Storytelling with Data

The Failure of Dr. Semmelweis

Bradley Warner and James Wisnowski

May 14, 2019

Contents

Introduction

This information comes from Wikipedia and the majority of the work is based on a project in DataCamp

Dr. Iganz Semmelweis worked at Vienna's General Hospital maternity ward for 3 years from 1846 to 1849. He was the equivalent of a chief resident and assigned the job of assistant to Professor Johann Klein. He prepared patients for the professor, assisted in difficult deliveries, and taught students.

Maternity wards were established as gratis, free, to help with the problem of infanticide of illegitimate children. The government would give free delivery and then make the newborn child a ward of the state in return the women provided training opportunities for the doctors and midwives.

At that time of Dr. Semmelweis' work in the hospital, Europe and North America had a problem with high infant and mother mortality due to an illness known as childbed fever. You must remember that at the time the theory of germs and disease transmittal was in its infancy. At some clinics, the mortality rate could be as high as 40%.

The Problem

There were two clinics at the hospital. The first clinic had a childbed fever mortality rate of almost 10% while the second had a childbed fever mortality of 4%. The clinics were identical except that the first was staffed by medical students while the second was staffed by midwives. In 1823, the doctors had begun the study of anatomy by dissecting cadavers. The midwives did not engage in this practice.

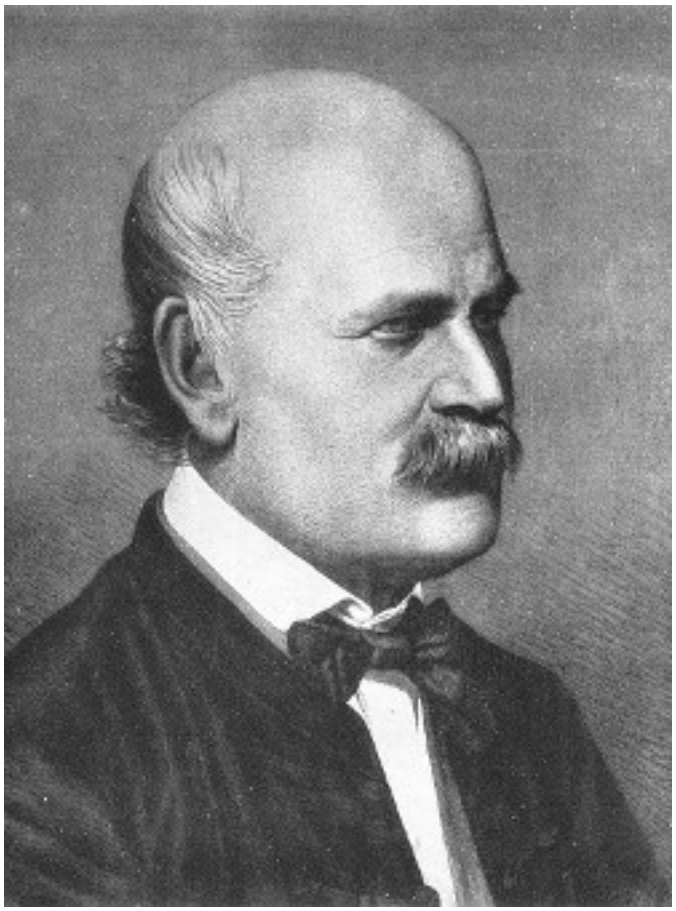


Figure 1: Dr. Semmelweis

The pregnant women were randomly assigned to the different clinics yet they knew the first was more dangerous. They would plead to be placed in the second. Some women would give birth in the street in front of the hospital as this still garnished support for the baby. The mortality rate was lower in the street than in clinic 1.

The Solution

Dr. Semmelweis noticed the problem, ran experiments, and collected data. He did not immediately know the cause. He investigated many different theories, overcrowding, climate, and even the incense used by priests. Ultimately, he found the solution in 1847 when one of the doctors, Jakob Kolletschka, died of childbed fever. This surgeon had accidentally cut his finger with a tool used in conducting an autopsy. Now Semmelweis knew the solution, wash your hands.

There were many publications demonstrating that washing hands reduced mortality rates. Semmelweis also collected data in the clinics where he implemented hand washing which also showed the efficacy of the intervention.

The Reception

Accepted medical practice at the time did not include hand washing. Doctors were considered gentlemen and thus clean. Semmelweis also did not have a scientific theory as to why hand washing worked. Finally, Semmelweis did not present his results in a convincing manner.

Semmelweis was mocked for his work and lost his job at the clinic. In the 1860s, he suffered a nervous breakdown and was committed to a mental institution. He died in after contracting an infection in a wound suffered after a beating from the guards.

More about Dr. Semmelweis can be found at

https://en.wikipedia.org/wiki/Ignaz_Semmelweis

https://en.wikipedia.org/wiki/Historical_mortality_rates_of_puerperal_fever#Yearly_mortality_rates_for_birthgiving_women_1833%E2%80%931858_for_first_and_second_clinics

The Data

In this section, we will look at some of the data that Dr. Semmelweis had to work with.

Prior to Arrival

The following data is the raw mortality data in the two clinics prior to Dr. Semmelweis' arrival. The autopsies were being conducted.

```
# Load in the tidyverse package
library(tidyverse)
# Load fonts
library(extrafont)
```

```
# Read datasets/yearly_deaths_by_clinic.csv into yearly
yearly <- read_csv("datasets/yearly_deaths_by_clinic.csv")

# Print out yearly
yearly
```

```
## # A tibble: 12 x 4
##     year births deaths clinic
##    <dbl>  <dbl>  <dbl> <chr>
##  1  1841   3036    237 clinic 1
##  2  1842   3287    518 clinic 1
##  3  1843   3060    274 clinic 1
##  4  1844   3157    260 clinic 1
##  5  1845   3492    241 clinic 1
##  6  1846   4010    459 clinic 1
##  7  1841   2442     86 clinic 2
##  8  1842   2659    202 clinic 2
##  9  1843   2739    164 clinic 2
## 10  1844   2956     68 clinic 2
## 11  1845   3241     66 clinic 2
## 12  1846   3754    105 clinic 2
```

Web Scrapping

This data can be scrapped from the Wikipedia website. The following code extracts all the data tables from the website.

```
# Load rvest
library(rvest)
```

```
## Loading required package: xml2
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:purrr':
##
##     pluck
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

The URL we need is too long to fit in the output file so we will break it down here

