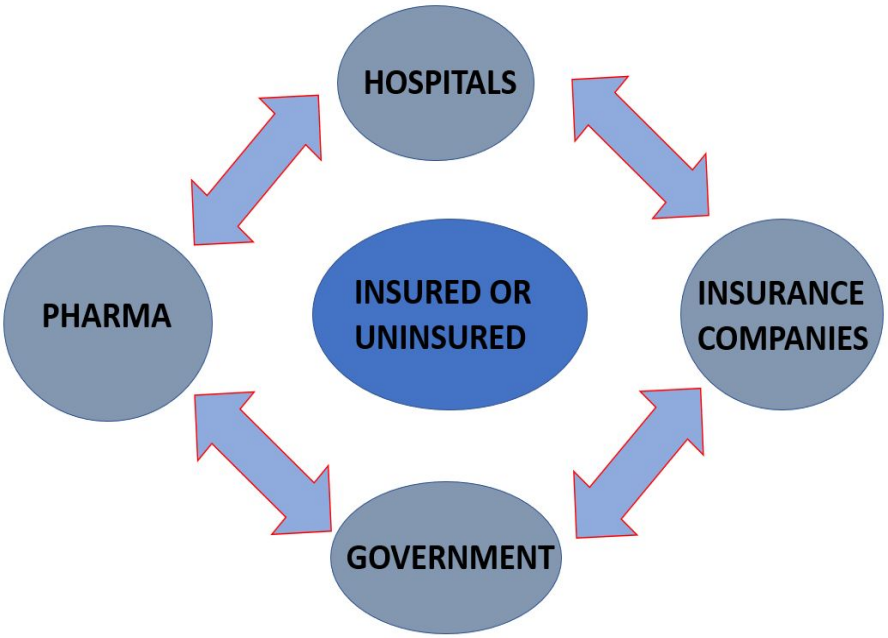
In order to achieve these goals, one major step was to ensure that there was enough information collected so that it is available to researchers and the public.

As per a Washington Post article,Vermont abandoned the idea of changing to a government-financed universal health-care system in late 2014 due to concerns over costs. Vermont wanted to move to a single payer model whereby healthcare providers are paid for their services by the government instead of private insurers. However, it could not embrace this model. Currently, it is retooling how government pays for the healthcare and reduce the underlying costs of healthcare system in Vermont.

The Vermont Hospitals datasets present information about patient health issues and hospital services provided in fourteen Vermont acute care hospitals within inpatient, revenue, and

emergency department as of 2016. Analysis of insurance claims are useful for understanding hospital utilization, the costs of admission and treatment, resource development and drug abuse. This analysis includes all people served by each hospital in the state of Vermont. In the datasets, individuals may have multiple records if they had multiple inpatient visits, procedures, and emergency department visits during the year (2016). As in any dataset, there are records with incomplete or missing information, which will be excluded from the analysis.

General Ecosystem in healthcare industry is depicted in the following image.




Some of the data that will be used include:

**Inpatient:** includes all patients that are billed as an inpatient stay, regardless of admission source.

**Emergency Department (ED)**: records that originated in the emergency room in terms of admission or non-admission to the inpatient facility.

**Diagnosis Groupings:** Inpatient admissions are often grouped by diagnoses using Medicare Severity Diagnosis Related Groups (MSDRGs) and Major Diagnostic Categories(MDCs). MS-DRG groupings describe conditions and procedures related to similar body systems or etiologies and are further grouped into 25 MDCs. We will be using the 2007 and above MSDRGs.

**Revenue Code:** provide information about costs breakdown of each visit, and also includes CPT codes that will be used for further analysis.

In this project, we will begin by understanding the claim data, and use the Vermont Discharge files to learn more about a patient admitted from different sources, and extract information relating to their basic demographics, diagnosis, medical procedures and the cost of each visit. In order to do that, we will be using the inpatient, emergency and revenue files for 2016 provided by the state of Vermont. In looking at diagnosis and costs, it is also important to study the kind of MDCs covered by different primary payers, which will enlighten us on the demographics each payer covers the most. Therefore, the second part of this project will study the top MDCs covered by Medicare, Medicaid, and Commercial insurance. Further studies will look into the demographics that might be related to the MDC covered and try to explain why a certain MDC occupies a certain cost percentage for the different payers. Finally, the drug abuse epidemic has gotten worse over the past decade not only in the United States but the world in general. It is therefore important to assess the costs (financially) associated with drug abuse, and how much the payers are experiencing a financial obligation towards drug use either through disorders or emergency admissions.

## Question 1

This question makes us think critically about each and every step taken by the doctors or hospitals for the patient. As a group we tried to question everything that the doctor has done for the patient. This helped us understand the sequence of events that were recorded for a patient. This question made us understand how claims data could be designed in future to help analysts like us to draw conclusions and find anomalies easily. Various variables used in answering this question include:

Here is a detailed analysis of the 7 unique patient ids given in the question.

**1. UNIQ: 507033**

A young female aged 25-29 was admitted in an elective fashion for baby delivery to Northwestern Medical Center. The patient had a full term uncomplicated delivery for a single live birth. There were no complications, and the patient was sent to home. This visit cost about $3233.29, and was paid for by BlueCross. A huge portion of the cost was from room and board for a semi private 2 bedroom for a one day stay.

**2. UNIQ: 40436**

A Medicare female patient aged between 70-74 was admitted to University of Vermont Medical Center from another hospital through the urgent route. After treatment, the patient was discharged after one day into own or family care. Patient presented a variety of symptoms ranging from heart issues to diabetes. Specifically, the patient was diagnosed with myocardial infarction, diabetes, asthma,atherosclerotic heart disease, and depression. The patient was treated by inserting a drug-eluting intraluminal device and conducting Fluoroscopy to test its functioning.The expense from this visit was about $70,275.41, whereby $23,275 was for the cardiovascular procedure and implants cost $10146.48 along with cardiology related tests amounting to $8387.29.

**3. UNIQ: 859382**

A male around the thirties was rushed to the emergency where he died (didn't survive) same day due to an overdose (drug user) from a non-physical facility. The patient suffered from acute respiratory failure, brain damage, ventricular fibrillation, and cardiac arrest. Treatments included respiratory ventilation and Insertion of endotracheal airway into the trachea. At the hospital, a lot of procedures were conducted and charges were about $13,000 where $3,339 was for clinical diagnostics and ventilation procedures cost $2575, and patient was self-pay. This visit will probably be covered by the uncompensated pool of the hospital.

**4. UNIQ: 1585831**

A female in her forties was admitted to the emergency room and the visit was paid for by Medicare. The patient didn't survive, and she only stayed one day at the hospital and was admitted from a non-physical facility due to heroin overdose (drug abuse). Patient probably didn't survive due to respiratory failure, heart issues, and poisoning by benzodiazepines, where intubation was performed as well as other respiratory procedures. Charges from this visit amounted to about $17093.79, and the breakdown is as follows for the major approximate costs:
Intensive care -$4,450
Pharmacy -$2,024
Laboratory tests -$2,392

**5. UNIQ: 200760**

This patient is aged between 18-24, and was admitted to the Vermont Medical Center for a through the emergency department due to a motor vehicle accident.. The patient was discharged after 4 days and transferred to Home Care services. The visit was paid for by commercial insurance. The patient suffered displaced fracture of medial malleolus of left tibia - which is a

broken ankle or fracture treatment, and had Gastro-esophageal reflux disease without esophagitis as well as major depressive disorder. The treatment included repositioning the Left Tibia with Internal Fixation Device. The total charges from this visit were $49, 533 with major cost components being Operating room services at $14,055 and Medical/Surgical Supplies charged at $ 10,696.

**6. UNIQ: 3692**

A male patient between the age 18-24 admitted through the urgent care route. The patient was discharged after 6 days to own or family care. Blue Cross paid for this visit. The patient was admitted for bipolar disorder combined with suicidal intentions, with a personal history of traumatic brain surgery, cannabis dependence, nicotine dependence, and constipation. The treatment was mainly for substance abuse and detoxification services. The total charges were $ 117, 895 where the major components of the charges were Psychiatric treatment and Pharmacy.

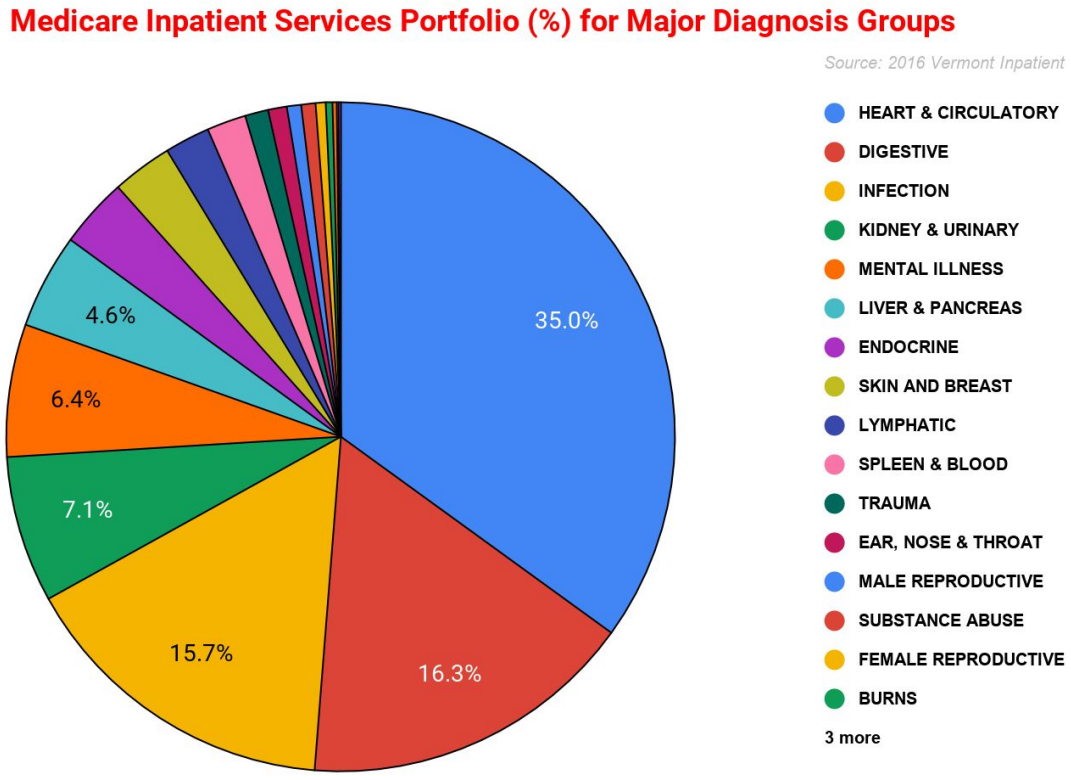
**7. UNIQ: 690326**

A female in her early forties admitted for cosmetic surgery. The patient was discharged after 8 days and will be taken care of by her family. The payment was made by the patient in this case. Patient had an alteration of Bilateral Breast and alteration of Abdominal Wall. The charges amounted to $43,425 and the breakdown for the major approximate costs are $25,919 for operating room services and $5,076 for Room and board.

