Death in New York City:  
The leading causes in 2007 and 2014 and the development through time

Ephraïm Veder 2603454

Sarra Souissi 2630604

Ingeborg van der Stuijt 2573043

October 22, 2018

Dr. Erika Kuijpers

VU Amsterdam

Digital Humanities & Social Analytics

Table of Contents

Introduction Page 2

Our Dataset and Methodology Page 5

Our Results Page 7

Conclusion Page 10

Recommendations for Further Research Page 12

Bibliography Page 13

Appendix Page 15

Introduction

**“Life is for the living. Death is for the dead. Let life be like music. And death a note unsaid.”**

* **Langston Hughes**

Death is inevitable. Death is natural. However, natural causes can be prevented. Dying a natural death doesn’t always mean someone’s time has simply come. The time one has to live, can easily be extended or cut short. Diseases, accidents and tragedies cost many lives, but what costs the most?

Our research is about the cause of death in New York City between 2007 and 2014, we want to know how it has changed over time and why. New York City is a representative city, since it has inhabitants of many cultural backgrounds, people of different ages, sexes and income levels. We can also compare the results of our data analysis to bigger researches, for example data of America as a whole or the world due to the diversity of the New York City Area. We can also compare our data to data of different years, to hopefully see bigger differences. It was also very helpful for our research that this dataset was fully accessible and complete.

More specifically we are interested in seeing the differences between males and females. There is not a lot of research on this topic. We think there could be important new findings because there are inherit biological differences, that could lead to higher risks of getting certain diseases. An article by Möller-Leimkühler (2002) describes one of these gender differences. “Suicide and premature death due to coronary heart disease, violence, accidents, drug or alcohol abuse are strikingly male phenomena” (Möller-Leimkühler, 2002). Moreover, due to the lack of research on the correspondence between gender and death rate, we decided to take this into consideration in our research.

It is very important to look at data that give us information about causes of death and there are many fields that focus on this kind of research, such as heart disease organisations and hospitals. The World Health Organisation is an organisation that is focused on the health of people from around the globe. They keep track of diseases and changes in health of the population worldwide. Of course, the World Health Organisation also wrote about causes of death. This is what they wrote on the importance of research about causes of death:

“Cause-of-death statistics help health authorities determine the focus of their public health actions. A country in which deaths from heart disease and diabetes rise rapidly over a period of a few years, for example, has a strong interest in starting a vigorous programme to encourage lifestyles to help prevent these illnesses. Similarly, if a country recognizes that many children are dying of pneumonia, but only a small portion of the budget is dedicated to providing effective treatment, it can increase spending in this area.” (World Health Organization, 2018)

To write our hypothesis we searched for articles that described the main causes of death worldwide. Since this is an average, we would expect to see these back in our data. This is helpful to get a better understanding where our data fits in with this as well. Some of our most interesting finds in literature are stated below:

The World Health Organisation (WHO) wrote the following:

“Of the 56.9 million deaths worldwide in 2016, more than half (54%) were due to the top 10 causes. Ischaemic heart disease and stroke are the world’s biggest killers, accounting for a combined 15.2 million deaths in 2016. These diseases have remained the leading causes of death globally in the last 15 years.”  
(World Health Organization, 2018)

The Stroke is a magazine that gave us more insight in if the results we get from datasets about causes of death really help us by preventing diseases. They state:

“The prevention by health care systems work! For example, the recent rank decline of stroke among leading causes of American death is testament to a half century of societal progress in cerebrovascular disease prevention and acute care. Renewed commitments are needed to preserve and broaden this historic achievement.”

(Towfighi & Saver in Stroke, 2011)

And so we asked ourselves if we would see the same results this literature would let us suspect. Does our dataset show a decline in the major causes of death as well? Our hypothesis was based on the assumption that we would find:

“The top 6 causes of death in NYC declined during the time period of 2007-2014”

Our research question was based on this hypothesis, with the main question being:

“How have the top 6 causes of death in NYC changed during the time period of 2007 and 2014?”

We have divided this question into several sub-questions so that it would have been easier to answer. This also formed the lay-out of our research.

We asked ourselves the following sub-questions:

* What are the top 6 leading causes of death in both 2007 and 2014 for women?
* What are the top 6 leading causes of death in both 2007 and 2014 for men?
* What does the comparison of both women and men say about the main cause of death?
* And how can we explain the change we see?

Our dataset and methodology

The data we used for our research is from NYC OpenData. This is a website owned by the City of New York. The data itself is provided by the Department of Health and Mental Hygiene. It is a quantitative dataset that describes all recorded deaths between 2007 and 2014 in New York City per year. There are 1094 rows in total. The dataset is in a CSV excel-format, which is common for large datasets like this one. There are seven variables used:

1. Year of death

2. Leading cause

3. Sex

4. Race Ethnicity

5. Deaths (in numbers)

6. Death Rate

7. Age Adjusted Death Rate

To answer the questions of this research, we modified the dataset. The original file holds the information of all the years between 2007 and 2014, so we isolated only the first and last year for we wanted to see what had changed over this large span of time.