```
https://en.wikipedia.org/wiki/Historical_mortality_rates_of_puerperal_
fever#Yearly_mortality_rates_for_birthgiving_women_1833%E2%80%931858_for_
first_and_second_clinics
```

```
# Semmelweis Wikipedia page
# URL is https://en.wikipedia.org/wiki/Historical_mortality_rates_of_puerperal_
# fever#Yearly_mortality_rates_for_birthgiving_women_1833%E2%80%931858_for_
# first_and_second_clinics
test_url <- "https://en.wikipedia.org/wiki/Historical_mortality_rates_of_puerperal_fever#Yearly_mortali

# Read the URL stored as "test_url" with read_html()
test_xml <- read_html(test_url)

# Print test_xml
test_xml
```

```
## {xml_document}
## <html class="client-nojs" lang="en" dir="ltr">
```

```
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset= ...
## [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-sub ...
```

```
table<-test_xml %>% html_nodes("table") %>% html_table(header=TRUE,fill=TRUE)
table[[1]][1:15,]
```

```
##    Year          Month Births Deaths Rate (%) Notes
## 1  1841   January 1841    254     37     14.6
## 2    NA  February 1841    239     18      7.5
## 3    NA     March 1841    277     12      4.3
## 4    NA     April 1841    255      4      1.6
## 5    NA       May 1841    255      2      0.8
## 6    NA      June 1841    200     10      5.0
## 7    NA      July 1841    190     16      8.4  <NA>
## 8    NA    August 1841    222      3      1.4  <NA>
## 9    NA September 1841    213      4      1.9  <NA>
## 10   NA   October 1841    236     26     11.0  <NA>
## 11   NA  November 1841    235     53     22.6  <NA>
## 12   NA  December 1841     na     na       na  <NA>
## 13 1842   January 1842    307     64     20.8  <NA>
## 14   NA  February 1842    311     38     12.2  <NA>
## 15   NA     March 1842    264     27     10.2  <NA>
```

Cleaning Data

The data does not have the proportion of deaths. Remember this data is the time from when midwives were exclusive to clinic 2 and prior to Semmelweis hand washing experiment.

```
# Adding a new column to yearly with proportion of deaths per no. births
yearly <- yearly %>% mutate(proportion_deaths = deaths/births)

# Print out yearly
yearly
```

```
## # A tibble: 12 x 5
##     year births deaths clinic   proportion_deaths
##    <dbl>  <dbl>  <dbl> <chr>                <dbl>
##  1  1841   3036    237 clinic 1            0.0781
##  2  1842   3287    518 clinic 1            0.158
##  3  1843   3060    274 clinic 1            0.0895
##  4  1844   3157    260 clinic 1            0.0824
##  5  1845   3492    241 clinic 1            0.0690
##  6  1846   4010    459 clinic 1            0.114
##  7  1841   2442     86 clinic 2            0.0352
##  8  1842   2659    202 clinic 2            0.0760
##  9  1843   2739    164 clinic 2            0.0599
## 10  1844   2956     68 clinic 2            0.0230
## 11  1845   3241     66 clinic 2            0.0204
## 12  1846   3754    105 clinic 2            0.0280
```

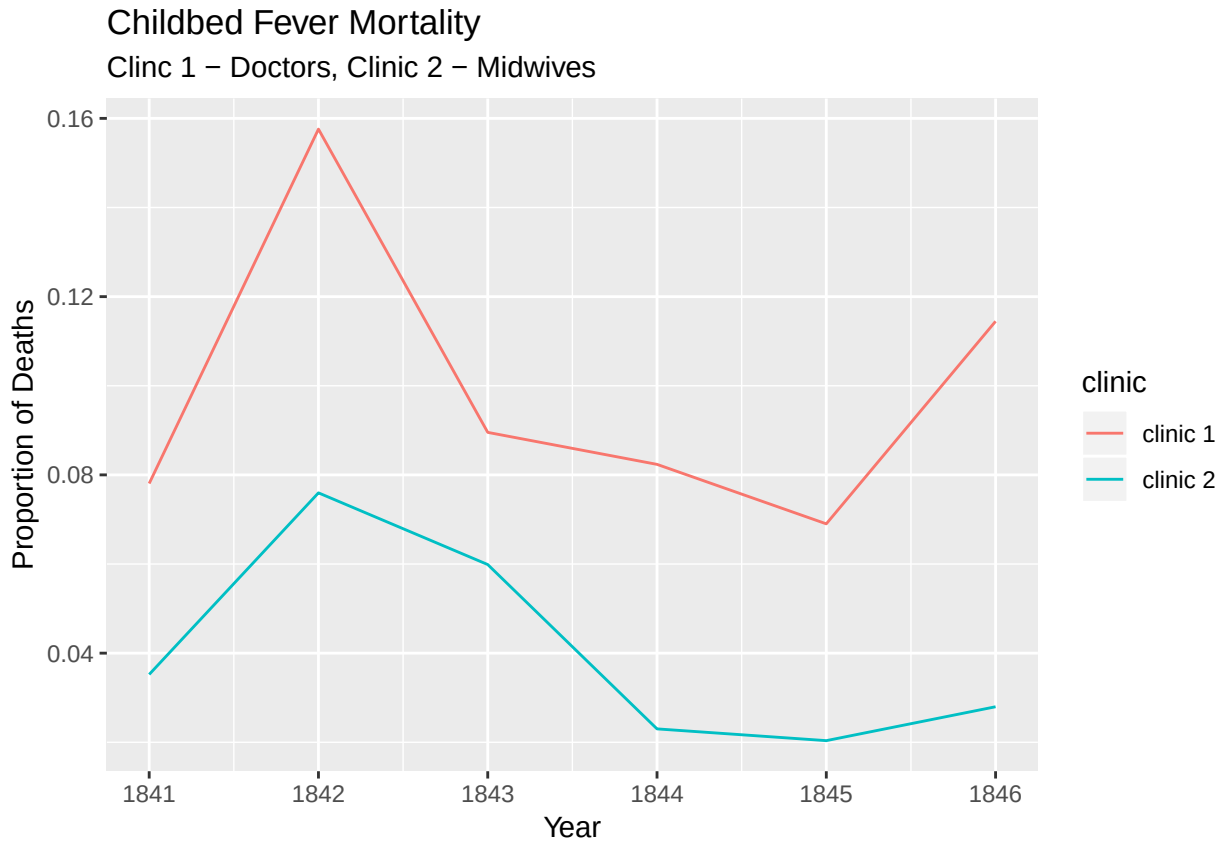
Visualization

The data in a table is too difficult to read. Let's plot the data to better draw conclusions.

Clinic 1 vs Clinic 2

The following plot shows the proportion of deaths in each clinic prior to the Semmelweis hand washing experiment. It is clear that clinic 1 has a higher mortality rate.