## Question 2

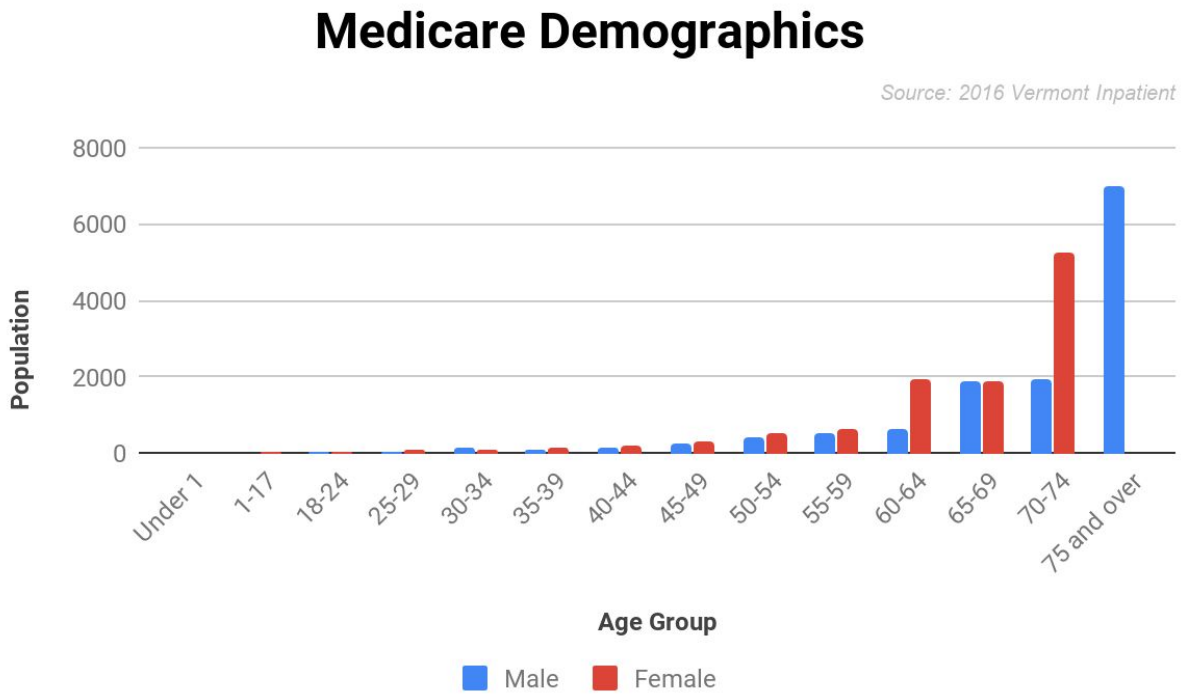
This question will investigate the popular MDCs each payer covers; Medicare, Medicaid and Commercial Insurance. For each payer, we will identify the MDCs and study the demographics of each payee to understand why costs and MDCs covered differ by payer.

### Medicare

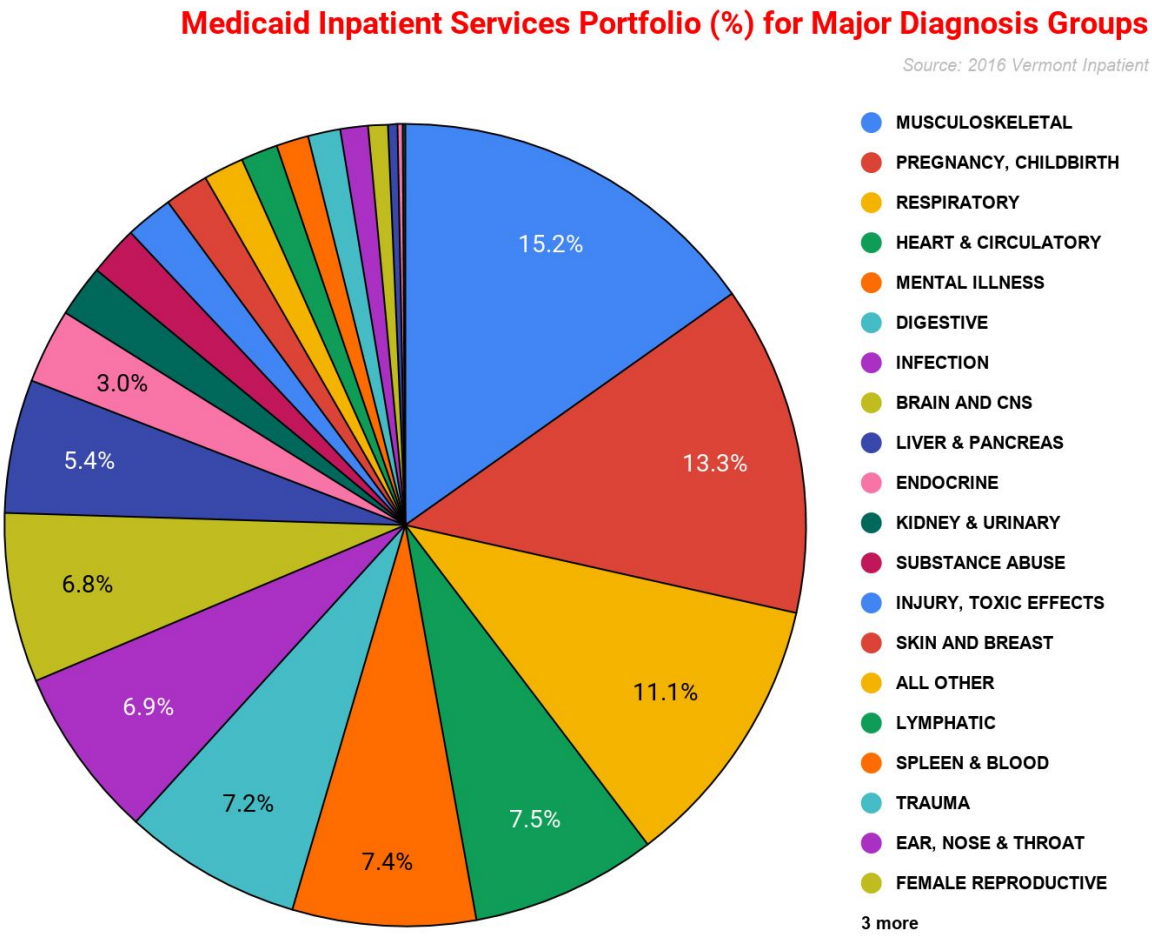



From the pie chart of Medicare we are observing that the top-5 MDCs are age-associated, whose vulnerability increases with age such as Heart and Circulatory, Digestive, Infection, Kidney and Urinary and Mental Illness. The reason these top five MDCs cost Medicare the most is because Medicare covers the elderly population, who are above 65 years and are prone to suffer from these diseases especially heart related diseases. On average, females tend to have the most

reported cases (please note that there is no data for 74 and over for females). The basic trend overall is that as people age, they report these cases, and for those over 65 are covered by Medicare so it makes sense for these diseases to cost Medicare the larger portion. Heart related diseases costs Medicare about 35%. This is because people who are 65 and over have a build up of fatty deposits in the walls of arteries over the years. In addition, older adults, tend to experience more stress than younger people, which negatively affects heart health. This applies to the other conditions since as people age, their health weakens in general, which causes issues such digestive problems, susceptibility to infections and mental illness e.g. depression.

