Midterm Project

MSQM Health Analytics - Team 2
Felecia Manning (fm150), **Spencer Millen** (smm246), **Kunal Shukla** (ks765),
Lauren Siewny (les62), **Ifrah Sohail** (is172), **Pengxi Zheng** (pz68)

10/28/2020

Wellness Product Market Sizing

Task 1: A first look at the data

1) Read in the data from 2014 and check the dimensions of this table.

```
#getwd()
#setwd()

full_2014 = read_csv("2014_data.csv")
dim(full_2014)
```

There are 77 columns and 2631171 rows in the dataset.

2) You decide that this data_frame takes much too long to manipulate and decide to take a subset of it in order to make later steps more manageable. Take a random sample of 100,000 rows from this data_frame.

```
# Create a random sample of 100,000 rows
sampled_data_2014 <- full_2014 %>%
  sample_n(100000)

# Check the dimensions of the sampled data
dim(sampled_data_2014)
view(sampled_data_2014)
```

There are 77 columns and 100000 rows in the dataset.

3) While this is still a sizeable sample, you worry that this subset may not be representative of the full dataset. To check this, you want to look at some summary statistics from both tables to ensure they are still representative. Look at the average age and %Female in both tables. In your opinion, is there significant variation? NOTE: A bit of cleaning is required before looking at average age. Per the data dictionary, *detail_age* shows the individual's age in years only when detail_age_type == 1. Do some quality checks (e.g. sort in descending/ascending order to find extreme values) to make sure this makes sense. Provide rationale for any other exclusions.

```
# Full_2014 detail_age to numeric
full_2014 <- full_2014 %>%
  mutate(detail_age = as.numeric(as.character(detail_age)))

# Changes M/F to numeric values
full_2014 <- full_2014 %>%
  mutate(sex_numeric = ifelse(sex == "F", 0, ifelse(sex == 'M', 1, NA)))
```

```
# Makes removes patients with no recorded age
full_2014 <- full_2014 %>%
  filter(detail_age_type != 9)

full_2014 <- full_2014 %>%
  filter(detail_age != 999)

# Makes patients coded for ages less than 1 year of life, and makes them 0
full_2014 <- full_2014 %>%
  mutate(detail_age = ifelse(detail_age_type > 1, 0, detail_age))

# Re-create a random sample of 100,000 rows
sampled_data_2014 <- full_2014 %>%
  sample_n(100000)

# Calculate the average age and percentage of females and males
# in the sampled dataset
sampled_data_summary_2014 <- sampled_data_2014 %>%
  summarize(
    Average_Age = mean(as.numeric(detail_age, na.rm = TRUE)),
    Percentage_Female = mean(sex_numeric == 0) * 100,
    Percentage_Male = mean(sex_numeric == 1) * 100
  )

# Summary statistics for full data set
full_data_summary_2014 = full_2014 %>%
  summarize(
    Average_Age = mean(as.numeric(detail_age, na.rm = TRUE)),
    Percentage_Female = mean(sex_numeric == 0) * 100,
    Percentage_Male = mean(sex_numeric == 1) * 100
  )

# Makes a list of data frames
list_of_dataframes <- list(
  `Full 2014 Data` = full_data_summary_2014,
  `Sampled Data 2014` = sampled_data_summary_2014
)

# Use bind_rows on the named list, specifying .id to create a new column
# with the dataframe names
full_vs_sample <- bind_rows(list_of_dataframes, .id = "DataFrameSource")
```

```
# Table to compare summary statistics
options(digits = 4)

knitr::kable(full_vs_sample)
```

DataFrameSource	Average_Age	Percentage_Female	Percentage_Male
Full 2014 Data	73.14	49.40	50.60
Sampled Data 2014	73.17	49.16	50.84

4) Satisfied with the degree of variability (or lack thereof) in the previous step, you remove

the larger data_frame (using rm()) and repeat steps 1-2 with the 2015 data_frame.

```r
# Drop full_2014
rm(full_2014)

# Read in 2015 dataset
full_2015 = read_csv("2015_data.csv")
dim(full_2015)

# Full_2015 detail_age to numeric
full_2015 <- full_2015 %>%
  mutate(detail_age = as.numeric(as.character(detail_age)))

# Changes M/F to numeric values
full_2015 <- full_2015 %>%
  mutate(sex_numeric = ifelse(sex == "F", 0, ifelse(sex == 'M', 1, NA)))

# Makes removes patients with no recorded age
full_2015 <- full_2015 %>%
  filter(detail_age_type != 9)

full_2015 <- full_2015 %>%
  filter(detail_age != 999)

# Makes patients coded for ages less than 1 year of life, and makes them 0
full_2015 <- full_2015 %>%
  mutate(detail_age = ifelse(detail_age_type > 1, 0, detail_age))
```