```
# Setting the size of plots in this notebook
options(repr.plot.width = 7, repr.plot.height = 4)

# Plot yearly proportion of deaths at the two clinics
ggplot(yearly, aes(x = year, y = proportion_deaths,
    color = clinic)) + geom_line() + labs(x = "Year",
    y = "Proportion of Deaths", title = "Childbed Fever Mortality",
    subtitle = "Clinc 1 - Doctors, Clinic 2 - Midwives")
```



This does show the actual number of women that died!

Monthly Deaths

We will repeat the analysis but this time include the data when Semmelweis started hand washing.

```
# Read datasets/monthly_deaths.csv into monthly
monthly <- read_csv("datasets/monthly_deaths.csv")
```

```
## Parsed with column specification:
## cols(
##   date = col_date(format = ""),
##   births = col_double(),
```

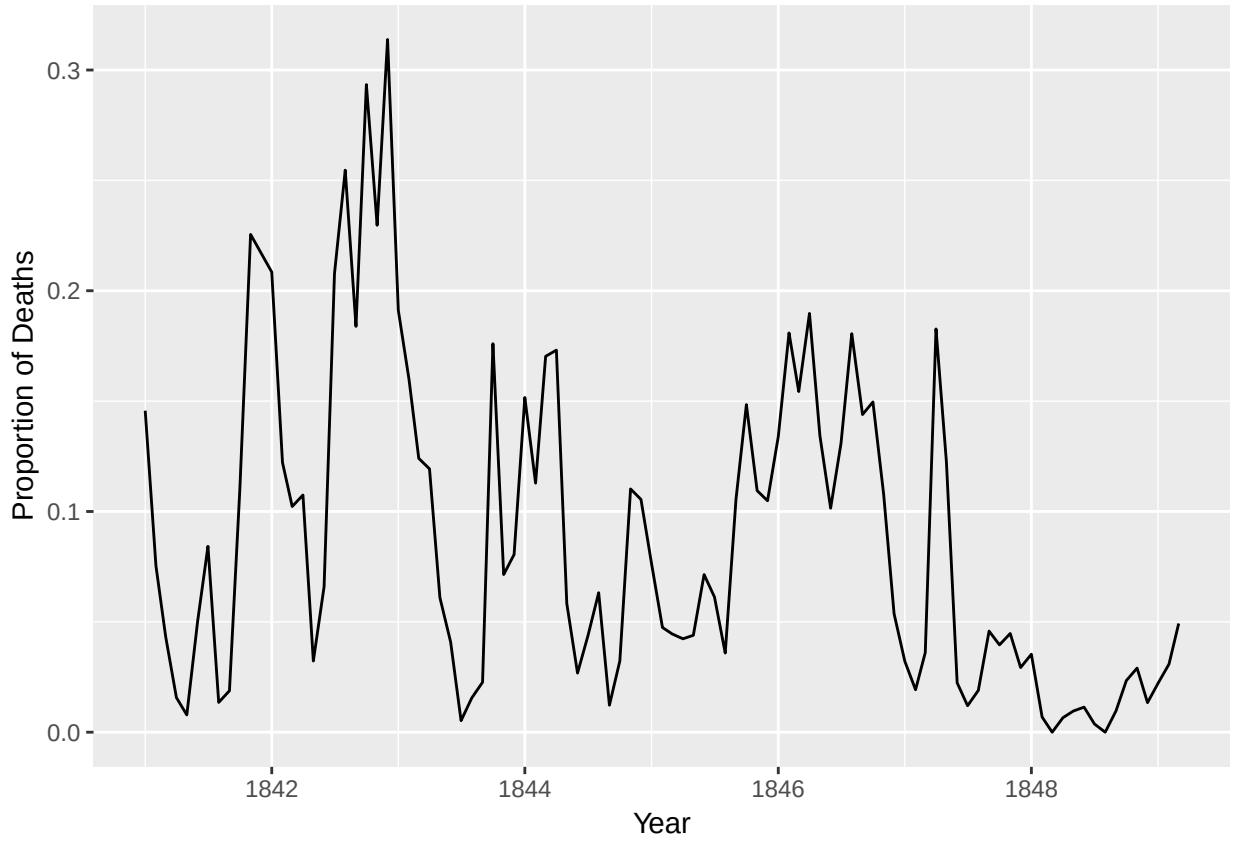
```
##   deaths = col_double()
## )
```

```
# Adding a new column with proportion of deaths per no. births
monthly <- monthly %>% mutate(proportion_deaths = deaths/births)