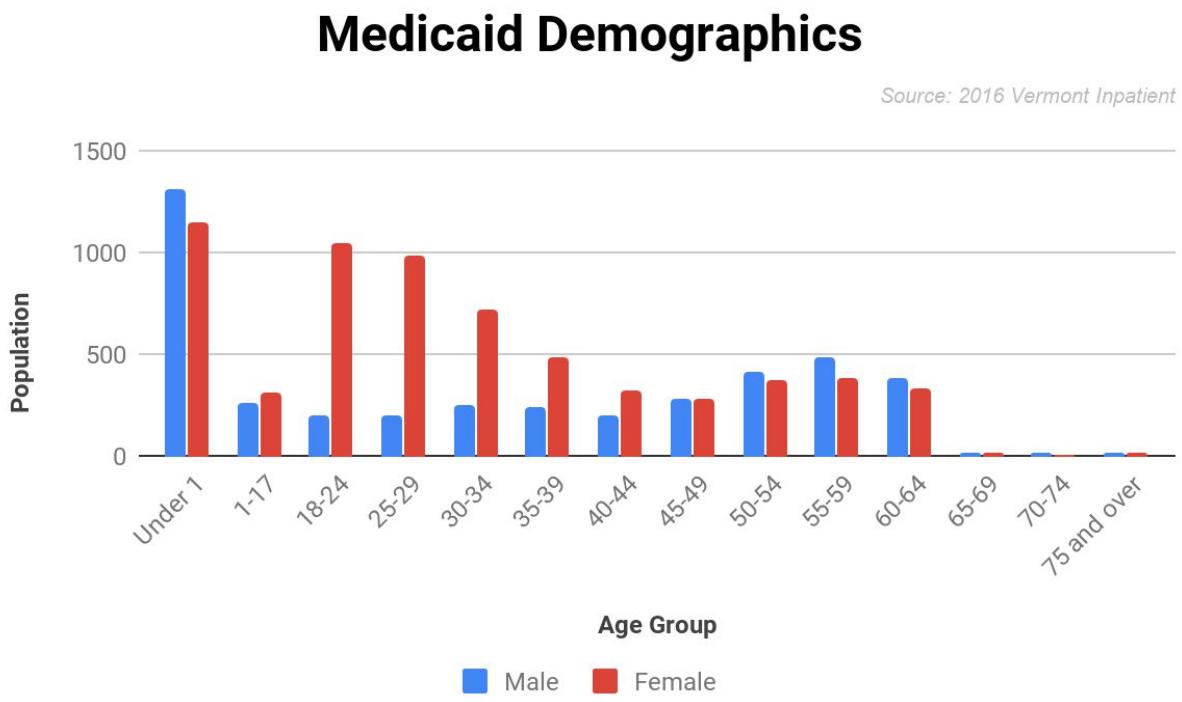


**Medicaid**

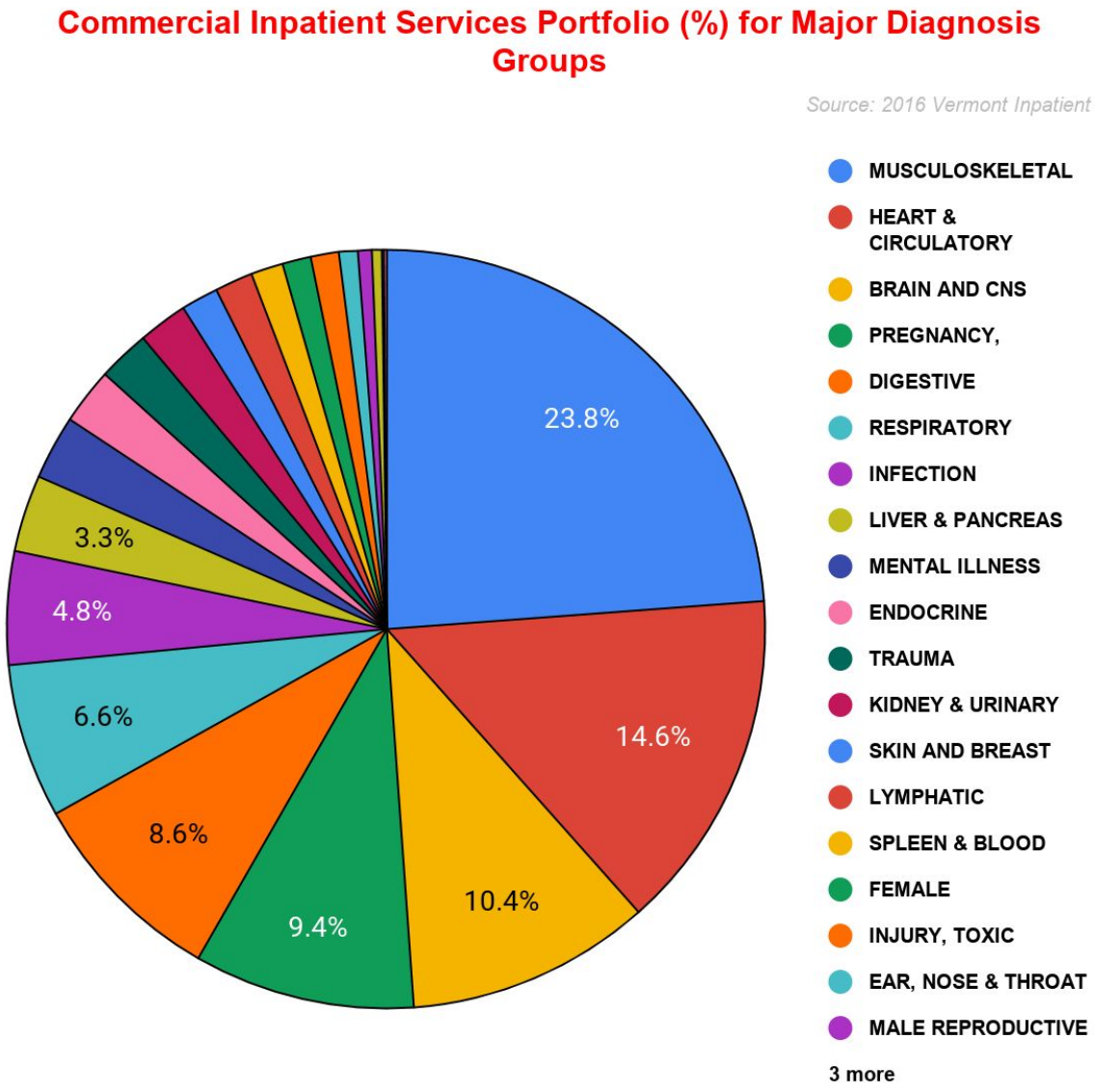



From the pie chart we can infer that the top MDCs are related to poverty and disability, and the limitations to access better healthcare and better living conditions. The pie charts shows that the top five MDCs are musculoskeletal, pregnancy and childbirth, respiratory, heart and circulatory and mental illness. The reason musculoskeletal is costing Medicaid the so much money at about 15.2% is because people usually don't have the necessary resources to be educated about this condition. Children aged under one year old, are likely to be impacted by musculoskeletal disorders due to falls and injuries, and since access to premium health is a challenge for low income families, this disorder affects one in three children each year

(https://www.rheumatologynetwork.com) due to involvement in activities without proper teaching and use of protective gear. Adults between eighteen and thirty-four are also affected the most due to injuries such as car accidents, whose victims have limited access to proper procedure and preventative methods. This notion applies to childbirth and pregnancy too due to lack of proper resources to both manage pregnancy and proper procedure during childbirth, which usually leads to complications, and younger females are affected the most across the board.

