EHR Database Exploration (Synthea)

Author: Marissa Munoz-Ruiz

Goal of Script: Explore an electric health records (EHR) database

How do hospitals track the information of their patients? Information such as electronic health records are commonly stored in relational databases.

The data used for this script consists of six excel files that mimic the basic structure of a relational database. Each file can be seen as a table and may be related to another table by a common column i.e. field. For example, the field PatientID is present in the Patient file and also occurs in the OutpatientVisit file. Due to this structure, we can determine the number of hospital visits that a specific patient had for a specific year even though the information is spread across two different files.

The six excel files were downloaded from Synthea, a synthetic data generator that models the medical history of synthetic patients and their associated health records (Synthea (https://synthetichealth.github.io/synthea/)).

I will use tidyverse to explore the data. Each section follows the same pattern:

- Question of interest
- Tidyverse code
- Brief explanation of code output

Which staff member makes the most money?

The summary function shows that the max value of Hourlyrate is $20. Therefore I will use use Salary to determine employee compensation. The minimum value of Salary is $1, so I will assume this is a data collection error and remove the observation from the Staff dataset.

```
summary(Staff)
```

```
##     StaffID       FirstName           LastName            Gender
##  Min.   : 1.00   Length:50          Length:50          Length:50
##  1st Qu.:13.25   Class :character   Class :character   Class :character
##  Median :25.50   Mode  :character   Mode  :character   Mode  :character
##  Mean   :25.50
##  3rd Qu.:37.75
##  Max.   :50.00
##
##     HireDate            HourlyRate        Salary          PayType
##  Min.   :2000-03-26   Min.   :13.00   Min.   :     1   Length:50
##  1st Qu.:2008-05-29   1st Qu.:15.00   1st Qu.: 56200   Class :character
##  Median :2010-05-21   Median :15.00   Median : 68329   Mode  :character
##  Mean   :2009-11-13   Mean   :15.43   Mean   :109566
##  3rd Qu.:2012-02-25   3rd Qu.:16.00   3rd Qu.: 94214
##  Max.   :2014-05-22   Max.   :20.00   Max.   :999999
##                       NA's   :29      NA's   :21
##   StaffType         StaffReportsTo
##  Length:50          Min.   : 1.00
##  Class :character   1st Qu.: 7.00
##  Mode  :character   Median :35.00
##                     Mean   :25.17
##                     3rd Qu.:44.00
##                     Max.   :46.00
##                     NA's   :2
```

```
Staff_new <- Staff %>%
  filter(!(Salary == min(Salary,na.rm=TRUE)))

Staff_new %>%
  filter(Salary == max(Salary))
```