# Print out the first rows in monthly
head(monthly)
```

```
## # A tibble: 6 x 4
##   date       births deaths proportion_deaths
##   <date>      <dbl>  <dbl>             <dbl>
## 1 1841-01-01    254     37           0.146
## 2 1841-02-01    239     18           0.0753
## 3 1841-03-01    277     12           0.0433
## 4 1841-04-01    255      4           0.0157
## 5 1841-05-01    255      2           0.00784
## 6 1841-06-01    200     10           0.05
```

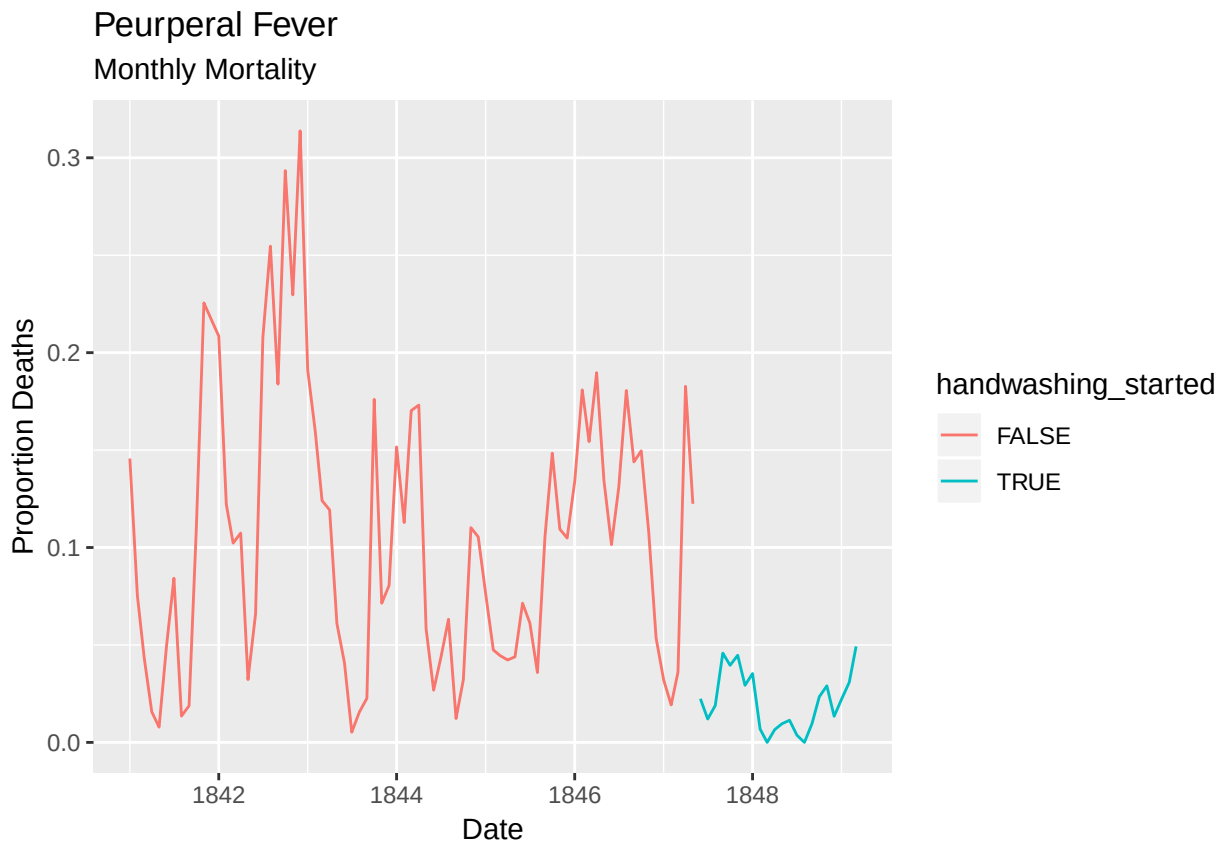
A plot of the monthly mortality rate for the hospital.

```
# Plot monthly proportion of deaths
ggplot(monthly,aes(x=date,y=proportion_deaths)) +
geom_line()+
labs(x="Year",y="Proportion of Deaths")
```



Let's put a marker in when the hand washing took place.

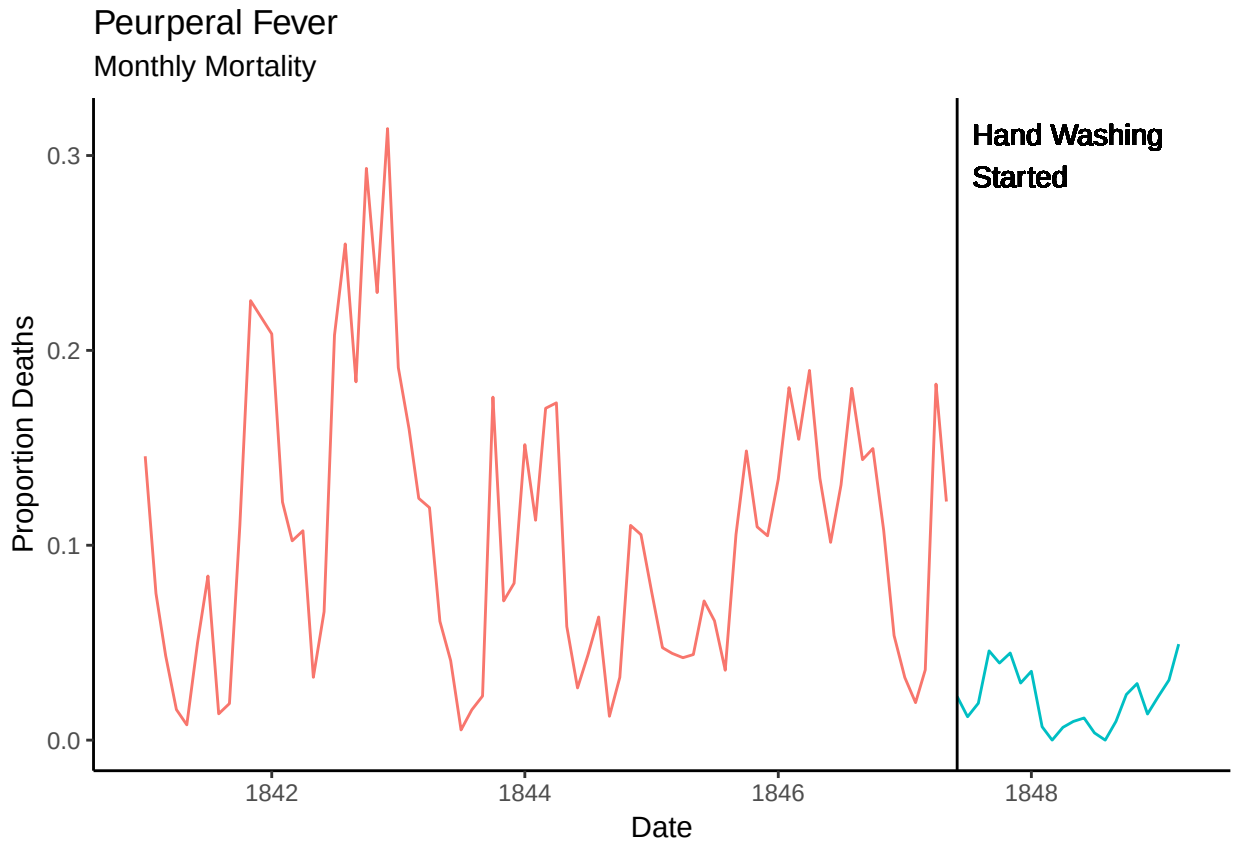
```
# From this date handwashing was made mandatory
handwashing_start = as.Date("1847-06-01")

# Add a TRUE/FALSE column to monthly called
# handwashing_started
monthly <- monthly %>% mutate(handwashing_started = date >=
    handwashing_start)

# Plot monthly proportion of deaths before and
# after handwashing
ggplot(monthly, aes(x = date, y = proportion_deaths,
    color = handwashing_started)) + geom_line() + labs(x = "Date",
    y = "Proportion Deaths", title = "Peurperal Fever",
    subtitle = "Monthly Mortality")
```



Let's clean it up. We removed the border and background. We used color and the vertical line to focus attention. Instead of a legend, we used text to highlight the change. We allow white space. We connected the data points.