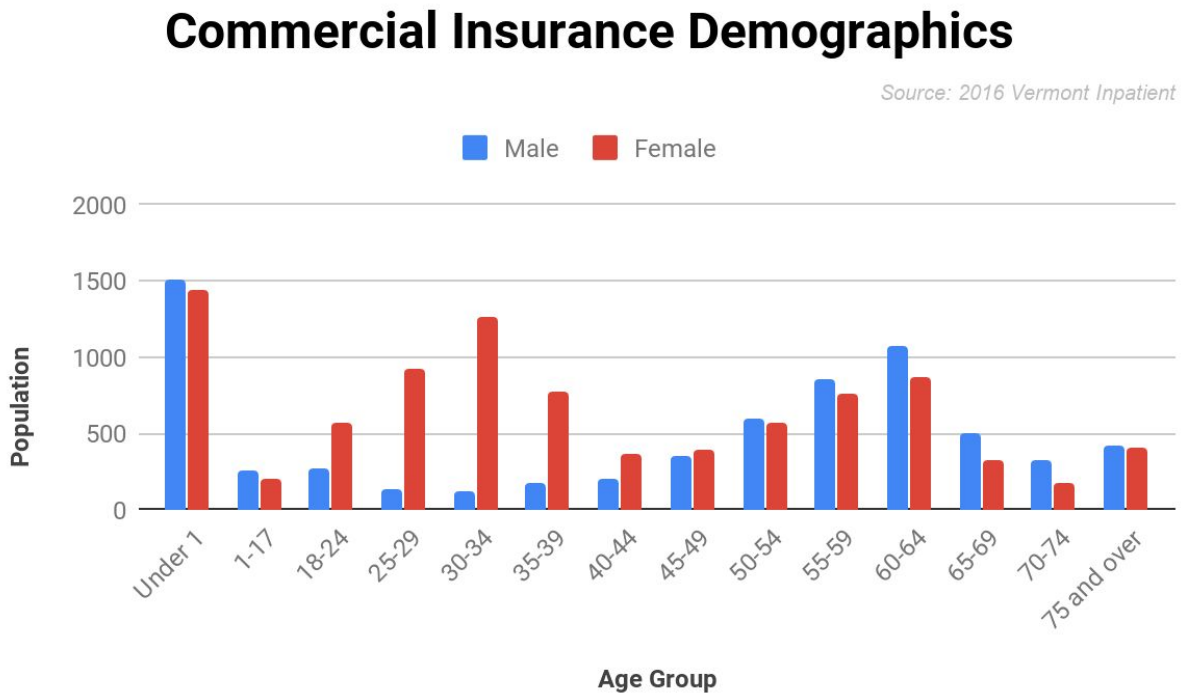


**Commercial Payers**



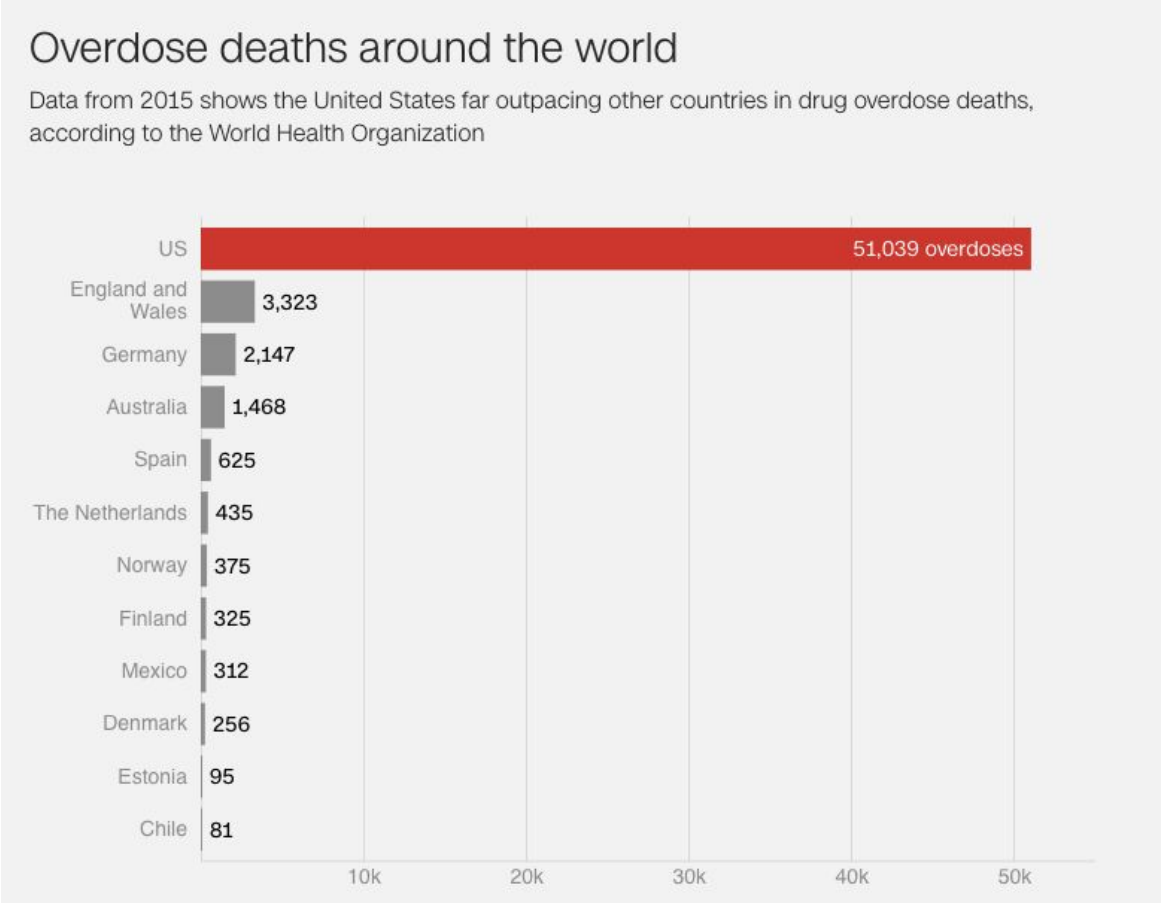

From the pie chart of Medicaid we are observing that the top-5 MDCs are musculoskeletal, heart and circulatory, Brain and CNS, pregnancy and digestive issues. Children younger than a year are more likely to suffer from musculoskeletal issues compared to females between 24 and 39 who could be affected by pregnancy and its complications. A lot of the population is younger than fifty years mainly because they are working or under parents' insurance and covered by

commercial insurance. Again, children under a year are more likely to suffer from falls causing musculoskeletal symptoms, and childbirth complications. Females between eighteen and thirty five are more likely to visit hospitals for pregnancy and childbirth related issues.



## Question 3

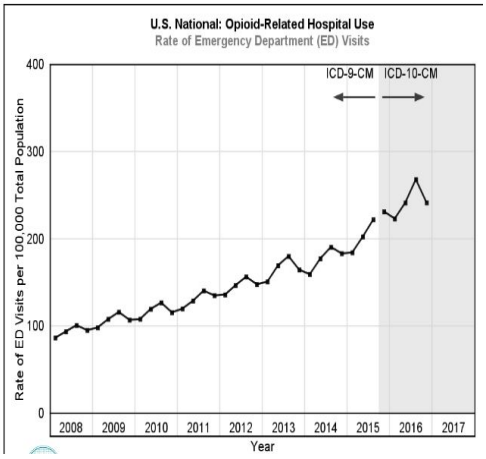
For this question, we will be examining the trends towards drug use and abuse, which sometimes lead to overdose using the 2016 Vermont ED. It is a known fact that in the US, there is a national health crisis in terms of drug overdose. This issue is also being experienced abroad, and therefore there is a need to study this problem so that the costs and losses associated with it could be curbed or lowered at the minimum. Research published in the Annals of Internal Medicine shows that there was approximately 63, 632 drug overdose deaths in the US in 2016 compared to about 53,000 in 2015 (see graph rom World Health Organization). Financially, substance abuse cost the US about $56 billion in 2015, an approximate increase of about 50% compared to $23 billion in 2010, which shows that the country faces an unprecedented drug crisis, and is worsening over the years.

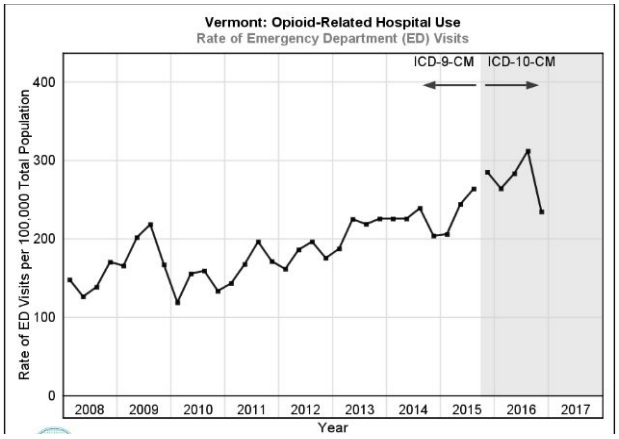


The first step to control this issue is by mandating healthcare providers to pay more attention attention when prescribing drugs, and keeping an updated database of information related to claims. In the previous homework, we observed that there are already systems in place to try to incentivize payers to minimize overprescription of drugs such as heroin. There should a deliberate, sustained and well coordinated education and prevention, which overtime could reduce the number of Americans using drugs. In reducing drug use and abuse, it is important to develop better prescribing practices, and educating both the healthcare provider and patient about alternative drugs, which in turn reduces addiction and the need to explore and use illicit drugs (National Drug Control Strategy AND U.S. Department of Health & Human Services).

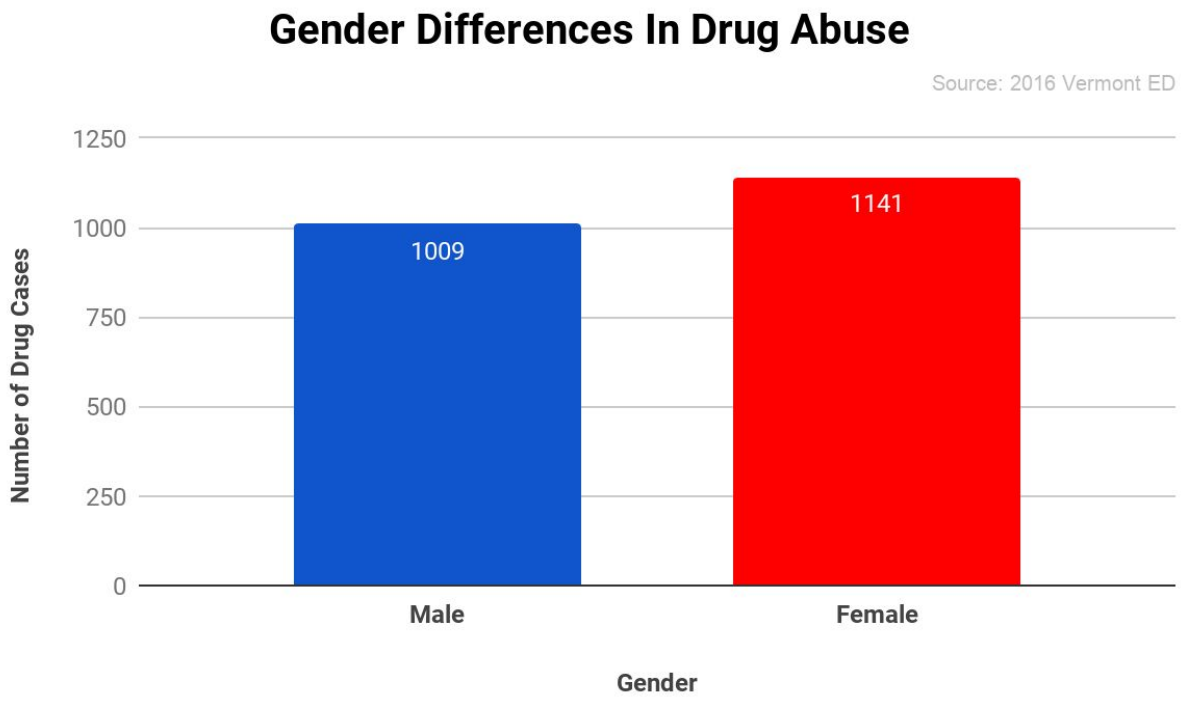
**Number of ED visits diagnosed as drug user/abuser:**

Between 2006 and 2014, the overall rate for substance abuse ED visit increased from 14.1% to 20.3% per 1,000 population in the United States (https://www.hcup-us.ahrq.gov). In Vermont, there are a lot of ED visits due to drug related issues at a rate of about 300 per 10,000 people, compared to the whole of the US at about 250 per 10,000 people (https://www.hcup-us.ahrq.gov: see graph below). The number of such cases whereby the visit was diagnosed as drug user or abuser is  2151 based on the data given.