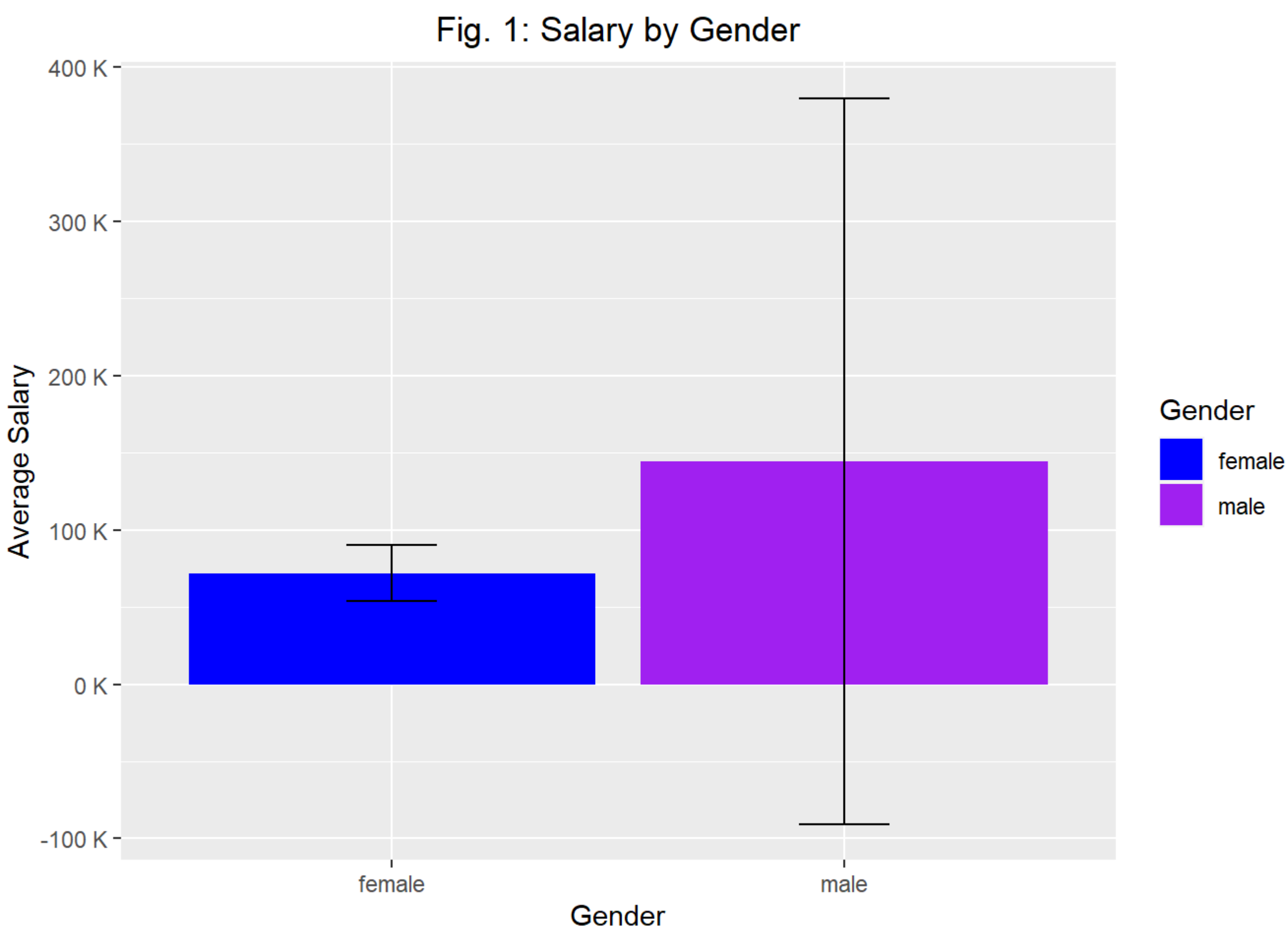
```
## # A tibble: 1 x 10
##   StaffID FirstName LastName Gender HireDate   HourlyRate Salary PayType
##     <dbl> <chr>     <chr>    <chr>  <date>          <dbl>  <dbl> <chr>
## 1       4 Joshua    Lucas    male   2011-10-06         NA 999999 Salary
## # ... with 2 more variables: StaffType <chr>, StaffReportsTo <dbl>
```

According to the output, the highest paid staff member is Joshua Lucas with a salary of $999,999, which is extremely high! Does Joshua’s salary point to a pay disparity?