There are 78 columns and 2717657 rows in the dataset.

```r
# Create a random sample of 100,000 rows
sampled_data_2015 <- full_2015 %>%
  sample_n(100000)

# Check the dimensions of the sampled data
dim(sampled_data_2015)
view(sampled_data_2015)
```

```r
# Calculate the average age and percentage of females and males in the
# sampled dataset
sampled_data_summary_2015 <- sampled_data_2015 %>%
  summarize(
    Average_Age = mean(as.numeric(detail_age, na.rm = TRUE)),
    Percentage_Female = mean(sex_numeric == 0) * 100,
    Percentage_Male = mean(sex_numeric == 1) * 100
  )

#Compare the summary statistics
sampled_data_summary_2015

# Summary statistics for full data set
full_data_summary_2015 = full_2015 %>%
  summarize(Average_Age = mean(as.numeric(detail_age, na.rm = TRUE)),
            Percentage_Female = mean(sex_numeric == 0) * 100,
            Percentage_Male = mean(sex_numeric == 1) * 100)
```

```
# Makes a list of data frames
list_of_dataframes <- list(`Full 2015 Data` = full_data_summary_2015,
                           `Sampled Data 2015` = sampled_data_summary_2015)

# Use bind_rows on the named list, specifying .id to create a new column
# with the dataframe names
full_vs_sample <- bind_rows(list_of_dataframes, .id = "DataFrameSource")

#Drop full_2015
rm(full_2015)
```

```
# Table to compare summary statistics
options(digits = 4)

knitr::kable(full_vs_sample)
```

DataFrameSource	Average_Age	Percentage_Female	Percentage_Male
Full 2015 Data	73.20	49.34	50.66
Sampled Data 2015	73.25	49.36	50.64

5) Combine these two data_frames into a single data_frame

```
# Convert 'detail_age' to integer in both data frames
sampled_data_2014 <- sampled_data_2014 %>%
  mutate(detail_age = as.integer(detail_age))

sampled_data_2015 <- sampled_data_2015 %>%
  mutate(detail_age = as.integer(detail_age))

# Combine with bind_rows
combined_sample_data <- bind_rows(sampled_data_2014, sampled_data_2015)

# Check the combined data
combined_sample_data
```

Task 2: How does age of death vary by Male/Female?

```
# Filter the data to remove missing age_of_death values
combined_sample_data_age_of_death <- combined_sample_data %>%
  group_by(sex) %>%
  summarize(Average_Age = mean(as.integer(detail_age, na.rm = TRUE)))

# Calculate summary statistics
summary_stats_for_age_of_death <- combined_sample_data %>%
  group_by(sex, current_data_year) %>%
  summarize(
    Average_Age_of_Death = mean(as.integer(detail_age)),
    Median_Age_of_Death = median(as.integer(detail_age)),
    Average_Age_by_Year = mean(as.integer(detail_age))
  )

# Create data frame
```

```r
age_data <- combined_sample_data %>%
  select(sex, detail_age)
```

1) You are aware of some basic medical knowledge that women tend to have a longer life expectancy than men, so you generate two more questions:

a) What is the average age of death for women versus men in these data?

```r
options(digits = 4)

knitr::kable(combined_sample_data_age_of_death)
```

sex	Average_Age
F	76.57
M	69.95

b) Does this relationship hold when we look at the average age by year? Note any differences.