```
# Plot monthly proportion of deaths before and
# after handwashing
ggplot(monthly, aes(x = date, y = proportion_deaths,
    color = handwashing_started)) + geom_line() +
    geom_vline(xintercept = as.Date("1847-06-01")) +
    labs(x = "Date", y = "Proportion Deaths", title = "Peurperal Fever",
        subtitle = "Monthly Mortality") +
    theme_classic() +
```

```
    theme(legend.position = "none") +
    geom_text(x = as.Date("1847-07-15"),
    y = 0.3, label = "Hand Washing \nStarted", color = "black",
    hjust = 0)
```



Summarizing the Data

Let's summarize the data both prior and after hand washing.

```
# Calculating the mean proportion of deaths
# before and after handwashing.

monthly_summary <- monthly %>%
group_by(handwashing_started) %>%
summarise(mean_proportion_deaths=mean(proportion_deaths))

# Printing out the summary.
monthly_summary
```

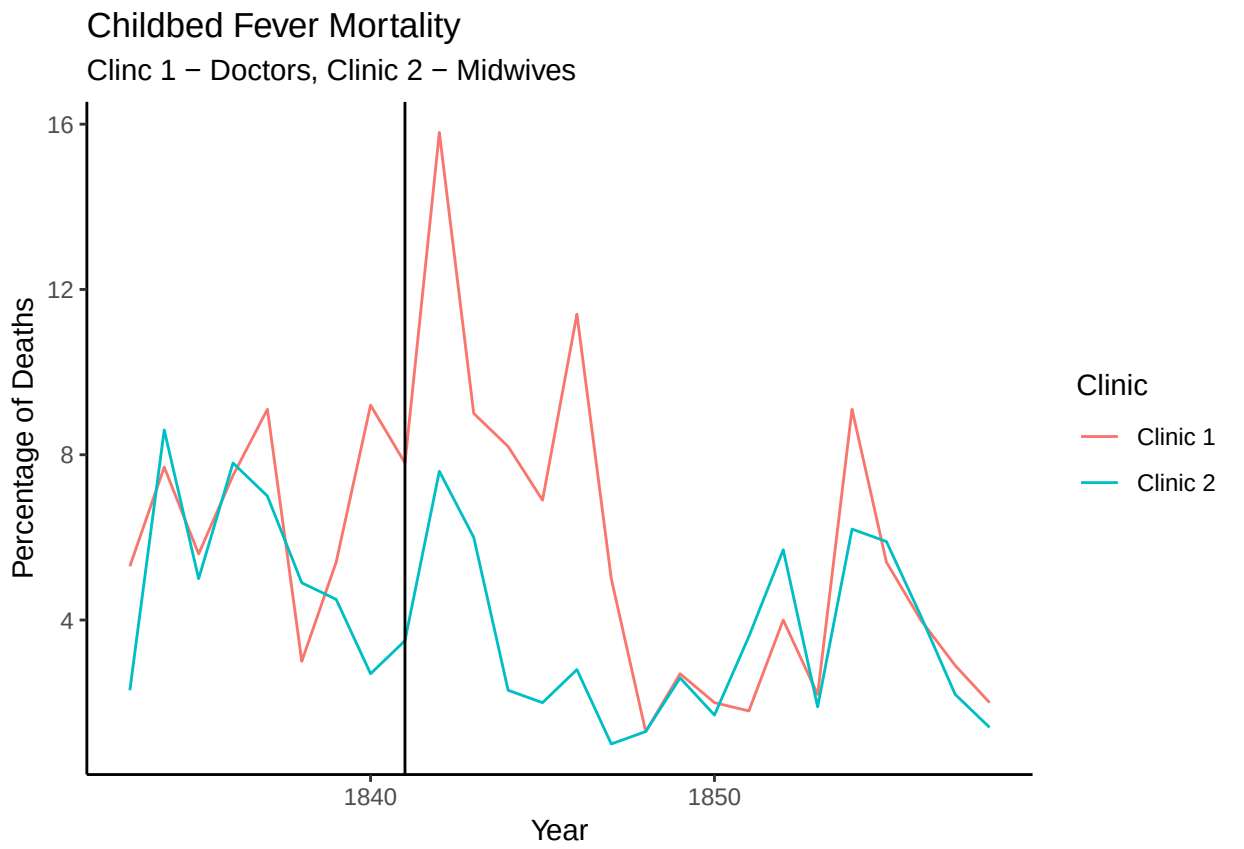
```
## # A tibble: 2 x 2
##   handwashing_started mean_proportion_deaths
##   <lgl>                                <dbl>
## 1 FALSE                               0.105
## 2 TRUE                                0.0211
```

Better Comparison

```
full_year<-as.data.frame(table[[2]][2:27,1:4])
names(full_year)<-c('Year','Births','Deaths','Rate')
temp<-table[[2]][2:27,c(1,6,7,8)]
names(temp)<-c('Year','Births','Deaths','Rate')
full_year<-rbind(full_year,temp)
rm(temp)
full_year$Clinic<-rep(c("Clinic 1","Clinic 2"),each=26)
full_year$Year<-as.integer(full_year$Year)
full_year$Rate<-as.double(full_year$Rate)
full_year$Births<-as.numeric(gsub(",","",full_year$Births))
head(full_year)
```

```
##   Year Births Deaths Rate   Clinic
## 2 1833   3737    197  5.3 Clinic 1
## 3 1834   2657    205  7.7 Clinic 1
## 4 1835   2573    143  5.6 Clinic 1
## 5 1836   2677    200  7.5 Clinic 1
## 6 1837   2765    251  9.1 Clinic 1
## 7 1838   2987     91  3.0 Clinic 1
```