Is there a pay disparity across gender among staff members?

```
Gender <- Staff_new %>%
  group_by(Gender) %>%
  summarize(Mean_Salary = mean(Salary, na.rm = TRUE),
            Median_Salary = median(Salary, na.rm = TRUE),
            SD_Salary = sd(Salary, na.rm = TRUE),
            Skew=e1071::skewness(Salary))

Gender %>%
  ggplot(aes(x=Gender,y=Mean_Salary, fill=Gender)) +
  geom_col() +
  geom_errorbar(aes(ymin=Mean_Salary-SD_Salary, ymax=Mean_Salary+SD_Salary),width=.2) +
  scale_fill_manual(values=c("blue","purple")) +
  labs(y="Average Salary", title = "Fig. 1: Salary by Gender") +
  theme(plot.title = element_text(hjust=0.5)) +
  #scale_y_continuous(labels = function(x) format(x, scientific = FALSE))
  scale_y_continuous(labels = label_number(suffix = " K", scale = 1e-3))
```




```
Gender
```

```
## # A tibble: 2 x 5
##   Gender Mean_Salary Median_Salary SD_Salary  Skew
##   <chr>        <dbl>         <dbl>     <dbl> <dbl>
## 1 female      72001.         67750     17923. 0.510
## 2 male       144586.         70508.   235059. 2.98
```

The average salary for each gender listed in the data is shown in Fig. 1. The plot suggests that males have a higher average salary compared to female staff. However the standard deviation for males is quite substantial ($\pm$ \$ 235,059) and is also heavily right-skewed (Skew = 2.98). A two-sample t-test is commonly used to determine if two means are statistically different. The t-test can be used when certain assumptions are met. Let's check the most important assumptions (outliers, normality, and heteroscedasticity) and also assume the samples are independent.

```
Staff_new %>%
  group_by(Gender) %>%
  identify_outliers(Salary)
```

```
## # A tibble: 2 x 12
##   Gender StaffID FirstName LastName HireDate   HourlyRate Salary PayType
##   <chr>    <dbl> <chr>     <chr>    <date>          <dbl>  <dbl> <chr>
## 1 male         4 Joshua    Lucas    2011-10-06         NA 999999 Salary
## 2 male         7 David     Mungo    2005-01-27         NA 259233 Salary
## # ... with 4 more variables: StaffType <chr>, StaffReportsTo <dbl>,
## #   is.outlier <lgl>, is.extreme <lgl>
```

```
Staff_new %>%
  group_by(Gender) %>%
  shapiro_test(Salary)
```

```
## # A tibble: 2 x 4
##   Gender variable statistic          p
##   <chr>  <chr>        <dbl>      <dbl>
## 1 female Salary       0.917 0.259
## 2 male   Salary       0.461 0.00000102
```

```
Staff_new %>%
  levene_test(Salary ~ Gender)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1    26      1.37 0.253
```

There are two outliers in the male group, most notably Joshua Lucas who is also the highest paid staff member. The Shapiro-Wilk normality test shows that the normality assumption does not hold (p-value < 0.5) and the Levene Test for Equality of Variances shows that the homogeneity assumption holds (p-value > 0.5). Since the normality assumption does not hold and the sample size is fairly small (n = 28), the t-test isn't the most optimal test to use.

However, not all is lost! The Mann-Withney-Wilcoxon test test does not assume the data is normally distributed and compares the median instead of the mean.

```
wilcox.test(Salary ~ Gender, data=Staff_new)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Salary by Gender
## W = 85, p-value = 0.6313
## alternative hypothesis: true location shift is not equal to 0
```

A two-sample Mann-Withney-Wilcoxon test suggests there was not a significant difference between male and females with regards to Salary (p > 0.5). What if Joshua Lucas's salary was also a data collection error? Would there be a difference in salary if Josha Lucas was removed?

```
Staff_NoJosh <- Staff %>%
  filter(!(Salary == max(Salary,na.rm=TRUE)))

Staff_NoJosh  %>%
  group_by(Gender) %>%
  shapiro_test(Salary)
```

```
## # A tibble: 2 x 4
##   Gender variable statistic      p
##   <chr>  <chr>        <dbl>  <dbl>
## 1 female Salary       0.917 0.259
## 2 male   Salary       0.848 0.0126
```

```
wilcox.test(Salary ~ Gender,data=Staff_NoJosh)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Salary by Gender
## W = 97, p-value = 0.9818
## alternative hypothesis: true location shift is not equal to 0
```

The Shapiro-Wilk normality test shows that the normality assumption does not hold (p-value < 0.5) for the male group. A two-sample Mann-Withney-Wilcoxon test suggests there was not a significant difference between male and females with regards to Salary (p > 0.5). Due to the small size of the data set and the synthetic nature of the data, this may not hold true for the general population of medical staff.