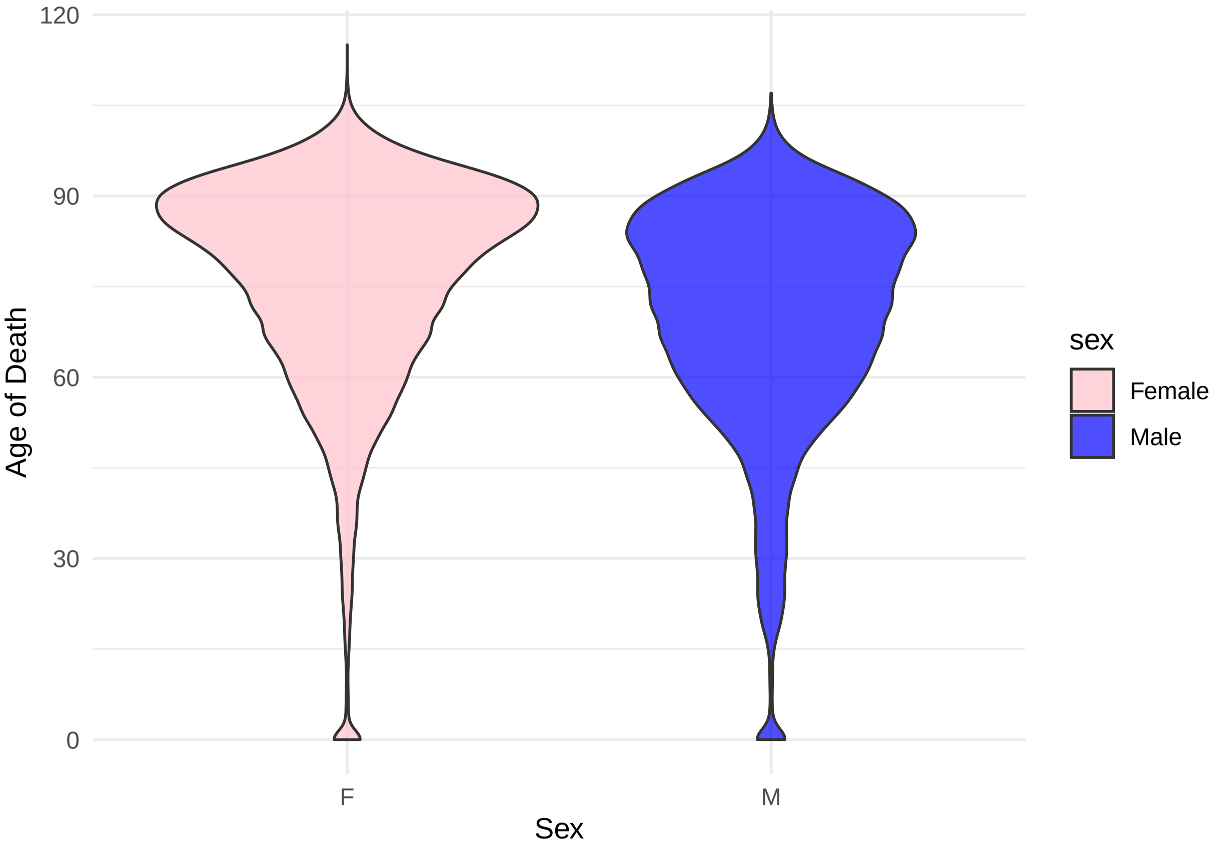
```r
options(digits = 4)

knitr::kable(summary_stats_for_age_of_death)
```

sex	current_data_year	Average_Age_of_Death	Median_Age_of_Death	Average_Age_by_Year
F	2014	76.47	81	76.47
F	2015	76.66	81	76.66
M	2014	69.99	73	69.99
M	2015	69.92	73	69.92

Women in the years 2014-2015 live longer then men, with an average age of death around 76 years, as compared to a male average age of death around 70 years. There is no apparent difference between years.

2) Unsatisfied with the simple averages, you would like to see the distribution of age of mortality over this population over both years. Plot the distribution of age for Male and Female and note any differences. What might explain the difference in averages?

```r
#Violin Plot
ggplot(age_data, aes(x = sex, y = detail_age, fill = sex)) +
  geom_violin(alpha = 0.7) +
  labs(x = "Sex", y = "Age of Death") +
  scale_fill_manual(values = c("pink", "blue"), labels = c("Female", "Male")) +
  theme_minimal()
```



Males have a lower average year of death, reflecting both increased infant mortality rate and a higher risk of death from late-teenage years through middle-age. Women's narrow distribution and smaller variance reflects their decreased risk of early mortality.

Task 3: What are the most prevalent diagnoses for cause of death?