To answer our question regarding gender we had to combine the data in a way that the causes of death are not divided by race, age or any other variables than sex. Also, as a limitation of our original dataset was that the column for “Death Rate” was separated by race, it was invaluable to our specific research questions. We therefore had to use total number of deaths to curate our results. Using Microsoft Excel and the SUMIF function, we were able to clean up the data by getting rid of duplicate “Causes of Death” that were a part of the original dataset due to the separation by race. This is explained more thoroughly in the Data Documentation (see Appendix).

To answer the questions for both sexes combined we added up all the deaths for 2007 and 2014 for an easy comparison.

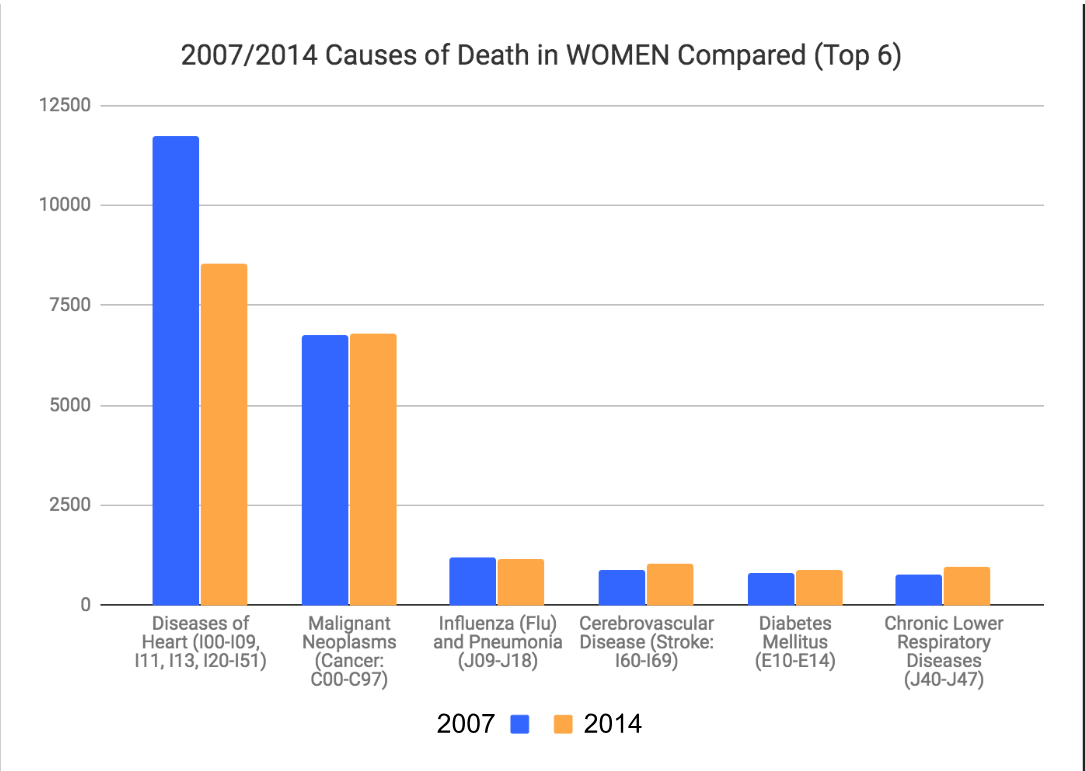
To visualize the data we wanted, we made bar charts of the top 6 causes of death. The charts show the 6 main causes of death with two bars, one for 2007 and another for 2014. First we did this for males and females separately and after that we made a combined chart.

Since the data comes from the New York State OpenSource data hub which is provided by the government in New York City, it is reasonable to argue that the provenance of this data is reputable. However, there could, perhaps be potential biases. Because the data comes from the department of health, there could be good reasons to skew the data. They are responsible for the health of their citizens, they would like to show that they are doing a good job at this. So they could have manipulated the data to show declines in the top causes of death for their own advantage.

Another potentially problematic point to address about this dataset is the generalizations of the causes of death. For example, a cause of death according to the data source is “Diseases of the Heart” rather than explicit heart conditions. Another example of this that for a cause of death that cannot be categorized, it is thrown into the category “All Other Causes” which is extremely vague and not useful. This also can lead to misrepresentation which must also be considered when dealing with the original dataset.