```
# Plot yearly proportion of deaths at the two
# clinics
ggplot(full_year, aes(x = Year, y = Rate, color = Clinic)) +
    geom_line() + labs(x = "Year", y = "Percentage of Deaths",
    title = "Childbed Fever Mortality",
    subtitle = "Clinc 1 - Doctors, Clinic 2 - Midwives") +
    geom_vline(xintercept = 1841) + theme_classic()
```



Long Term Look

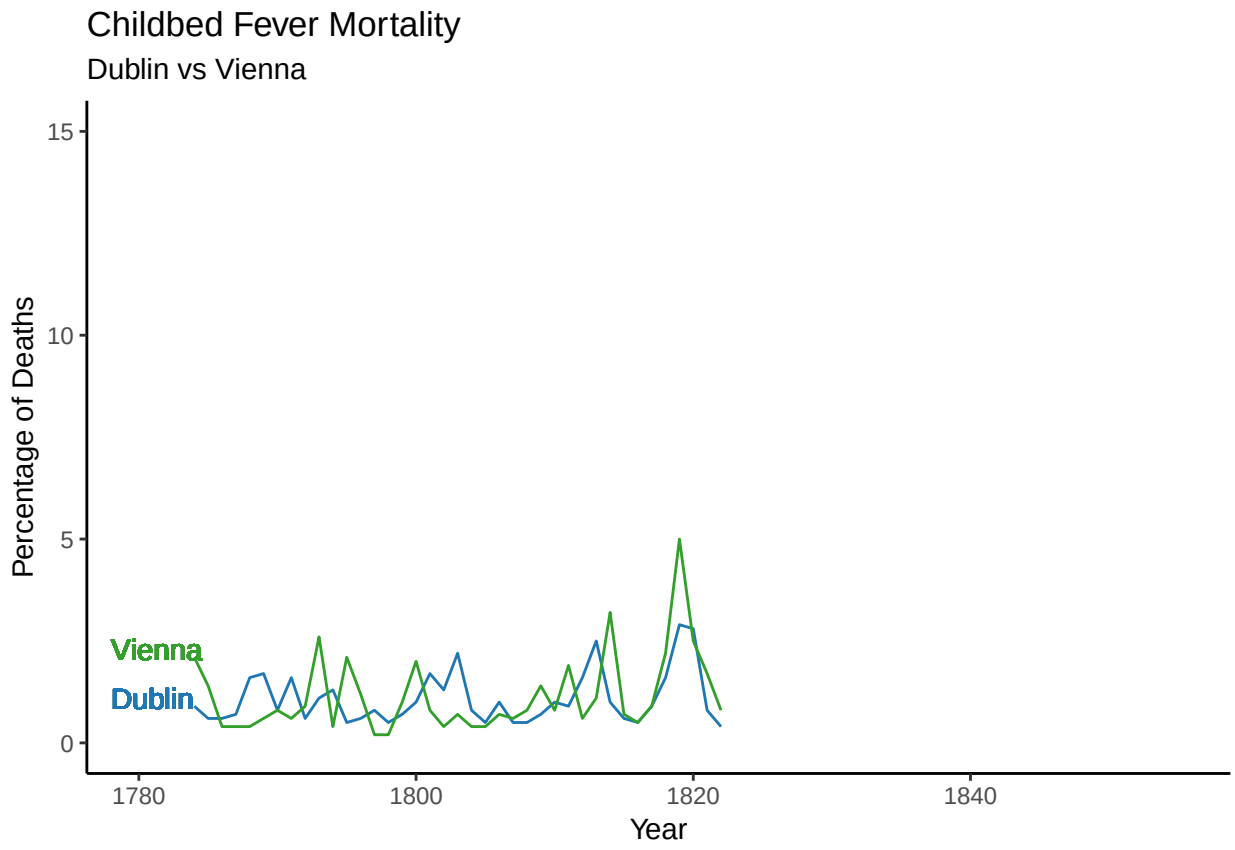
Get the data from Vienna hospital and Dublin maternity ward. In Dublin the doctors did not work in the morgue.

```
long_data <- as.data.frame(table[[3]][,1:4])
long_data$Births<-as.numeric(gsub(",","",long_data$Births))
long_data <- rbind(long_data,as.data.frame(table[[5]]))
long_data$Clinic<-rep(c("Vienna","Dublin"),each=66)
names(long_data)[4]<-"Rate"
head(long_data)
```

```
##   Year Births Deaths Rate Clinic
## 1 1784    284      6  2.1 Vienna
## 2 1785    899     13  1.4 Vienna
## 3 1786   1151      5  0.4 Vienna
## 4 1787   1407      5  0.4 Vienna
## 5 1788   1425      5  0.4 Vienna
## 6 1789   1246      7  0.6 Vienna
```

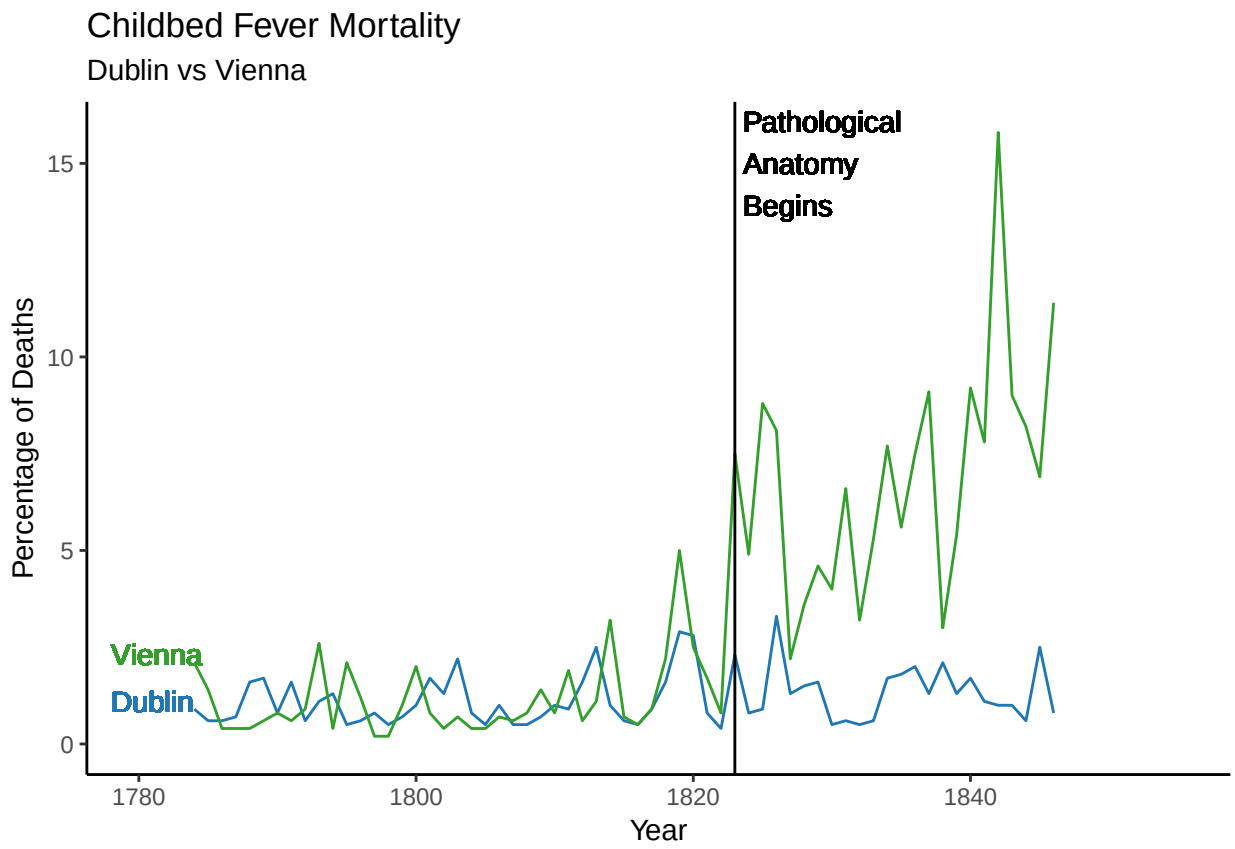
```
# Plot yearly proportion of deaths at the two hospitals
long_data %>% filter(Year <= 1822) %>% ggplot(aes(x = Year,
    y = Rate, color = Clinic)) + geom_line() + labs(x = "Year",
    y = "Percentage of Deaths", title = "Childbed Fever Mortality",
    subtitle = "Dublin vs Vienna") +
  theme_classic() +
```