1) You would like to see which diseases, over the two years, are most prevalent. Consider the ICD10 code that describes the underlying cause of death (icd_code_10th_revision). What are the top 5 most prevalent diseases?

```
# Counting occurrences of each ICD-10 code
icd_count <- combined_sample_data %>%
  filter(!is.na(icd_code_10th_revision)) %>%
  group_by(icd_code_10th_revision) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

# Selecting the top 5 most prevalent diseases
top_5_diseases <- head(icd_count, 5)

# Creating a table for ICD-10 code and descriptions with disease names
icd_descriptions <- data.frame(
  icd_code_10th_revision = c("I251", "C349", "F03", "I219", "J449"),
  description = c(
    "Atherosclerotic heart disease of a native coronary artery: CAD",
    "Malignant neoplasm of lung, unspecified: Lung cancer",
    "Unspecified dementia, unspecified severity: Dementia",
    "Acute myocardial infarction: MI",
    "Chronic obstructive pulmonary disease: COPD"))

# Joining with the lookup table to include descriptions
top_5_diseases_with_desc <- top_5_diseases %>%
  left_join(icd_descriptions, by = "icd_code_10th_revision") %>%
  select(icd_code_10th_revision, description, count)
```

```
# Displaying the top 5 most prevalent diseases with descriptions
knitr::kable(top_5_diseases_with_desc)
```

icd_code_10th_revision	description	count
I251	Atherosclerotic heart disease of a native coronary artery: CAD	12236
C349	Malignant neoplasm of lung, unspecified: Lung cancer	11617
F03	Unspecified dementia, unspecified severity: Dementia	8729
J449	Chronic obstructive pulmonary disease: COPD	8378
I219	Acute myocardial infarction: MI	8357

2) You decide that the ICD 10 code is too granular and instead would like to look at a grouping of these codes. You opt for the *Elixhauser* comorbidity groupings from HCUP (https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp). Read in the .csv called "Elixhauser.csv" that relates ICD codes to the Elixhauser comorbidity group and join this with the dataset used in 1 (above) to find the top 5 most common categories.

```
#load Elixhauser data , rename the column
Elixhauser = read.csv('Elixhauser.csv')
colnames(Elixhauser) <- c('icd_code_10th_revision','disease')

# join Elixhauser data in sample data
summary_stats_icd = combined_sample_data %>%
  left_join(Elixhauser, join_by(icd_code_10th_revision))

# Selects columns of interest
selected_columns <- summary_stats_icd %>%
  select(icd_code_10th_revision, disease, detail_age)

# Count the occurrences of each disease category
disease_count = selected_columns %>%
  filter(!is.na(disease)) %>%
  group_by(disease) %>%
  summarize(
    count = n(),
    mean_age = mean(detail_age, na.rm = TRUE),
    median_age = median(detail_age, na.rm = TRUE),
  ) %>%
  arrange(desc(count))

# Select the top 5 most common categories
top_5_diseases_categories <- head(disease_count, 5)
```

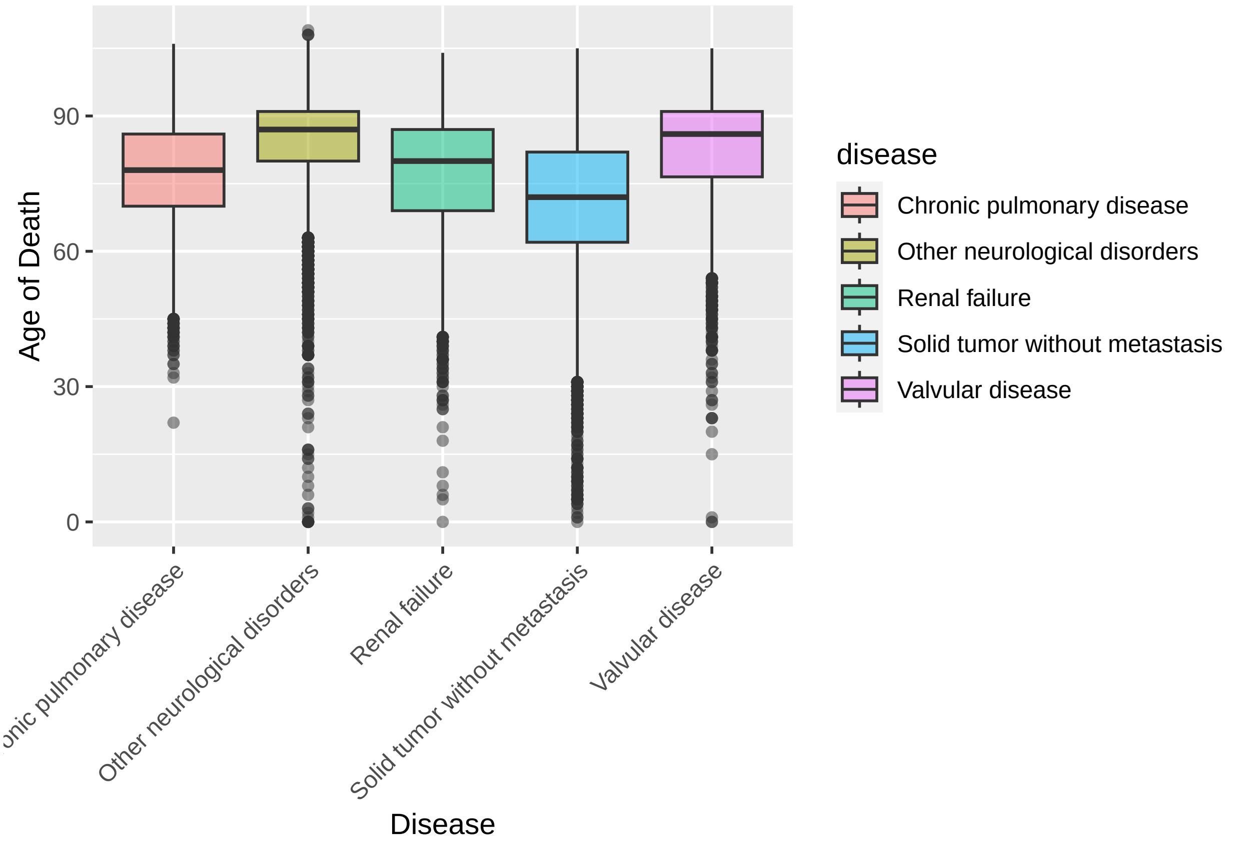
```
# Displaying the top 5 most common categories
options(digits = 4)

knitr::kable(top_5_diseases_categories)
```