**Gender Bias:**

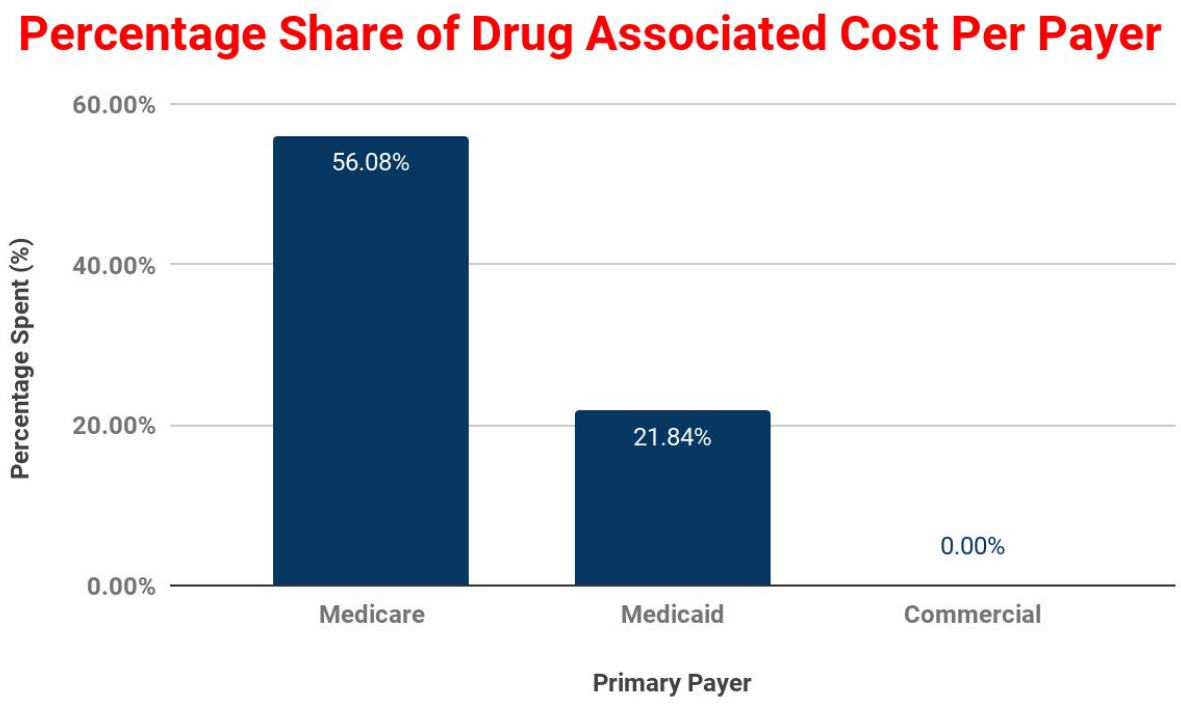



From our analysis, we observed that in terms of drug use or abuse, women are more likely to engage in these activities. Of the total drug cases, about 53.1% of admissions have been female users compared to the 46.9% who are male. Therefore, the myth that drug abuse or use is a male problem is just a gender bias myth on ED admissions. Women may be more susceptible to cravings and relapse, thus a higher ED admission rate compared to males. This is because females even though careful, begin administering drugs at lower doses than males (NIH), and use escalates more, rapidly leading addiction, which increases likelihood of relapse following abstinence. At the same time, the gender differences trend that we observed could just be differences in opportunity and access to drugs than vulnerability to drug use.

**Total Spent:**

In 2016, the state of Vermont spent about $30,741,220 in drug related cases. For a state that had about 623,354 , where by the median age is about 42.6 years, and the sex ratio is 50.7% females and 49.3% males (http://worldpopulationreview.com), this is a huge amount. It is clear then that the state of Vermont is in a crisis, especially opioid use. Since Vermont is on the older range of the spectrum, it makes sense that the percentage share of drug costs of about 56.08% is covered by Medicare. This could be from prescribed drugs since old people are more likely to suffer from pain related issues as we saw in question 2. Medicaid on the other hand covers about 21.84% of drug related expenses. Vermont has about 29% (http://worldpopulationreview.com) of the population with only a high school degree, who might be from low income places, and thus covered by Medicaid.

**By Primary Payers:**

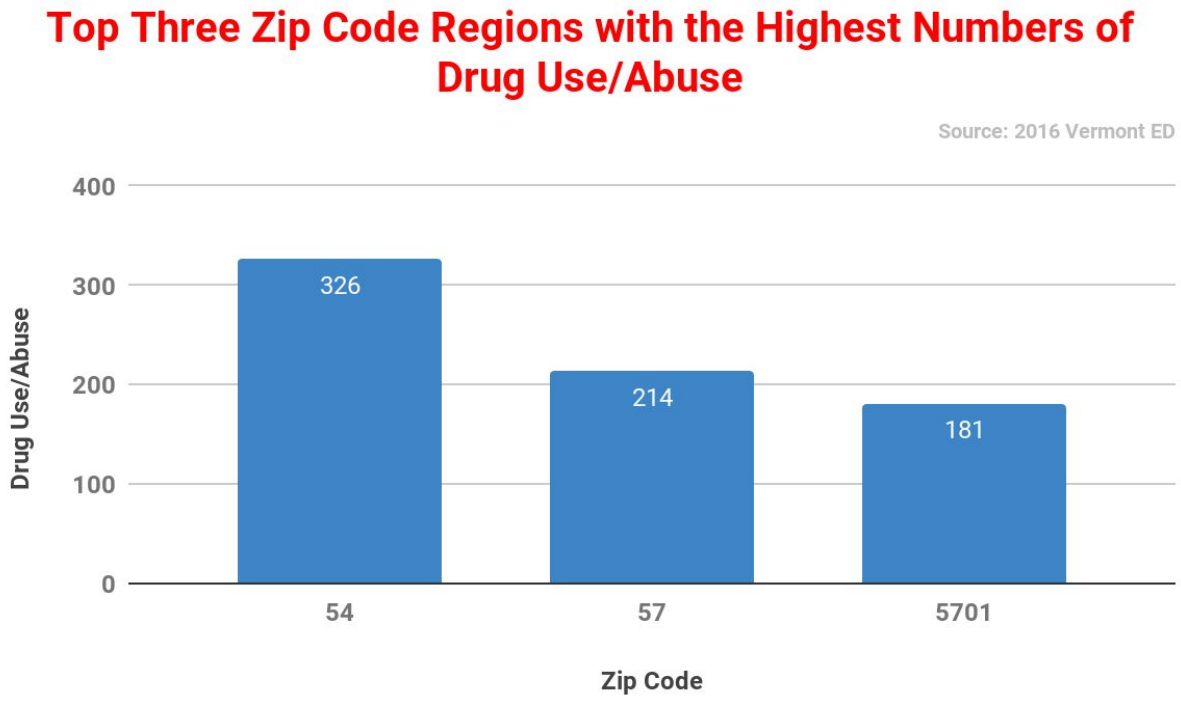


**ED for diagnosis related to synthetic narcotics or amphetamines:**

As per NCIT *(https://ncit.nci.nih.gov* ), Amphetamine is a synthetic substance related to natural sympathomimetic amines. Amphetamine appears to exert its central nervous system (CNS) and peripheral effects indirectly by inducing the release of biogenic amines from their storage sites in nerve terminals. This agent is a commonly abused psychostimulant drug, which may be snorted, taken orally, smoked, or injected. Synthetic opioids are fully synthesized, meaning they're manmade. These drugs, despite the distinctions in how they're made, all act on the brain and body in the same way. They also all have the potential for addiction and physical dependence. Risks of using naturally-derived or synthetic opiates or opioids include addiction and physical dependence because of the opioid receptors they interact with in the central nervous system. All forms of opioids and opiates, natural or otherwise, bind to opioid receptors and, in so doing, relieve pain and create a sense of euphoria.
In Vermont the Emergency Department cases for diagnosis related to synthetic narcotics or amphetamines is 156 for the year 2016.

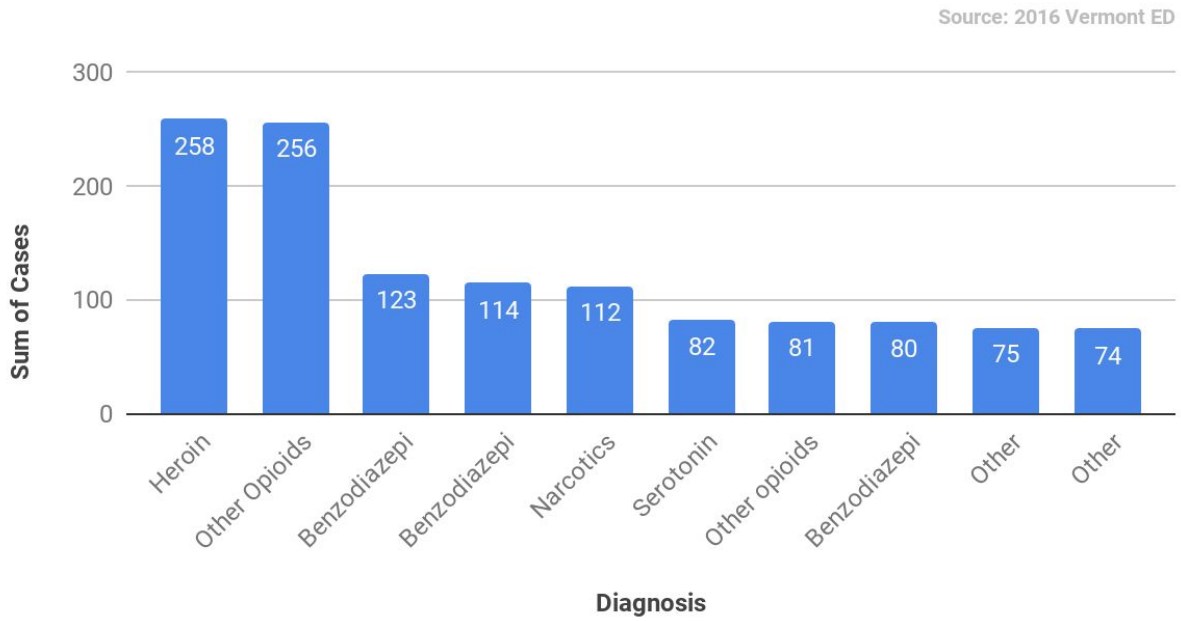
**Zip code regions with the highest numbers of drug use/abuse:**

In Vermont, the three zip codes with the highest numbers if drug use and abuse are 054, 057, and 05701. These zip code regions have similar income distribution statistics, which is estimated to be below the per capita income for the rest of the state. For example, in the 054 zip code regions, the average income as reported in tax returns, is about $39,000 compared to statewide income of about $45,000. Further analysis show that there is about 86% of individuals with an income below $25,000 and about 89% for income below $50,000 (http://www.city-data.com/zips). This low income distribution and unemployment rate of 5.6% (http://www.city-data.com/zips) could explain why the drug usage in 054 for example is as high as 326 cases.




