* **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. More specifically we were 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.

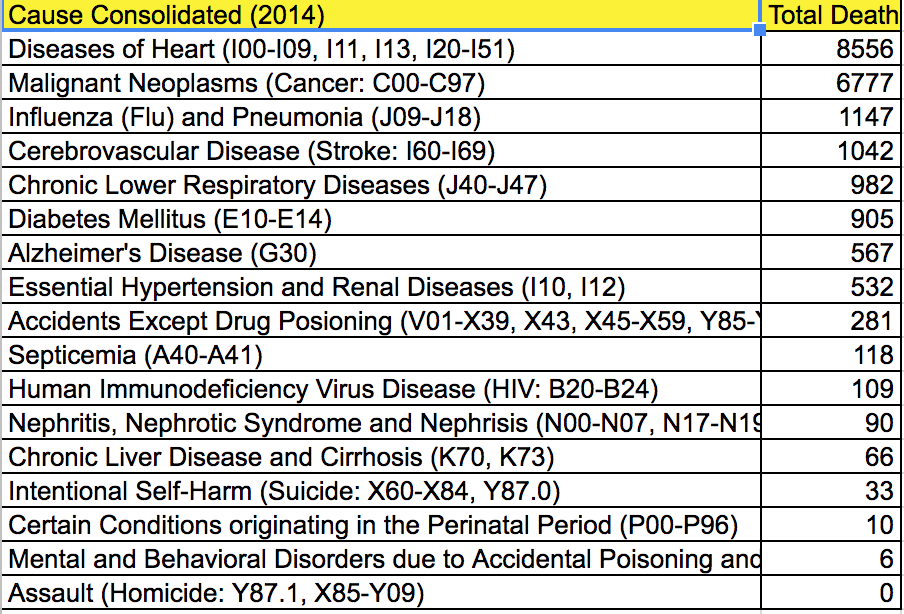
From the original dataset the steps taken to curate the dataset for our needs are as follows:

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

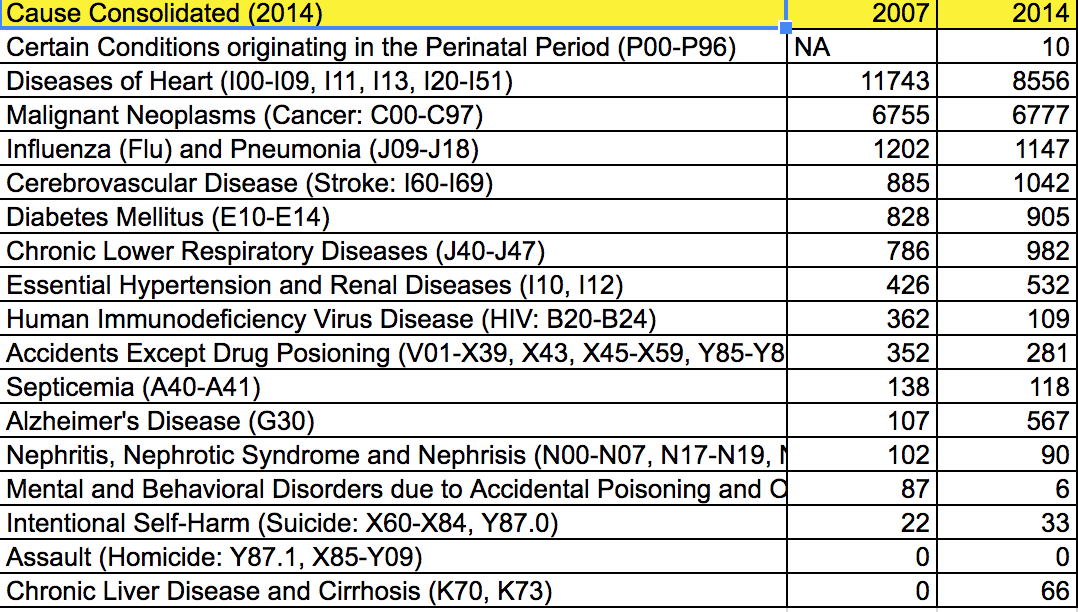
2. In the original dataset, the causes of death and total deaths were separated by racial background, making it difficult to see the total number of deaths by gender, as we were aiming to achieve. So, in order to get total deaths for each cause of death, all of the racial groups per cause of death were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