What is the salary breakdown when staff type is considered?

```
#Salary by gender & staff type
Gender_Staff <- Staff_new %>%
  group_by(Gender,StaffType) %>%
  summarize(Mean_Salary = mean(Salary, na.rm = TRUE),
            Median_Salary = median(Salary, na.rm = TRUE),
            SD_Salary = sd(Salary, na.rm = TRUE),
            Skew=e1071::skewness(Salary))

Gender_Staff %>%
  ggplot(aes(x=Gender,y=Mean_Salary, fill=Gender)) +
  geom_col() +
  facet_wrap(~StaffType) +
  scale_fill_manual(values=c("blue","purple")) +
  geom_errorbar(aes(ymin=Mean_Salary-SD_Salary, ymax=Mean_Salary+SD_Salary),width=.2) +
  labs(y="Average Salary", title="Fig. 2: Salary by Gender and Staff Type") +
  theme(plot.title = element_text(hjust=0.5)) +
  scale_y_continuous(labels = label_number(suffix = " K", scale = 1e-3))
```


```
Gender_Staff
```

```
## # A tibble: 8 x 6
## # Groups:   Gender [2]
##   Gender StaffType           Mean_Salary Median_Salary SD_Salary     Skew
##   <chr>  <chr>                     <dbl>         <dbl>     <dbl>    <dbl>
## 1 female Clinic Manager          104765         104765        NA  NaN
## 2 female Nurse                    68654.         68485    15010.  -0.0290
## 3 female Physician                70609.         56459    24734.   0.385
## 4 female Physician Assistant      67750          67750      819.   0
## 5 male   Clinic Manager          103953         103953        NA  NaN
## 6 male   Nurse                    58419.         64662    21352.   0.121
## 7 male   Physician               180881         161728    70747.   0.251
## 8 male   Physician Assistant     380339.         71395   536642.   0.385
```

In Fig.2, the average salary is similar between female and male staff members when the position is clinic managers, nurses, or physicians. However, physicians assistants show a a large jump. As mentioned earlier, Joshua Lucas is mostly driving this difference.

Which staff member saw the most patients in 2016?

```
Staff %>%
  inner_join(OutpatientVisit, by = 'StaffID') %>%
  mutate(Year = year(VisitDate),.after='VisitDate') %>%
  group_by(Year,StaffID,StaffType,FirstName,LastName) %>%
  summarize(Visits = n()) %>%
  filter(Year == 2016) %>%
  arrange(desc(Visits)) %>%
  head()
```

```
## `summarise()` regrouping output by 'Year', 'StaffID', 'StaffType', 'FirstName' (override with
`.groups` argument)
```

```
## # A tibble: 6 x 6
## # Groups:   Year, StaffID, StaffType, FirstName [6]
##    Year StaffID StaffType           FirstName LastName Visits
##   <dbl>   <dbl> <chr>               <chr>     <chr>     <int>
## 1  2016      12 Nurse               Juliann   Williams    479
## 2  2016       4 Physician Assistant Joshua    Lucas       459
## 3  2016      30 Nurse               Mark      Carman      456
## 4  2016      37 Physician Assistant Lisa      Willis      436
## 5  2016      35 Physician           Steven    Bechtel     433
## 6  2016      32 Nurse               Elizabeth Schell      431
```

In 2016, Juliann Williams had 479 outpatient visits as a nurse.