disease	count	mean_age	median_age
Solid tumor without metastasis	19955	70.91	72
Chronic pulmonary disease	10982	77.37	78
Other neurological disorders	10741	84.85	87
Renal failure	2913	77.10	80
Valvular disease	2103	81.86	86

3) Visualize the distribution of age of death for each of the top 5 disease categories found in 2 (above).

```r
selected_columns

# Vector for naming the ICD-10 codes
icd_codes <- c("Solid tumor without metastasis", "Chronic pulmonary disease",
               "Other neurological disorders", "Renal failure",
               "Valvular disease")

# Filter your_data for rows with ICD codes in icd_codes
filtered_ICD_data <- selected_columns %>%
  filter(disease %in% icd_codes)
```

```r
# Create plot to visualize the distribution of age of the death
ggplot(filtered_ICD_data, aes(x = disease, y = detail_age, fill = disease)) +
  geom_boxplot(alpha = 0.5) +
  labs(x = "Disease", y = "Age of Death") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

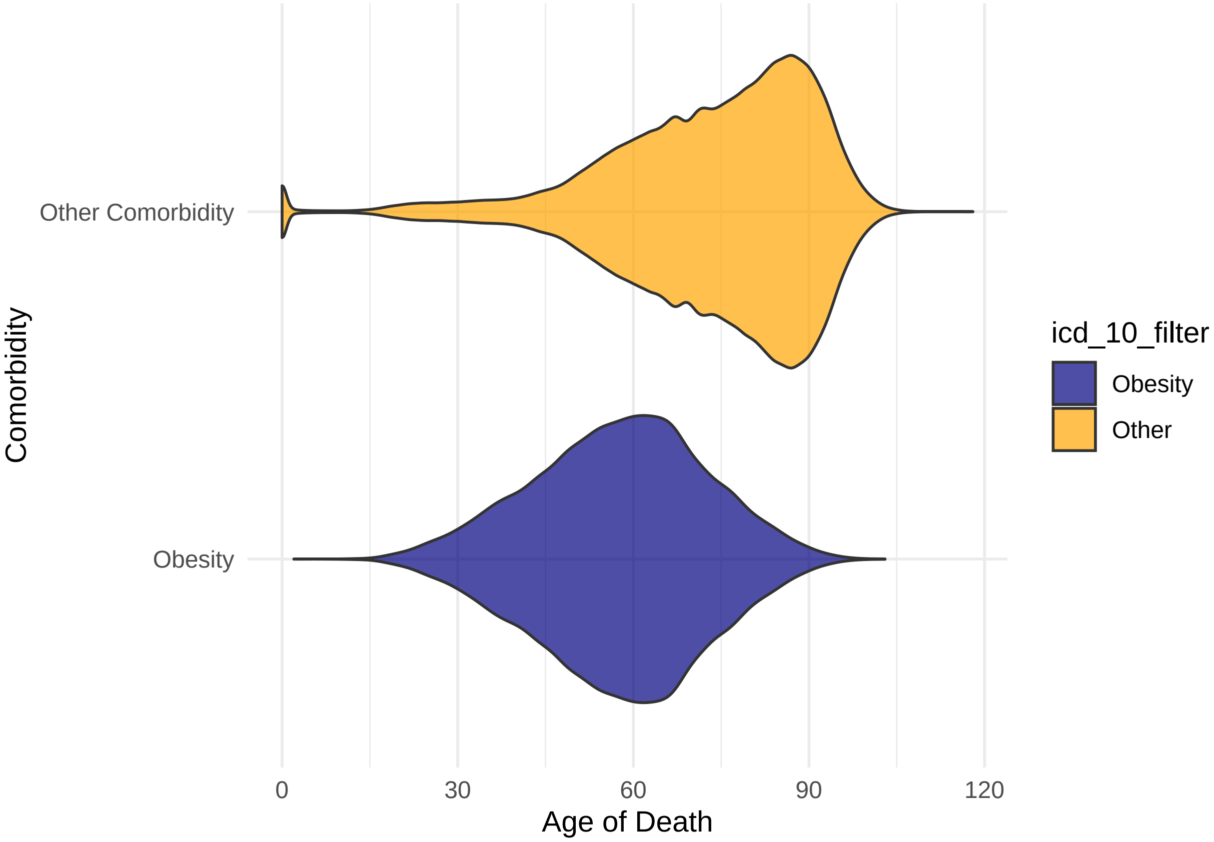


Task 4: Create a recommendation for a target market/product

Building off your analyses in the previous tasks, consider a subset of the total population and a cause of death disproportionately impacting that population (e.g. young people dying from heart disease). Create a visualization contrasting this population with the population as a whole and explain how this informs your final target market/product recommendation.