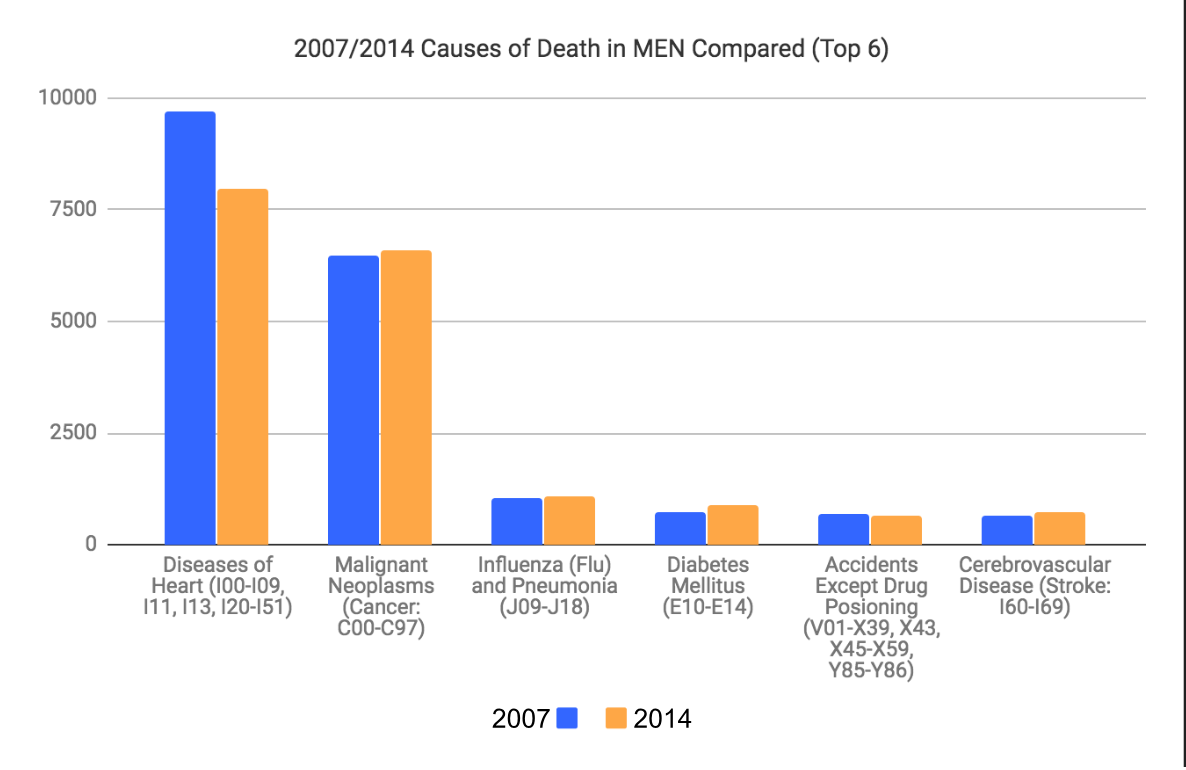
Our results

Using the methods we described previously to curate the original spreadsheet, we created charts:

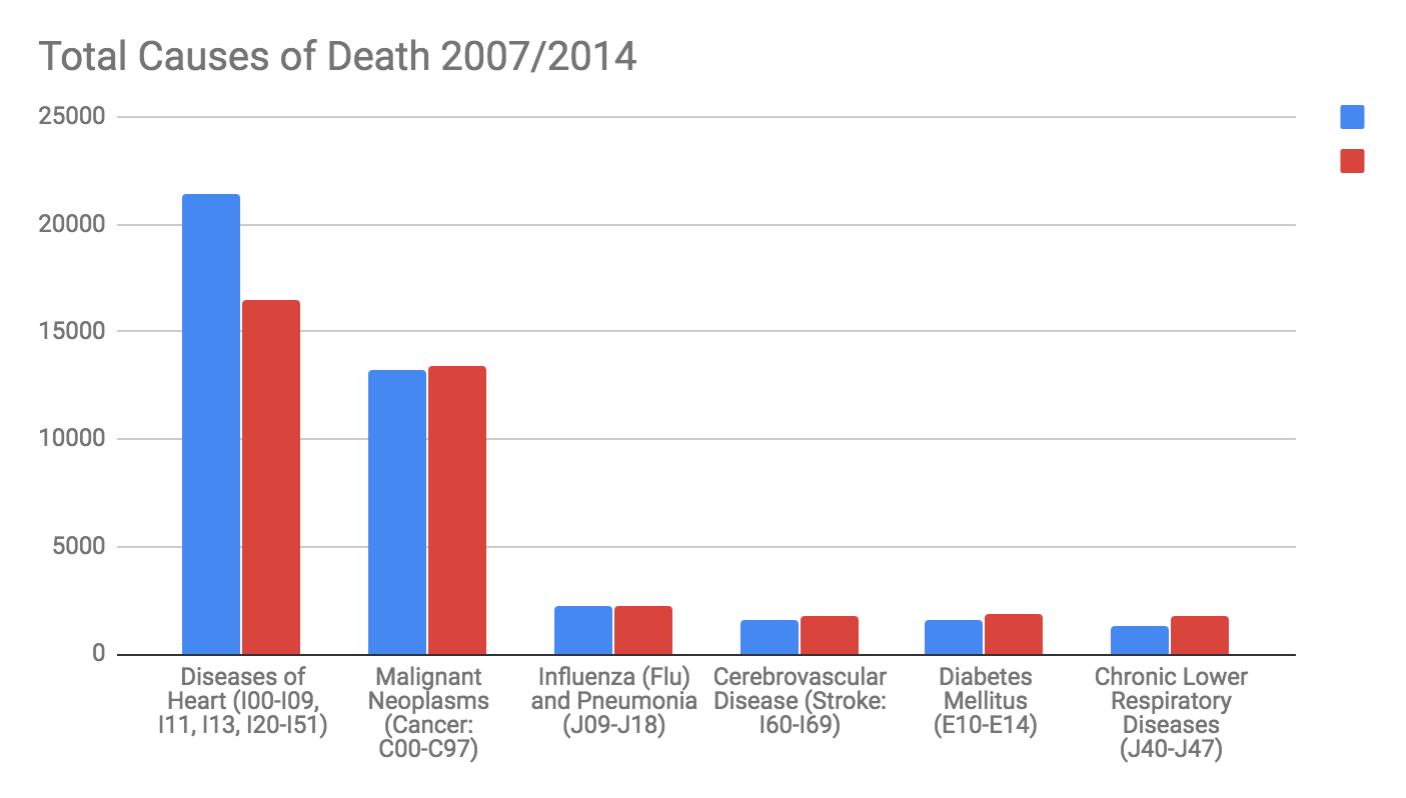


This chart shows the following results:

* The leading cause of death in both 2007 and 2014 for women is Diseases of the Heart
* There was a significant decline in death by Disease of the Heart in women from 2007 to 2014
* Compared to the rest of the top 5 causes of death in women, Disease of the Heart is the greatest cause of death and has the most significant change over time.



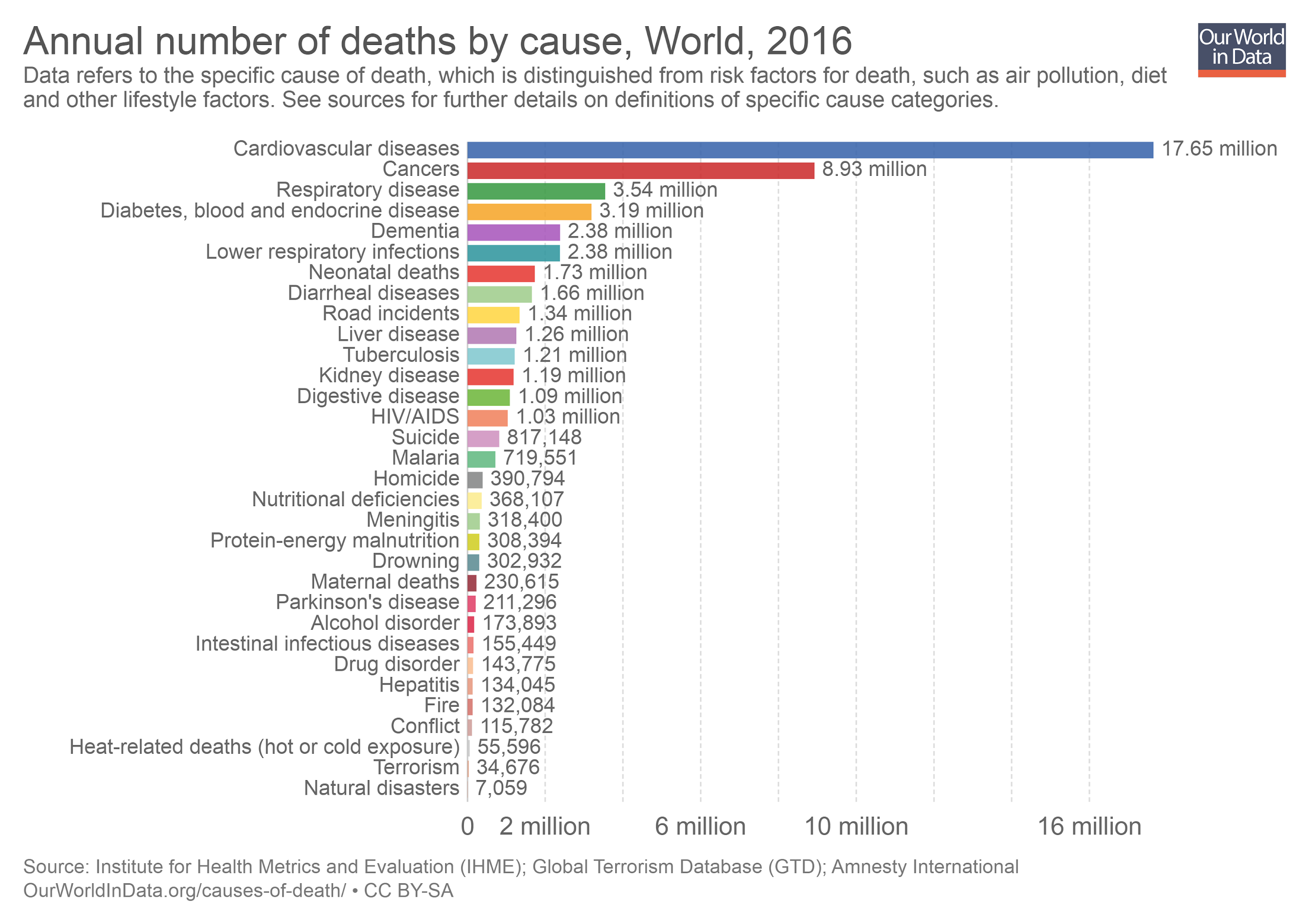
The chart for men shows similar results as for women. However, when we examine the Y-axis of both bar graphs, we see that significantly more women are impacted by Diseases of the Heart than men.