**Common Drug Abuse/Use Diagnosis:**

T401X1A and T402X5A are the two most common diagnoses of drug use or abuse, with 258 (20%) and 256 cases (20% approx) respectively. Thus 40% drug abuse in top 10 is by the use of two drugs. T401X1A means poisoning by heroin, accidental (unintentional), initial encounter, and T402X5A denotes adverse effect of other opioids, initial encounter. The next three are T424X2A, T424X5A and T40605A, which are 123, 114 and 112 respectively which comprise 28% in top 10 and the drugs are benzodiazepines and narcotics.




| Diagnosis code | Number | Type |
|---|---|---|
| T401X1A | 258 | Poisoning by heroin, accidental (unintentional), initial encounter. |
| T402X5A | 256 | Adverse effect of other opioids, initial encounter |
| T424X2A | 123 | Poisoning by benzodiazepines, intentional self-harm, initial encounter |
| T424X5A | 114 | Adverse effect of benzodiazepines, initial encounter |
| T40605A | 112 | Adverse effect of unspecified narcotics, initial encounter |
| T43222A | 82 | Poisoning by selective serotonin reuptake inhibitors, intentional self-harm, initial encounter |
| T402X1A | 81 | Poisoning by other opioids, accidental (unintentional), initial encounter |
| T424X1A | 80 | Poisoning by benzodiazepines, accidental (unintentional), initial encounter |
| T426X5A | 75 | Adverse effect of other antiepileptic and sedative-hypnotic drugs, initial encounter |

| | | |
|---|---|---|
| T426X2A | 74 | Poisoning by other antiepileptic and sedative-hypnotic drugs, intentional self-harm, initial encounter |

## Conclusion

This project aimed to analyse the insurance claim data in the healthcare market. In doing our analysis, we used the All Payer Claim Data published by the State of Vermont, specifically, the Vermont Uniform Hospital Discharge Data Sets. These database consists of inpatient discharge data, outpatient procedures and services data, and emergency department data. In each dataset, the website includes case-specific diagnostic discharge data, demographics of the patient, reason for admission, treatment and services provided for each visit, hospital duration and service specific charges. In our analysis, we constructed a report of each patient according to the patient's specific visit and unique number. We also identified how the cost of each MDC is allocated within three major insurance payers in Vermont. Finally, Vermont has an opioid crisis, among the highest in the country, so we further analysed the different effects and costs of drug usage and abuse.

To observe differences and similarities in expenses incurred by payers, we identified the three top payers and analyzed the data to identify which MDCs cost each payer the most. In analyzing the spread of costs for each for MDC as settled by primary payer, we observed that the top MDCs for Medicare are age related medical issues experienced by adults. For example, Medicare spends about 35% on heart related while Medicaid and Commercial insurance payers spends the most in children and childbearing related issues such as musculoskeletal and pregnancy.

Since Vermont Just like the rest of the us is experiencing a drug usage and abuse epidemic, it is important to analyze the data for trends such as the most issues diagnosed as it related to drug use and abuse using the emergency department data. Drug overdose is a national crisis, and

causes unnecessary costs and deaths that affect us financially, socially and economically as well as the quality of life. There is a huge push to regulate drug use by demanding that healthcare providers to be more careful when submitting drug abuse related data. We discovered that Medicare covers the most in drug related cases, which could be because older people use a lot of pain related medicine. Medicaid on the other hand, spends about 21% on drug related cases. This could be because Medicaid covers mostly families from low income demographics. Therefore, a lot of the people, either through prescription or illicit use lack education on how to manage pain related drugs for example. Overall, drug use and abuse costs the state of Vermont almost $30 million, therefore there is an urgency to reduce the devastating economic effects of such cases. From the analysis, it is also clear that demographics and income distribution also plays a role in drug use or abuse, which is important for the state to introduce programs where the individuals could find alternate ways of income. Most importantly, it appears that heroine is the most used drug leading to poisoning especially for initial encounters, so it is important for the state to pinpoint the origin of this drug, and how the population procures it.

Question 3 gives us some directions on how to implement a planned  proposal for Vermont state to tackle the drug overuse and over-dose problem.We can see from previous analysis in the project that this problem exists more with age band 18 to 24 males. A good start can be by educating the kids in high school with the ill effects of drugs and how to not be an addict to such drugs.We can focus therapy sessions, give support to the community level organizations who are trying to solve this issue. We also see large number of cases are due to poisoning by heroin accidentally and poisoning by other opioids. According to http://www.healthvermont.gov, we understood that most opioid-related fatalities involve multiple substances. The most common combination of substances was involving fentanyl and heroin, which accounted for nearly half (47%) of all opioid-related fatalities. Over one-third of deaths involved fentanyl and cocaine (35%). The third most common combination was cocaine and heroin, found in 21% of deaths. These three substances together – fentanyl, heroin and cocaine – were found in 19% of opioid-related accidental and undetermined fatalities. Government can ask pharma companies to

direct their research in this direction to understand more about this three substances and their occurance in various medicines.

# Appendix of Healthcare HW3

Written by Zhiqi Chen, Nkosingiphile Shongwe, Salil Redkar, Debarati Mazumdar and Lakshmi Malavika Andavilli

2/24/2019

## Question 1

```
library(readxl)
library(dplyr)

ED = read.delim("VTED16.TXT", sep = ",")
INP = read.delim("VTINP16_upd.TXT", sep = ",")
REV = read.delim("VTREVCODE16.TXT", sep = ",")
REVCODE = read_excel("HS_2016VT_REVCODE_FILE_LAYOUT_and_CODES.xls", sheet =
"REVCODE")

REV$REVCODE = formatC(REV$REVCODE, width=4, flag="0")
```

### 1. UNIQ = 507033

```
uniq1_rev = REV %>%
  filter(Uniq == "507033")
uniq1_revcode = merge(uniq1_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq1_inp = INP %>%
  filter(UNIQ == "507033")
```

### 2. UNIQ = 40436

```
uniq2_rev = REV %>%
  filter(Uniq == "40436")
uniq2_revcode = merge(uniq2_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq2_inp = INP %>%
  filter(UNIQ == "40436")
```

### 3. UNIQ = 859382

```
uniq3_rev = REV %>%
  filter(Uniq == "859382")
uniq3_revcode = merge(uniq3_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq3_ed = ED %>%
  filter(UNIQ == "859382")
uniq3_inp = INP %>%
  filter(UNIQ == "859382")
```

### 4. UNIQ = 1585831

```
uniq4_rev = REV %>%
  filter(Uniq == "1585831")
uniq4_revcode = merge(uniq4_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq4_ed = ED %>%
```

```
  filter(UNIQ == "1585831")
uniq4_inp = INP %>%
  filter(UNIQ == "1585831")
```

### 5. UNIQ = 200760

```
uniq5_rev = REV %>%
  filter(Uniq == "200760")
uniq5_revcode = merge(uniq5_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq5_ed = ED %>%
  filter(UNIQ == "200760")
uniq5_inp = INP %>%
  filter(UNIQ == "200760")
```

### 6. UNIQ = 3692

```
uniq6_rev = REV %>%
  filter(Uniq == "3692")
uniq6_revcode = merge(uniq6_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq6_ed = ED %>%
  filter(UNIQ == "3692")
uniq6_inp = INP %>%
  filter(UNIQ == "3692")
```

### 7. UNIQ = 690326

```
uniq7_rev = REV %>%
  filter(Uniq == "690326")
uniq7_revcode = merge(uniq7_rev, REVCODE, by = "REVCODE", all.x = TRUE)
uniq7_inp = INP %>%
  filter(UNIQ == "690326")
```

## Question 2

```
library(readxl)
library(tidyverse)
library(data.table)
library(ggplot2)
library(grid)
library(gridExtra)
library(lattice)
library(ggplotify)
library(stringr)
library(data.table)

# File to use
inpatient <- read.delim("VTINP16_upd.TXT",sep=',')
ed_vermont <- read.delim("VTED16.TXT",sep=',')

# Lists of each for reference only
drg_list <- read_excel("HS_2016VT_PUF_FILE_LAYOUT_and_CODES.xls",sheet="MSDRG
2007 forward")
mdc_list <- read_excel("HS_2016VT_PUF_FILE_LAYOUT_and_CODES.xls",sheet="MDC")
```

```r
payer_list <-
read_excel("HS_2016VT_PUF_FILE_LAYOUT_and_CODES.xls",sheet="PRINCIPAL_PAYMENT
_SOURCE")

# Selecting the rows with payer that is either Medicare/Medicacid/Commerical
ppay_reqd <- c(1,2,6,7)
inpatient1 <- inpatient %>% filter (PPAY %in% ppay_reqd)

# All the commercial players are now labelled as '3', Medicare and Medicaid
are '1', '2' respectively
inpatient1$PPAY[inpatient1$PPAY=='6' | inpatient1$PPAY =='7'] <- "3"

# Cross-tabulation
cross_tab <-inpatient1 %>% group_by(PPAY,MDC) %>%
summarize(Cost_per_cat=sum(CHRGS)) %>%
spread(PPAY,Cost_per_cat)

list <- c("1","2","3")
for (i in list){
        cross_tab[i] <- round(cross_tab[i]/(10^6), digits = 2)
}

cross_tab<-cross_tab %>% filter(!is.na(MDC))
cross_tab<-merge(cross_tab,mdc_list, by="MDC", all.x =TRUE)
cross_tab<- cross_tab[,c(1,5,2,3,4)]

# Demographic information
demo_medicare = inpatient1 %>%
  filter(PPAY=='1') %>%
  group_by(intage,sex) %>%
  summarise(n = n())
demo_medicare = demo_medicare[order(-demo_medicare$n),]
head(demo_medicare)

## # A tibble: 6 x 3
## # Groups:   intage [3]
##   intage   sex     n
##    <int> <int> <int>
## 1     14     2  7006
## 2     14     1  5227
## 3     12     1  1925
## 4     13     2  1919
## 5     13     1  1869
## 6     12     2  1856

demo_medicaid = inpatient1 %>%
  filter(PPAY=='2') %>%
  group_by(intage,sex) %>%
  summarise(n = n())
demo_medicaid = demo_medicaid[order(-demo_medicaid$n),]
head(demo_medicaid)
```