Which staff member saw the most patients in primary care settings in 2016?

```
Staff %>%
  inner_join(OutpatientVisit, by = 'StaffID') %>%
  inner_join(Clinic, by = 'ClinicCode') %>%
  mutate(Year = year(VisitDate),.after='VisitDate') %>%
  filter(ClinicDescription == "Primary Care", Year == 2016) %>%
  group_by(StaffID,StaffType,ClinicDescription,FirstName,LastName) %>%
  summarize(Visits = n()) %>%
  arrange(desc(Visits)) %>%
  head()
```

```
## # A tibble: 6 x 6
## # Groups:   StaffID, StaffType, ClinicDescription, FirstName [6]
##   StaffID StaffType           ClinicDescription FirstName LastName Visits
##     <dbl> <chr>               <chr>             <chr>     <chr>     <int>
## 1      12 Nurse               Primary Care      Juliann   Williams    479
## 2       4 Physician Assistant Primary Care      Joshua    Lucas       459
## 3      30 Nurse               Primary Care      Mark      Carman      456
## 4      37 Physician Assistant Primary Care      Lisa      Willis      436
## 5      35 Physician           Primary Care      Steven    Bechtel     433
## 6      32 Nurse               Primary Care      Elizabeth Schell      431
```

In 2016, Juliann Williams had 479 outpatient visits, all of which were in a primary care setting.

Is there a difference in mortality between men and women?

```
Patient %>%
  inner_join(Mortality, by = "PatientID") %>%
  group_by(Gender) %>%
  summarize(Count = n()) %>%
  filter(Gender %in% c('female','male')) %>%
  mutate(Proportion = round(Count/sum(Count),2))
```

```
## # A tibble: 2 x 3
##   Gender Count Proportion
##   <chr>  <int>      <dbl>
## 1 female  2494       0.39
## 2 male    3946       0.61
```

Of all the people who were deceased in the data, 39% were female and 61% were male. It seems that there is a difference in mortality between gender. However, a two-proportions test would have to be conducted to determine if this difference is significant. The two-proportions test can be used when the sample size is large. The total number of patients in the data is 9045.

```
Props <- Patient %>%
  left_join(Mortality, by = "PatientID") %>%
  mutate(Deceased = ifelse(is.na(DateOfDeath),'Not Deceased','Deceased')) %>%
  filter(Gender %in% c('female','male'))

table(Props$Gender, Props$Deceased)
```

```
##
##          Deceased Not Deceased
##   female     2494         6887
##   male       3946         5099
```

```
prop.test(table(Props$Gender, Props$Deceased),correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  table(Props$Gender, Props$Deceased)
## X-squared = 588.17, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1839851 -0.1568281
## sample estimates:
##    prop 1    prop 2
## 0.2658565 0.4362631
```

The two-proportions test strongly suggests that there is a difference in mortality rates (p-value < 0.5) between males and females.

Which disease is most prevalent? Which disease is least prevalent?

```
#Assumption: Each visit counts regardless if it was the same patientt

Outpatient <- OutpatientVisit %>%
  mutate(ICD10 = ifelse(
    (ICD10_1 %in% DiseaseMap$ICD10) |
    (ICD10_2 %in% DiseaseMap$ICD10) |
    (ICD10_2 %in% DiseaseMap$ICD10),
    c(ICD10_1,ICD10_2,ICD10_3), NA), .after = 'ICD10_3')

Prevalence <- Outpatient %>%
  inner_join(DiseaseMap, by = "ICD10") %>%
  group_by(Condition) %>%
  summarize(Visits = n()) %>%
  mutate(Percent = round(Visits/sum(Visits, na.rm = TRUE),4))

Prevalence %>%
  filter(Percent == max(Percent))
```

```
## # A tibble: 1 x 3
##   Condition Visits Percent
##   <chr>      <int>   <dbl>
## 1 Paralysis  80702   0.468
```

```
Prevalence %>%
  filter(Percent == min(Percent))
```

```
## # A tibble: 1 x 3
##   Condition Visits Percent
##   <chr>      <int>   <dbl>
## 1 HIV          419 0.00240
```

Of all outpatient visits where a condition was listed, the most common condition was paralysis (46.8%) and the least common condition was HIV (0.2%).