```r
# Pull in both datasets
full_2014 = read_csv("2014_data.csv")
full_2015 = read_csv("2015_data.csv")

# Combine to one large dataset
full_combinded = bind_rows(full_2014, full_2015)

# Drop full_2014/2015
rm(full_2014, full_2015)

# detail_age to numeric
full_combinded = full_combinded %>%
  mutate(detail_age = as.numeric(as.character(detail_age)))

# Changes M/F to numeric values
full_combinded = full_combinded %>%
  mutate(sex_numeric = ifelse(sex == "F", 0, ifelse(sex == 'M', 1, NA)))

# Makes removes patients with no recorded age
full_combinded = full_combinded %>%
  filter(detail_age_type != 9)

full_combinded = full_combinded %>%
  filter(detail_age != 999)

# Makes patients coded for ages less than 1 year of life, and makes them 0
full_combinded = full_combinded %>%
  mutate(detail_age = ifelse(detail_age_type > 1, 0, detail_age))

# Joins the Elixhauser data to the full data set
full_combinded = full_combinded %>%
  left_join(Elixhauser, join_by(icd_code_10th_revision))

# Selects columns of interest
full_obesity_all = full_combinded %>%
  select(icd_code_10th_revision, disease, detail_age)

# Renames NA to allow for comparison
full_obesity_all = full_obesity_all %>%
  mutate(disease =
           ifelse(is.na(disease), "Other Comorbidity", disease))

# Selects disease of interest
full_obesity = full_obesity_all %>%
  mutate(icd_10_filter =
           ifelse(disease == "Obesity", "Obesity", "Other Comorbidity"))
```

```r
#Creates the violin plot
ggplot(full_obesity,
       aes(x = detail_age, y = icd_10_filter, fill = icd_10_filter)) +
  geom_violin(alpha = 0.7) +
  labs(x = "Age of Death", y = "Comorbidity") +
  scale_fill_manual(values = c("navy", "orange"),
                    labels = c("Obesity", "Other")) +
  theme_minimal()
```



This graph demonstrates earlier mortality for patients with documented obesity as a comorbidity.

Obesity is an epidemic that, as a comorbidity, is associated with an early mortality. Cognitive Behavioral Therapy (CBT) has been associated with positive outcomes in obesity management (PMID: 32175002). We recommend an app that targets obese patients in their late 20s, to connect them with CBT certified therapists. Ideally, this early-intervention product will reduce early mortality for patients affected by obesity.