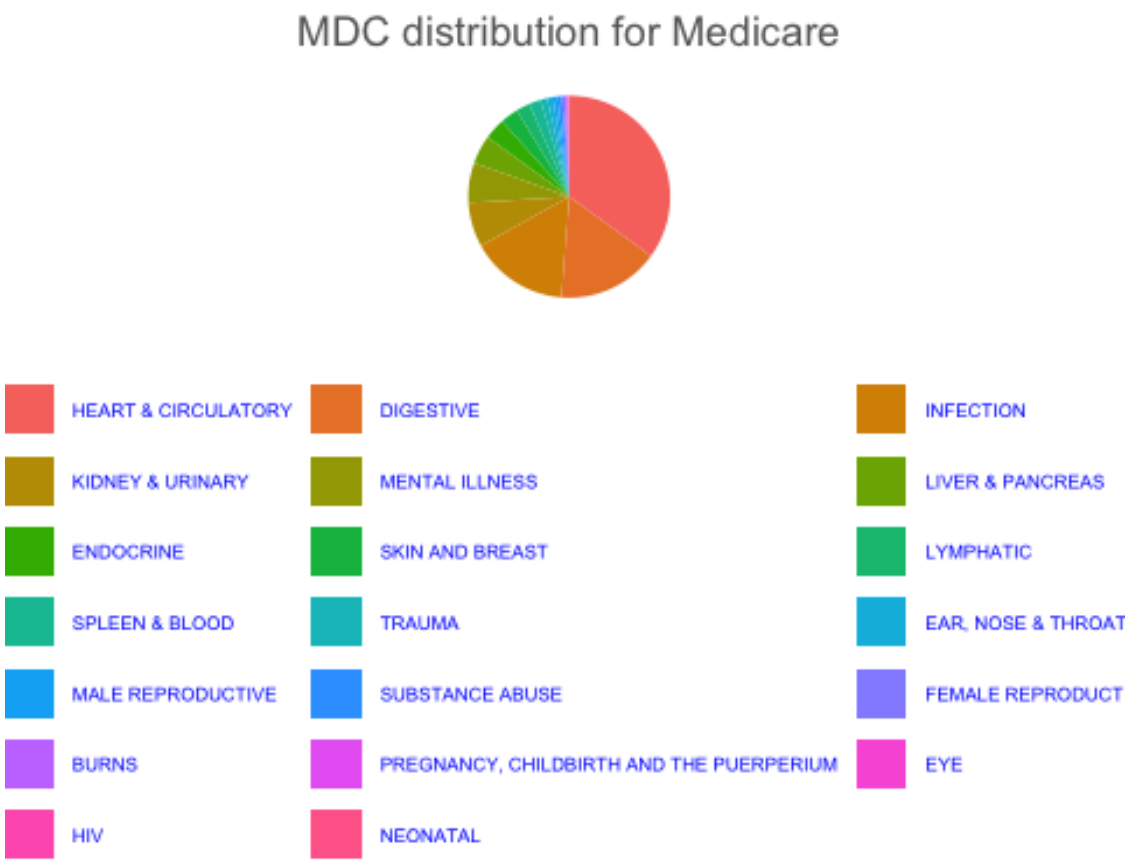
```
## # A tibble: 6 x 3
## # Groups:   intage [5]
##   intage   sex     n
##    <int> <int> <int>
## 1      1     1  1307
## 2      1     2  1148
## 3      3     2  1050
## 4      4     2   988
## 5      5     2   718
## 6     10     1   489
```

```
demo_commercial = inpatient1 %>%
  filter(PPAY=='3') %>%
  group_by(intage,sex) %>%
  summarise(n = n())
demo_commercial = demo_commercial[order(-demo_commercial$n),]
head(demo_commercial)
```

```
## # A tibble: 6 x 3
## # Groups:   intage [4]
##   intage   sex     n
##    <int> <int> <int>
## 1      1     1  1502
## 2      1     2  1438
## 3      5     2  1265
## 4     11     1  1070
## 5      4     2   923
## 6     11     2   865
```

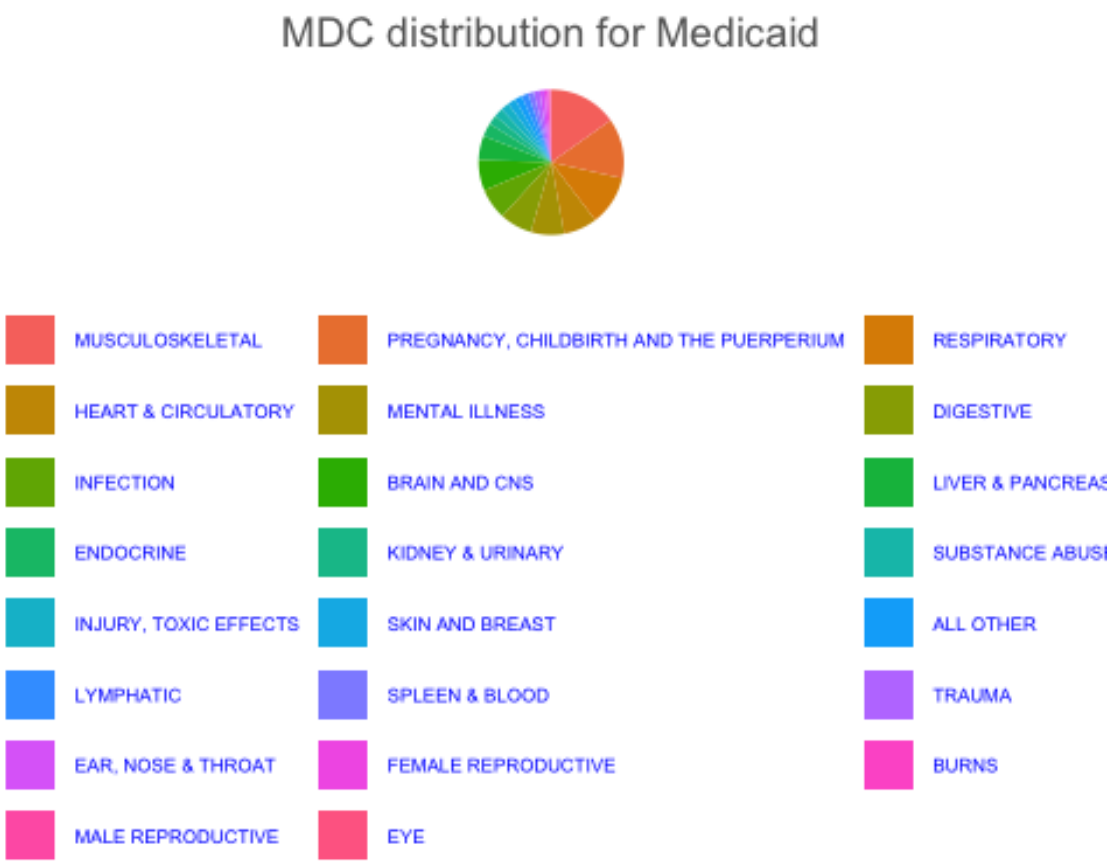
```
# Making the pie-charts
aa<- cross_tab %>% filter(!is.na(`1`))%>%select(MDC,MDC_CAT_NAME,`1`)%>%
mutate(percentage=round((`1`/sum(`1`)*100),digits = 1)) %>%
arrange(desc(percentage))

pie_medicare<- aa %>% ggplot(aes(x="", y=aa$percentage,
fill=factor(MDC_CAT_NAME,levels = aa$MDC_CAT_NAME))) +
geom_bar(stat="identity", width=1)+ coord_polar(theta = "y",
start=0,direction = -1)+
        labs(x = NULL, y = NULL, fill = NULL, title = "MDC distribution for
Medicare")

pie_medicare<- pie_medicare + theme_classic() + theme(axis.line =
element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          plot.title = element_text(hjust = 0.5, color =
"#666666"))+theme(legend.position="bottom")+
        theme(legend.text = element_text(colour="blue", size=6))+
        guides(fill = guide_legend(nrow = 7,byrow = TRUE))
```

```
pie_medicare
```




```
aa1 = aa[order(-aa$percentage),]
aa1 = aa1 %>%
  select(MDC_CAT_NAME, percentage)
head(aa1, 5)

##          MDC_CAT_NAME percentage
## 1 HEART & CIRCULATORY       35.0
## 2           DIGESTIVE       16.3
## 3           INFECTION       15.7
## 4    KIDNEY & URINARY        7.1
## 5      MENTAL ILLNESS        6.4

# Making the pie-charts
aa<- cross_tab %>% filter(!is.na(`2`))%>%select(MDC,MDC_CAT_NAME,`2`)%>%
mutate(percentage=round((`2`/sum(`2`)*100),digits = 1)) %>%
arrange(desc(percentage))

pie_medicaid<- aa %>% ggplot(aes(x="", y=aa$percentage,
fill=factor(MDC_CAT_NAME,levels = aa$MDC_CAT_NAME))) +
geom_bar(stat="identity", width=1)+ coord_polar(theta = "y",
start=0,direction = -1)+
```