The results of the top six causes of death in total, when combining the results of men and women, again show similar results. The leading cause of death, despite the change over time is Diseases of the Heart and there has been an overall drastic decline in these deaths over the seven year period.

Because of the severity of the change in this one specific cause of death, we decided to focus the remainder of our research on this specific result. Also, our conclusions shift to finding out more about Diseases of the Heart, why they are the leading cause of death in New York City by large, and the significant decline in deaths by this cause despite it remaining the leading cause of death. Moreover, we spend some time discussing the significance of sex and why women are dying at a

higher rate than men from Diseases of the heart.

It is interesting to see the same results in the annual number of deaths by cause in this graph about the whole world population in 2016. This shows us that our results are comparable to the world numbers and New York City indeed is representative for the world population as we suspected in our introduc

source: Ritchie & Roser, Our World In Data (2018)

Conclusion

Our data show a pretty clear result: The main causes of death have overall stayed the same through the years 2007-2014. The top cause of death, which is heart disease, has declined for both genders but remains to be the biggest cause of death. Heart disease is an overarching concept and includes many different conditions that are be damaging for the heart, for example coronary artery disease, arrhythmia, cardiomyopathy, and heart failure. This was what we expected to find since it is the same result that research has found in the rest of the world before these years and also up to recent years. The other five causes of death all show change over time as well, but those are very minor differences.

We found that there were more women affected by heart disease than men. We wanted to know why, and found very few, controversial results. A couple articles found through research that heart disease risks rise after menopause. Some experts believe that naturally occuring hormones protect young women from heart disease.

A way to solve this is also described: “Many observational studies have found lower rates of coronary heart disease (CHD) in women who take postmenopausal estrogen than in women not receiving this therapy.” (Stephen Hulley et al., 1998) But there was also an article that said: “Since 1984, more women have died of heart disease than men each year, although that is partly because women generally live longer than men.” (Belchetz, 2018). So it is not very clear why women had a higher risk according to our research.

We also found it interesting to understand why there is such a decline in heart diseases. The main reasons that heart disease is the leading cause all over the world are obesity and smoking. This is why there are a lot of preventative measures taken to bring back the amount of cases that can be traced back to either obesity and smoking in the last few years, which we expect are a direct cause of the decline. There is an analysis of Begin Unal c.s. that studied the number of deaths caused by coronary heart diseases in the UK during the period 1981 - 2000. They state that there was a drop of 68.230 cases, which is a decrease of 54 percent. Approximately half of these cases are prevented by primary prevention such as quitting smoking, reducing blood pressure and the level of cholesterol in the entire population. The primary prevention was four times more effective than the secondary prevention. This made it clear that there should be a population wide prevention strategy involving the tobacco use and encouraging a healthy lifestyle (Medisch Contact, 2005). However, the decline in heart disease related deaths could even be greater than it is. The number of deaths due to heart diseases is declining because of preventative treatment, according to Koopman (UMC Utrecht, 2015), because the number of people that smoke is declining and people get medicines to help control their cholesterol sooner. However, there are factors that counterfact as well, such as the increasing number of people that suffer from diabetes, mostly due to obesity. According to Koopman (UMC Utrecht, 2015) the number of people that died from heart diseases during the time period of 1997-2007 decreased by 11.000 people. However, she states that this number could be over 12.000 if obesity and diabetes hadn’t increased this much.

In recent years there are more sanctions being taken to take away the power of brands and make sure the awareness is the highest it can be. This is very apparent with the relatively new cigarette covers, that have been implemented in multiple countries. These show that cigarettes lead to many complications including heart disease.