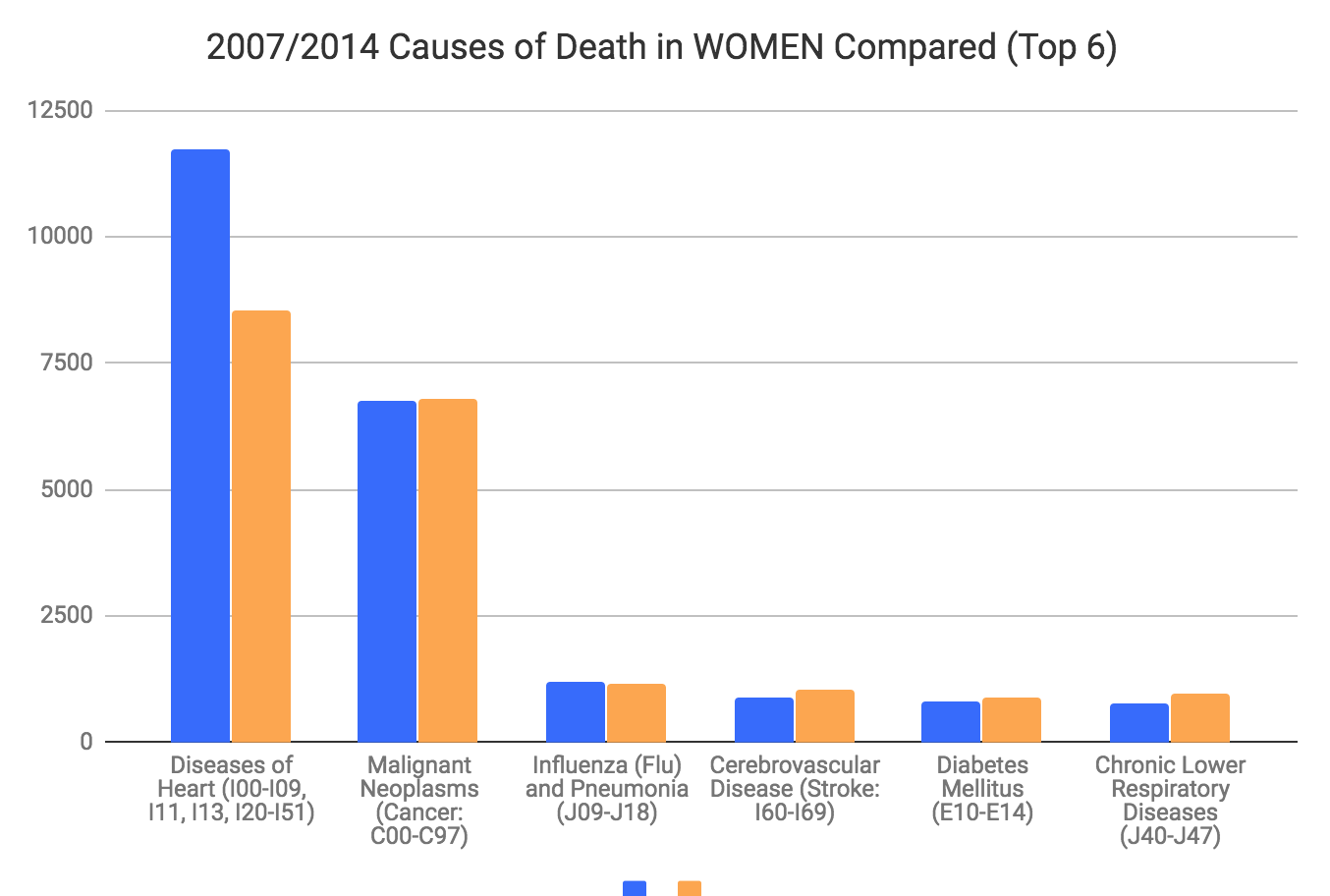
The resulting curated data is four separate spreadsheets showing in descending order the number of deaths for each of the leading causes of death. Each spreadsheet gives data for only one year and one sex. E.g. there is one datasheet for Causes of Death in Women in 2007, one datasheet for Causes of Death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