Are there any diseases that are unevenly distributed across races?

```
#Assumption: Each visit counts regardless if it was the same patient

Outpatient %>%
  inner_join(Patient, by = "PatientID") %>%
  inner_join(DiseaseMap, by = "ICD10") %>%
  group_by(Race, Condition) %>%
  summarize(Visits = n()) %>%
  filter(Race %in% c("white","hispanic","black")) %>%
  spread(.,Race,Visits) %>%
  rowwise() %>%
  mutate(total = sum(c(black,hispanic,white)),
         perc_black = round(black/total,2),
         perc_hispanic = round(hispanic/total,2),
         perc_white = round(white/total,2)) %>%
  arrange(desc(perc_white))
```

```
## # A tibble: 22 x 8
## # Rowwise:
##    Condition      black hispanic white total perc_black perc_hispanic perc_white
##    <chr>          <int>    <int> <int> <int>      <dbl>         <dbl>      <dbl>
##  1 Pulmonary        247      586  3697  4530       0.05          0.13       0.82
##  2 Metastatic_so~   173      207  1064  1444       0.12          0.14       0.74
##  3 Peptic_ulcer_~    51       64   311   426       0.12          0.15       0.73
##  4 Peripheral_va~   139      144   698   981       0.14          0.15       0.71
##  5 Diabetes_with~   206      320  1108  1634       0.13          0.2        0.68
##  6 Alcohol          583      610  2211  3404       0.17          0.18       0.65
##  7 Cancer           292      385  1181  1858       0.16          0.21       0.64
##  8 LiverMild         48      123   305   476       0.1           0.26       0.64
##  9 Paralysis       7440    11588 33143 52171       0.14          0.22       0.64
## 10 Dementia         184      313   854  1351       0.14          0.23       0.63
## # ... with 12 more rows
```

Across all the conditions, there were more outpatient visits by the white population than any other ethnic group. There were large amounts of outpatients visits for pulmonary conditions, metastatic solid tumors, and peptic ulcer diseases in the white population.

Are there any diseases that are unevenly distributed across gender?

```
#Assumption: Each visit counts regardless if it was the same patient

Outpatient %>%
  inner_join(Patient, by = "PatientID") %>%
  inner_join(DiseaseMap, by = "ICD10") %>%
  group_by(Gender, Condition) %>%
  summarize(Visits = n()) %>%
  filter(Gender %in% c("female","male")) %>%
  spread(.,Gender,Visits) %>%
  rowwise() %>%
  mutate(total = sum(c(female,male)),
         perc_female = round(female/total,4),
         perc_male = round(male/total,4)) %>%
  arrange(desc(perc_female))
```

```
## # A tibble: 22 x 6
## # Rowwise:
##    Condition                    female  male total perc_female perc_male
##    <chr>                         <int> <int> <int>       <dbl>     <dbl>
##  1 Depression                     6387  1862  8249       0.774     0.226
##  2 Dementia                       1435   559  1994       0.720     0.280
##  3 Peptic_ulcer_disease            416   164   580       0.717     0.283
##  4 Peripheral_vascular_disease    1010   404  1414       0.714     0.286
##  5 Drugs                          1636   718  2354       0.695     0.305
##  6 Cancer                         1879   842  2721       0.691     0.309
##  7 Pulmonary                      4312  1946  6258       0.689     0.311
##  8 Metastatic_solid_tumour        1428   672  2100       0.68      0.32
##  9 Renal                          1413   673  2086       0.677     0.323
## 10 Diabetes_with_complications    1720   820  2540       0.677     0.323
## # ... with 12 more rows
```

Across all conditions, there were more outpatient visits by the female population than the male population. The most outpatients visits for female patients were for depression, dementia, and peptic ulcer disease.