Recommendations for further research

Because our research only uses data from one city in one specific country, it’s external validity is a problem. It would be very useful to use multiple datasets that research the same timeframe as we did but from different countries that are in different continents. This way you could find out what changes have occurred in different parts of the world, to see if they match for example the living condition of the population.

Another idea for future research could be to use a bigger timeframe. For example from 1960 to now. In 1964 an important research was published by by the Surgeon General's Advisory Committee on Smoking and Health, it discussed the negative health effects of smoking. This report has had a huge influence on the tobacco industry. There have been other major changes in society since then so it would be more interesting to know if differences in technology and health awareness made an impact on the causes of death.

Bibliography

Belchetz, B. (7 February 2018). *The biggest myths about heart attacks that need to be busted.* Marilyn. Available on 20-10-2018 from

<https://www.marilyn.ca/Health/Articles/February2018/The-biggest-myths-about-heart-attacks-that-need-to>

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., ... & Heart and Estrogen/progestin Replacement Study (HERS) Research Group. (1998).

Medisch Contact. (26 August 2005). Hartziekten: primaire preventie werkt. Available on 10-10-2018 from

<https://www.medischcontact.nl/nieuws/laatste-nieuws/artikel/hartziekten-primaire-preventie-werkt.htm>

Möller-Leimkühler, A. M. (2003). The gender gap in suicide and premature death or: why are men so vulnerable?. European archives of psychiatry and clinical neuroscience, 253(1), 1-8.

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., ... & Heart and Estrogen/progestin Replacement Study (HERS) Research Group. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Jama, 280(7), 605-613.

Ritchie, H. & Roser, M. (February 2018). *Cause of Death.* Our World in Data. Available on 20-10-2018 from

<https://ourworldindata.org/causes-of-death>

Towfighi, A. & Saver, J.L. (21 july 2011). Stroke Declines From Third to Fourth Leading Cause of Death in the United States: Historical Perspective and Challenges Ahead. *Stroke*, 42, 2351-2355. Available on 19-9-2018 from<https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.111.621904>

UMC Utrecht. (2 April 2015). Overgewicht remt succes aanpak hartziekten. Available on 10-10-2018 from <https://www.umcutrecht.nl/nl/Ziekenhuis/Afdelingen/Hart-en-vaatcentrum/Nieuws/Overgewicht-remt-succes-aanpak-hartziekten>

WHO. (24 May 2018). The Top 10 Causes of Death. Available on 19-9-2018 from

<http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>

Appendix

Data Documentation

* **The provenance of the data:**

<https://data.cityofnewyork.us/d/jb7j-dtam?category=Health&view_name=New-York-City-Leading-Causes-of-Death>

The original dataset used for this research was from the New York City Open Data hub provided by the City of New York as a public source of large datasets about the city. From this we were able to access an exhaustive list of the causes of death in New York City from 2007 until 2014. Since the data comes from the New York State OpenSource data hub which is provided by the government in New York City, it is reasonable to argue that the provenance of this data is reputable. However, there could, perhaps be potential biases. Because the data comes from the department of health, there could be good reasons to skew the data. They are responsible for the health of their citizens, they would like to show that they are doing a good job at this. So they could have manipulated the data to show declines in the top causes of death for their own advantage.

* **The data model we used:**

We chose to represent our data using bar graphs to show change over time. We used bar graphs as opposed to a line graph or another traditional “change-over-time” data model because of our choice to only compare two years (2007 and 2014). Because of this, by using a bar graph, there is a clear visual presentation of whether or not there was a change between the years 2007 or not and the severity of the change.

* **The curation of the data you carried out and the choices made:**

The goal of our research was mainly to see both if there was a change over time in the leading causes of death in men and women between the years 2007 and 2014 (note: we used these years due to the limitations of our dataset only providing death rates from the years 2007-2014) as well as to see if there was any significant difference in the leading causes of death for men and women separately.

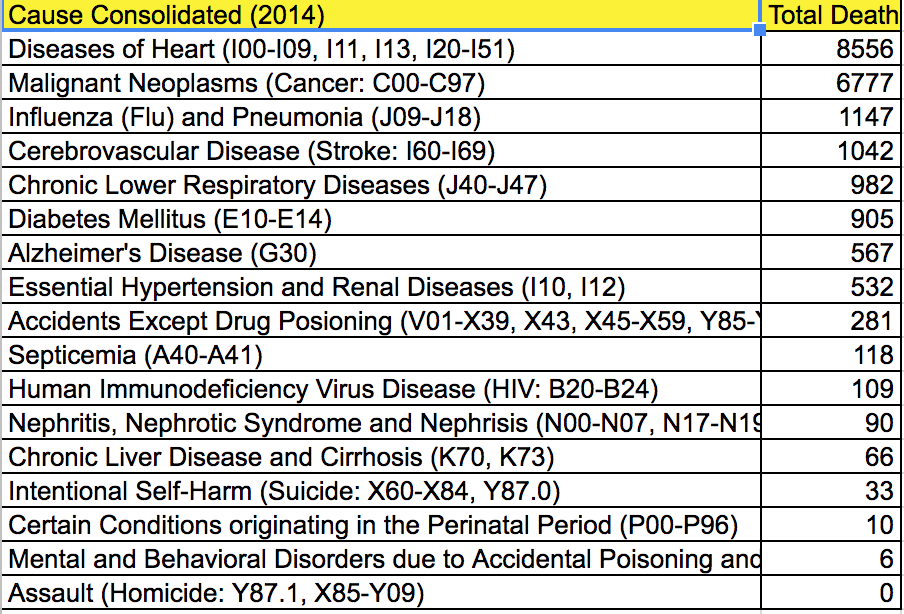
From the original dataset the steps taken to curate the dataset for our needs are as follows:

1. Using Microsoft excel, the data for the years 2007 and 2014 were isolated respectively.

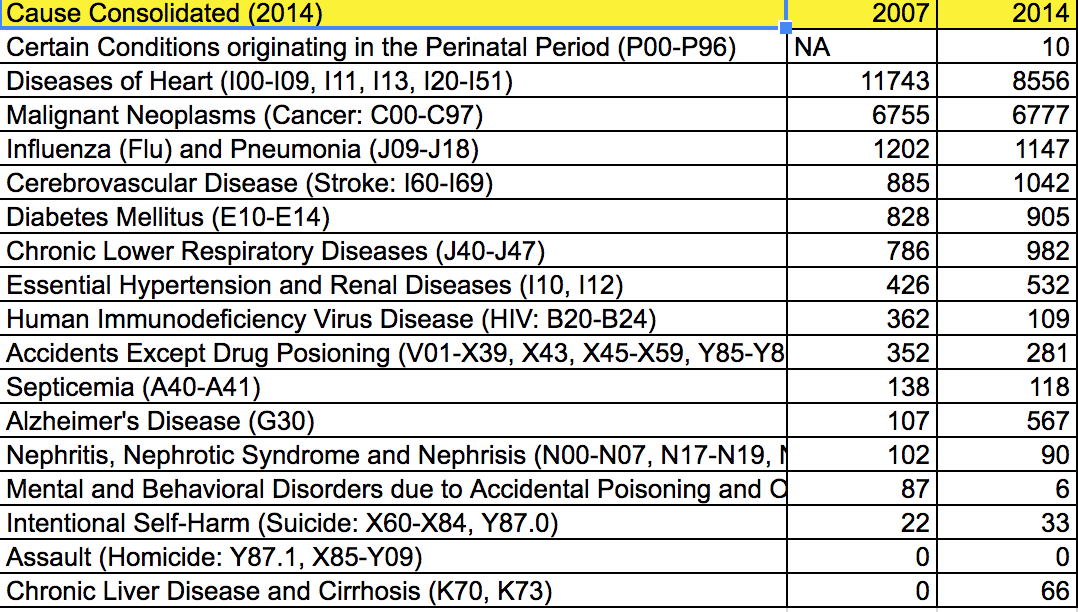
2. In the original dataset, the causes of death and total deaths were separated by racial background, making it difficult to see the total number of deaths by gender, as we were aiming to achieve. So, in order to get total deaths for each cause of death, all of the racial groups per cause of death were grouped together to get a total number of deaths per sex for each cause rather than a total number of deaths per race for each cause.  


This was done using the SUMIF function in excel. After removing the “Race Ethnicity” column, SUMIF was called per leading cause to sum the total number of deaths based on if the sex was “M” or “F”. If there was a “.” in place of a number for “Deaths”, it was replaced with a 0, assuming there was no significant data for the number of deaths for this ethnicity.