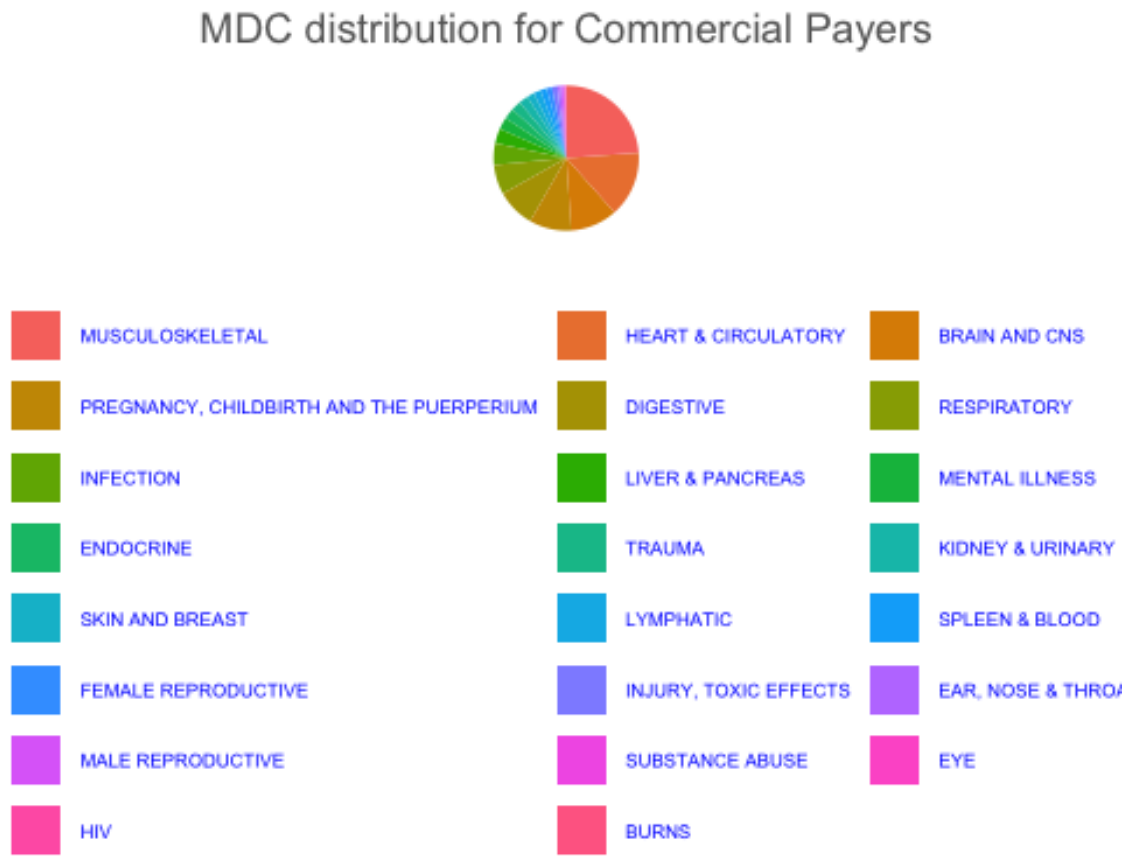
```
        labs(x = NULL, y = NULL, fill = NULL, title = "MDC distribution for
Medicaid")

pie_medicaid<- pie_medicaid + theme_classic() + theme(axis.line =
element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          plot.title = element_text(hjust = 0.5, color =
"#666666"))+theme(legend.position="bottom")+
        theme(legend.text = element_text(colour="blue", size=6))+
        guides(fill = guide_legend(nrow = 8,byrow = TRUE))

pie_medicaid
```




```
aa1 = aa[order(-aa$percentage),]
aa1 = aa1 %>%
  select(MDC_CAT_NAME, percentage)
head(aa1, 5)

##                               MDC_CAT_NAME percentage
## 1                            MUSCULOSKELETAL       15.2
## 2 PREGNANCY, CHILDBIRTH AND THE PUERPERIUM       13.3
## 3                               RESPIRATORY       11.1
```

```
## 4                       HEART & CIRCULATORY        7.5
## 5                            MENTAL ILLNESS        7.4
```

```r
# Making the pie-charts
aa<- cross_tab %>%filter(!is.na(`3`))%>%select(MDC,MDC_CAT_NAME,`3`)%>%
mutate(percentage=round((`3`/sum(`3`)*100),digits = 1)) %>%
arrange(desc(percentage))

pie_commercial <- aa %>% ggplot(aes(x="", y=aa$percentage,
fill=factor(MDC_CAT_NAME,levels = aa$MDC_CAT_NAME))) +
geom_bar(stat="identity", width=1)+ coord_polar(theta = "y",
start=0,direction = -1)+
        labs(x = NULL, y = NULL, fill = NULL, title = "MDC distribution for
Commercial Payers")

pie_commercial<- pie_commercial + theme_classic() + theme(axis.line =
element_blank(),
          axis.text = element_blank(),
          axis.ticks = element_blank(),
          plot.title = element_text(hjust = 0.5, color =
"#666666"))+theme(legend.position="bottom")+
        theme(legend.text = element_text(colour="blue", size=6))+
        guides(fill = guide_legend(nrow = 8,byrow = TRUE))

pie_commercial
```




```
aa1 = aa[order(-aa$percentage),]
aa1 = aa1 %>%
  select(MDC_CAT_NAME, percentage)
head(aa1, 5)
```

```
##                                MDC_CAT_NAME percentage
## 1                            MUSCULOSKELETAL       23.8
## 2                      HEART & CIRCULATORY       14.6
## 3                            BRAIN AND CNS       10.4
## 4 PREGNANCY, CHILDBIRTH AND THE PUERPERIUM        9.4
## 5                                 DIGESTIVE        8.6
```

## QUESTION 3

```
total_drug <- ed_vermont %>% filter_at(vars(starts_with("DX")),
any_vars(str_detect(., pattern = "T40|T41|T42|T43")))

# 1. How many ED visits exactly have been diagnosed as drug user/abuser?
nrow(total_drug)
```

```
## [1] 2151
```

```
# 2. Men vs Women
counts<-total_drug %>% filter(!is.na(sex)) %>% group_by(sex) %>%
```

```r
summarize(counts=n())
head(counts)
```

```
## # A tibble: 2 x 2
##     sex counts
##   <int>  <int>
## 1     1   1009
## 2     2   1141
```

```r
# 3. Money spent on the drug abusers
m = total_drug %>%
  summarise(Total = sum(CHRGS))
total_amount = m[1,1]
total_amount = total_amount/(10^6)
total_amount = round(total_amount, digits = 2)
print(sprintf("The total cost in millions of dollars for the identified
patients is %s", total_amount))
```

```
## [1] "The total cost in millions of dollars for the identified patients is
30.74"
```

```r
m1 <-total_drug %>%
  filter(PPAY == "1") %>%
  summarize(total_amount_1 = sum(CHRGS))
total_amount_1 = m1[1,1]
total_amount_1 = total_amount_1/(10^6)
percent_1 = total_amount_1/total_amount*100
percent_1 = round(percent_1, digits = 2)
print(sprintf("The percentage share of total by Medicare is %s", percent_1))
```

```
## [1] "The percentage share of total by Medicare is 56.08"
```

```r
m2 <-total_drug %>%
  filter(PPAY == "2") %>%
  summarize(total_amount_2 = sum(CHRGS))
total_amount_2 = m2[1,1]
total_amount_2 = total_amount_2/(10^6)
percent_2 = total_amount_2/total_amount*100
percent_2 = round(percent_2, digits = 2)
print(sprintf("The percentage share of total by Medicaid is %s", percent_2))
```

```
## [1] "The percentage share of total by Medicaid is 21.84"
```

```r
m3 <-total_drug %>%
  filter(PPAY == "3") %>%
  summarize(total_amount_3 = sum(CHRGS))
total_amount_3 = m3[1,1]
total_amount_3 = total_amount_3/(10^6)
percent_3 = total_amount_3/total_amount*100
percent_3 = round(percent_3, digits = 2)
print(sprintf("The percentage share of total by Commercial Payers is %s",
percent_3))
```

```
## [1] "The percentage share of total by Commercial Payers is 0"
```

```
# 4. How many of patients have been brought to ED for diagnosis related to
synthetic narcotics or amphetamines?
total_syn <- ed_vermont %>%
filter_at(vars(starts_with("DX")),any_vars(str_detect(., pattern =
"T404|T4362")))
nrow(total_syn)
```

```
## [1] 156
```

```
# 5. Name the 3 zip code regions with the highest numbers of drug use/abuse.
zip = total_drug %>%
  group_by(TXTZIP) %>%
  summarise(n = n())
zip = zip[order(-zip$n),]
head(zip, 3)
```

```
## # A tibble: 3 x 2
##   TXTZIP     n
##   <fct>  <int>
## 1 054      326
## 2 057      214
## 3 05701    181
```

```
# 6. What are the 10 most common diagnoses of drug use/abuse?
df1 = data.frame(total_drug[, c(10:29)])
df2 = data.frame(DX = unlist(df1, use.names = FALSE))
df3 = df2 %>%
  group_by(DX) %>%
  summarise(n = n())
df4 = df3[-1,]
df5 = df4 %>%
  filter(str_detect(DX, paste(c("T40","T41","T42","T43"), collapse = "|")))
df5 = df5[order(-df5$n),]
head(df5, 10)
```

```
## # A tibble: 10 x 2
##    DX          n
##    <fct>   <int>
##  1 T401X1A   258
##  2 T402X5A   256
##  3 T424X2A   123
##  4 T424X5A   114
##  5 T40605A   112
##  6 T43222A    82
##  7 T402X1A    81
##  8 T424X1A    80
##  9 T426X5A    75
## 10 T426X2A    74
```