```
  scale_color_manual(values = c("#1f78b4", "#33a02c")) +
  geom_text(x = 1778, y = 2.3, label = "Vienna",
        color = "#33a02c", hjust = 0) + geom_text(x = 1778,
    y = 1.1, label = "Dublin", color = "#1f78b4", hjust = 0) +
  expand_limits(x = c(1780, 1855), y = c(0, 15)) +
  theme(legend.position = "none")
```

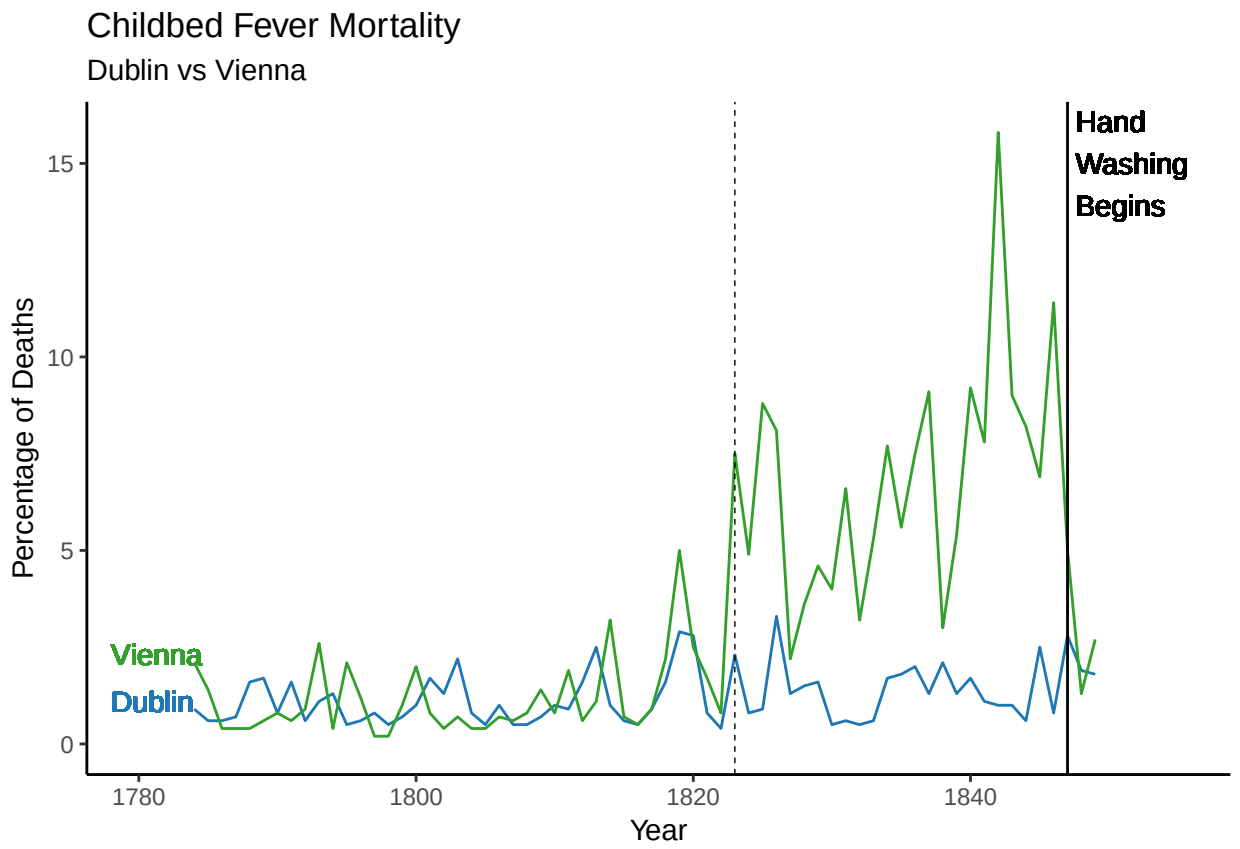


The use of doctors to perform autopsies at the Vienna clinic starts in 1823.