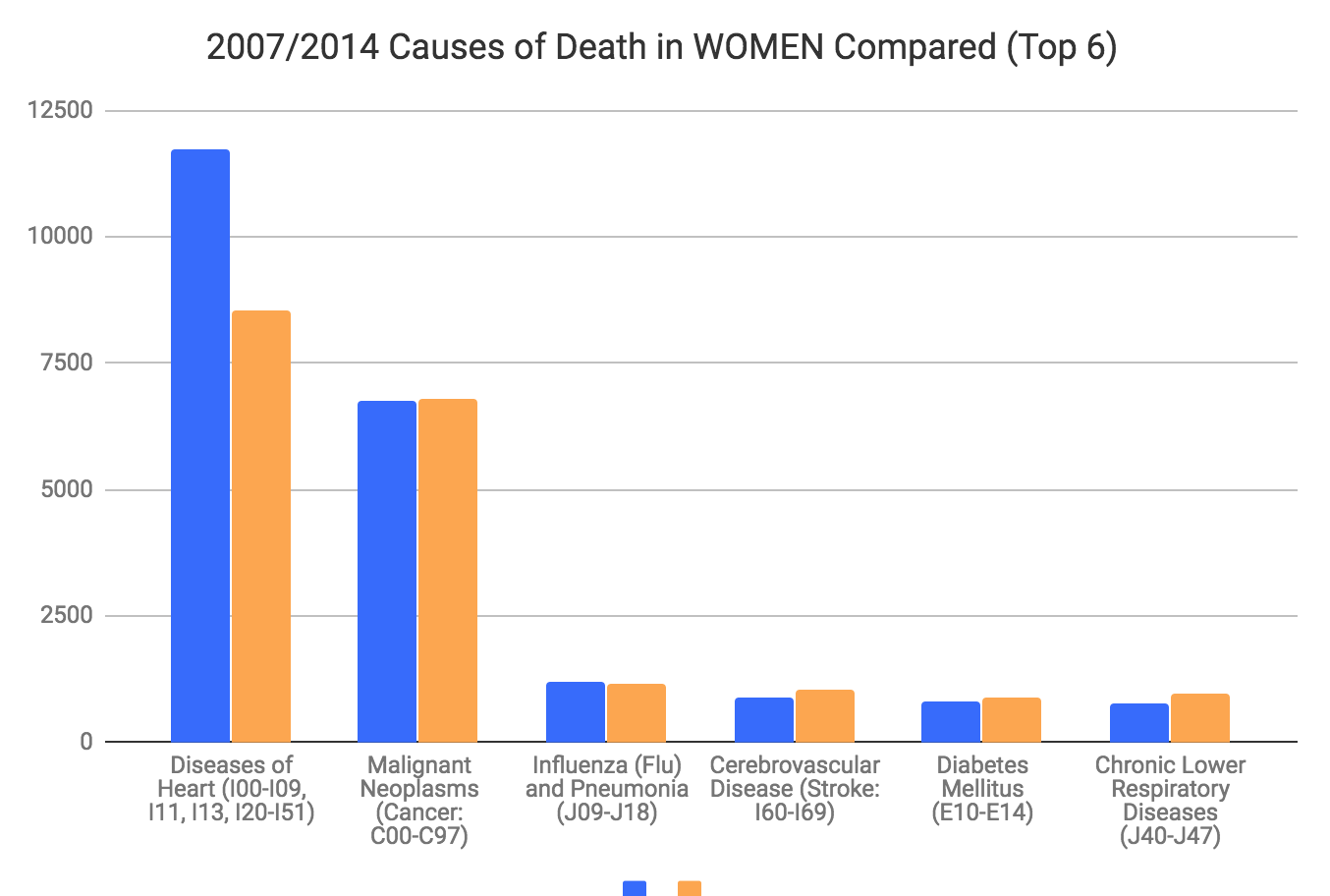
The resulting curated data is four separate spreadsheets showing in descending order the number of deaths for each of the leading causes of death. Each spreadsheet gives data for only one year and one sex. E.g. there is one datasheet for Causes of Death in Women in 2007, one datasheet for Causes of Death in Women in 2014, one datasheet for Causes of Death in Men in 2007, and one datasheet for Causes of Death in Men in 2014.



3. From this point, in order to compare the change over time in causes of death for men and women separately, the total number of deaths from both years were combined on a table in order to produce the bar graph which shows the change in number of deaths over the 7 year time period. This was done for both men and women separately because we wanted to see if there were internal changes (within the sex) and compare those to the changes between the sexes.



4. Finally, when modeling the data, we decided to use three bar graphs; one showing the number of deaths for women in 2007 and 2014, one showing the number of deaths for men in 2007 and 2014 and a third which combines the total per cause of death of men and women combined. We did this in order to be able to visually compare the numbers internally but also, despite the separation, it is clear to see the similarities and differences in the causes of death for men and women. Additionally, by combining the deaths of men and women in the third bar graph, we get an overall visual of the leading causes of death.



* **The steps you took to annotate or enrich the data:**

We believe that the data was very concise and self explanatory, specifically with the labeling of the causes of death. Other than the means of curation and the explanation of the curation provided above, there was no need for annotation or enrichment of the data. We used a dataset which provided the causes of death in New York City based on race and sex. We wanted to see the change over time in causes of death by sex only. After curation and graphical modeling, our results yielded are clear and do not necessarily need annotation or enrichment any further, based on our research questions.

Other than this, perhaps, the decision to visually demonstrate the top 6 causes of death rather than all of the causes given by the original dataset was enriching because it allows a clear and concise visualization of our results. By limiting our graphical representation to the most significant causes of death, we are left with more clear and straightforward results.

* **The tools used:**

Microsoft Excel:

* 1. The dataset was sorted by year
  2. Data per year was cut and pasted into separate spreadsheets within the same project
  3. SUMIF function was used to sum the total deaths of each cause based on sex
  4. Charts function (based on selected data i.e “Recommended Charts”) used to create bar graph data model.