```
# Plot yearly proportion of deaths at the two hosptials
long_data %>% filter(Year <= 1846) %>% ggplot(aes(x = Year,
    y = Rate, color = Clinic)) + geom_line() + labs(x = "Year",
    y = "Percentage of Deaths", title = "Childbed Fever Mortality",
    subtitle = "Dublin vs Vienna") + geom_vline(xintercept = 1823) +
  theme_classic() +
  scale_color_manual(values = c("#1f78b4","#33a02c")) +
  geom_text(x = 1778, y = 2.3, label = "Vienna",
    color = "#33a02c", hjust = 0) +
  geom_text(x = 1778, y = 1.1, label = "Dublin", color = "#1f78b4", hjust = 0) +
  geom_text(x = 1823, y = 15, label = " Pathological\n Anatomy\n Begins",
        hjust = 0, color = "black") +
  expand_limits(x = c(1780,1855), y = c(0, 15)) +
  theme(legend.position = "none")
```




```
# Plot yearly proportion of deaths at the two hospitlas
long_data %>% filter(Year <= 1850) %>%
ggplot(aes(x=Year,y=Rate,color=Clinic))+
geom_line() +
  labs(x="Year",y="Percentage of Deaths",title="Childbed Fever Mortality",
       subtitle="Dublin vs Vienna") +
  geom_vline(xintercept = 1823,size=.25,linetype=2)+
  geom_vline(xintercept = 1847) +
  theme_classic() +
  scale_color_manual(values=c('#1f78b4','#33a02c'))+
  geom_text(x=1778, y=2.3, label="Vienna",color='#33a02c',hjust=0) +
  geom_text(x=1778, y=1.1, label="Dublin",color='#1f78b4',hjust=0)+
  geom_text(x=1847,y=15,label=' Hand\n Washing\n Begins',hjust=0,color='black') +
  expand_limits(x=c(1780,1855),y=c(0,15))+
  theme(legend.position = "none")
```



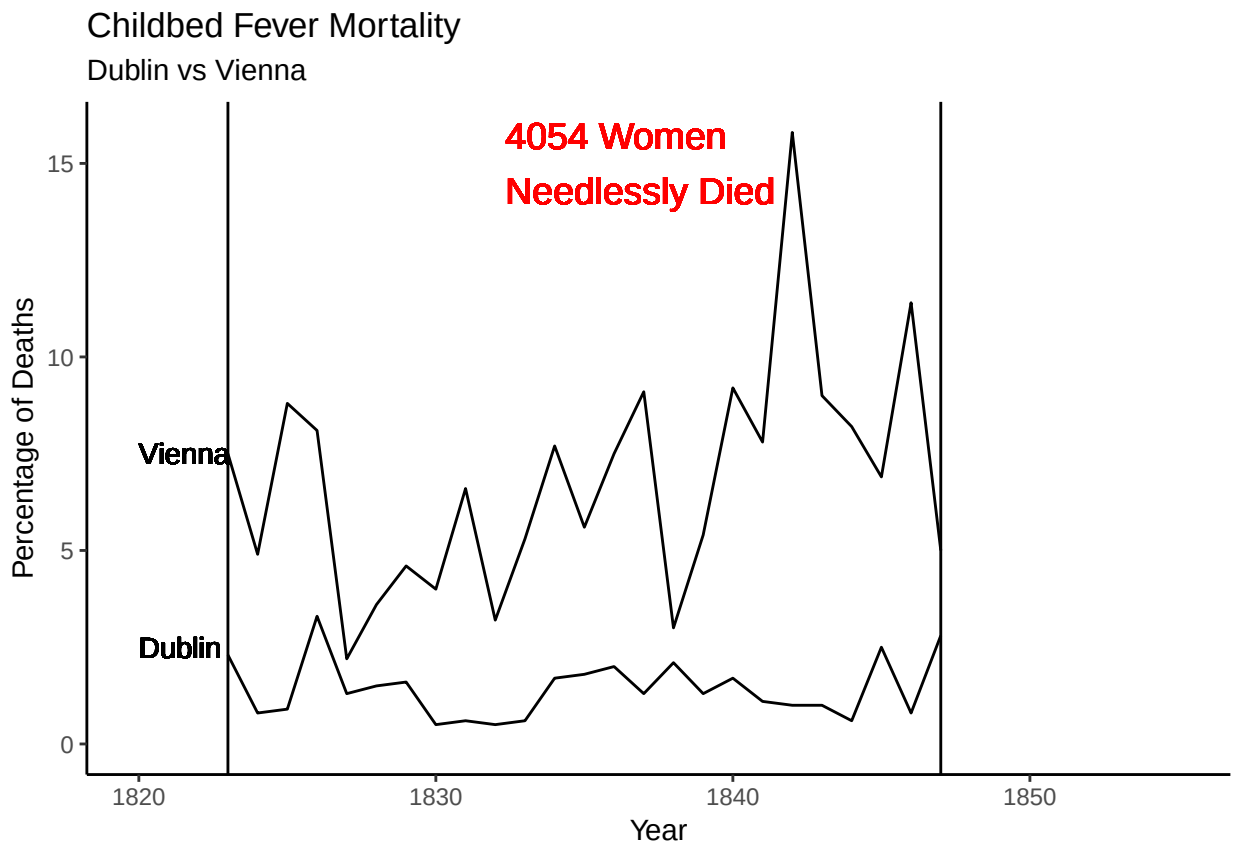
How many women died in excess of what would be expected if hands had been washed from 1823 until 1847?

```
tempr<-long_data %>% filter(Year>=1823,Year<1847,Clinic=='Dublin') %>% select(Rate)
tempb<-long_data %>% filter(Year>=1823,Year<1847,Clinic=='Vienna') %>% select(Births)
tempd<-long_data %>% filter(Year>=1823,Year<1847,Clinic=='Vienna') %>% select(Deaths)
sum(tempd-tempb*tempr/100)
```

```
## [1] 4053.676
```

That is 4054 more women died than is expected because doctors did not wash their hands.

```
# Plot yearly proportion of deaths at the two clinics
long_data %>% filter(Year <= 1847,Year >= 1823) %>%
ggplot(aes(x=Year,y=Rate,color=Clinic))+
geom_line() +
  labs(x="Year",y="Percentage of Deaths",title="Childbed Fever Mortality",
       subtitle="Dublin vs Vienna") +
  geom_vline(xintercept = 1823)+
  geom_vline(xintercept = 1847) +
  theme_classic() +
  scale_color_manual(values=c('black','black'))+
  geom_text(x=1820, y=7.5, label="Vienna",color='black',hjust=0) +
  geom_text(x=1820, y=2.5, label="Dublin",color='black',hjust=0)+
  geom_text(x=1832,y=15,label=' 4054 Women\n Needlessly Died',hjust=0,color='red',size=5) +
  expand_limits(x=c(1820,1855),y=c(0,15))+
  theme(legend.position = "none")
```



Conclusion

Semmelweis was not able to convince his peers and superiors of his findings. This was due to a problem of communication and publication. Ultimately he was proven to be correct but how many women needless died in the intervening period?