Mortality Analysis for the city of New York

Muhammad Bilal
AIT 580
George Mason University
Fairfax Virginia
mbilal3@gmu.edu

***Abstract*— The prime objective of the work is aimed at studying and revealing insights based on the major causes of death in the city of New York dataset. The information can then be used to improve the health care of the city's residents. An understanding of the death rates for the top major causes would greatly help the city's health officials set up disease prevention goals, priorities, and strategies. New York city was chosen as the subject of research because of its high and diverse population. The city's population is still on the rise necessitating a mortality analysis be performed on the city. The research work is targeted around the top causes of deaths in the city's residents of all ethnicities and racial backgrounds. The work also explored and answered a few specific questions pertaining to how the mortality rate varies by gender and time. The paper also discussed the limitations of the study, and highlight valuable findings obtained from the analysis.**

***Index Terms*—Mortality Analysis, New York City, Death count, Death rate, Ethnicity analysis**

I. INTRODUCTION

The purpose of the research was to analyze major causes of death that have the greatest impact on the mortality rate of the city. Since New York is a city with a very diverse population, there would be fair representation of people across different ethnicities and genders. A basic google search reveals that the coronary diseases, cancer and accidents have been the major causes of death for the first two decades of the 21st century.

As per the Centre for Disease Control and Prevention, coronary diseases are the number one culprit for the deaths of the NYC residents across all age groups. The senior citizens are more prone to the disease than the younger population. A staggering 44,092 [1] deaths occurred in the city in the year 2017 alone due to heart diseases as per the CDC's official website. Cancer remains the second prominent causes of death with 34,956 [1] deaths occurring in 2017. It is hypothesized that smoking is directly correlated with cancer as cigarettes contain carcinogenic elements that play a critical role in development of the disease. Accidents took up the third spot in this horrific chart. As per the CDC, 7,687 [1] people lost their lives in the city due to sudden accidents. Younger population is the group most prone to such type of deaths since the majority of such incidents are related to car collisions with drivers driving under the influence of drugs or alcohol.

Chronic Lower Respiratory Disease was also one of the top causes of mortality in the city and the country as a whole. The CDC reports 7,258 [1] deaths occurring in 2017 due to this disease of the lungs. Passive smoking and other unhealthy lifestyle choices seems to act as the catalyst for onset of this disease. Stroke, diabetics and Alzheimer's are some of the other leading causes of mortality amongst old residents of the city. Influenzas, Renal diseases, and suicide have also significantly impacted the mortality rate of the city and the country.

With pandemic creating havocs across the country, it is more important now than ever before in the history to run thorough mortality analysis on the population of one of its major metropolises to identity and protect the vulnerable segments of its population. This study of the top causes of mortality could prove to be a great help to that end. The study is focused on the major causes of death in the city of New York based on a dataset collected by the Department of Health and Mental Hygiene (DOHMH). The data set includes the major causes of death by across ethnicities and genders since 2007. The research work was centered around answers the following 5 questions.

- *What cause had the highest contribution to the overall mortality count of the city of New York?*
- *How does the Mortality cause and counts differ by the gender of the residents of the city of New York?*
- *Is there a disparity between the death reported data by different ethnicities of the city of New York?*
- *Are specific ethnicities more prone to a certain cause of mortality in the New York city?*
- *How has time impacted the major cause of mortality?*

II. LITERATURE REVIEW

Moy et al. performed a statistical analysis on the leading causes of death in the United States in their journal article titled "Leading Causes of Death in Nonmetropolitan and Metropolitan Areas — United States, 1999–2014". The papers made a direct mortality rate comparison between Metropolitan and Nonmetropolitan regions of the United States. The analysis was performed on the age-adjusted data for the five leading causes of deaths based on the International Classification of Diseases, 10th revision. The data included deaths from heart diseases, cancer, chronic lower respiratory disease,

unintentional injuries, and stroke. The trend followed a Poisson distribution. Standard Errors for both the number and percentage of excess deaths were also calculated in order to incorporate the variance so as to provide a good comparison between the expected and the observed counts. Joinpoint regression was used to analyse the national mortality trends for nonmetropolitan and metropolitan areas around the country. The only results that were considered statistically significant were the ones that had a p value of less than 0.05. The article concluded with the surprising findings that the mortality and death rates of people in the Nonmetropolitan areas of the country were significantly greater than that of the Metropolitan region of the country. [2]

Espay et.al also carried out a similar type of research work and published their findings in the American Journal of Public Heath in the article titled "Leading Causes of Death and All-Cause Mortality in American Indians and Alaska Natives". The work was heavily centred towards regional patterns and trends in all-cause mortality and leading causes of death in American Indians and Alaska Natives. County-level population estimates gathered by the US Census Bureau were used as the baseline for the rate calculations. The fact that the US Census Bureau bridges race categories into a single-race annual population estimate greatly helped the researchers segregate the targeted race death data. The research employed the SEER*Stat software version 8.0.4 in order to express the per 100,000 population to the 2000 US standard population. The already adjusted-age death rates were used to calculate confidence intervals and Standard Errors. The article further utilized the Joinpoint regression techniques and performed the analysis assuming results that had a $p<0.05$ as statistically significant. The research concluded with very alarming insights uncovering the fact the death rates of American Indians and Asian Natives were nearly 50% greater than rates in Whites. [3]

A group of researchers based in San Francisco led by Hastings et al. published a research paper titled "Leading Causes of Death among Asian American Subgroups (2003–2011)". The fact that Asian American population is one of the fastest growing racial/ethnic groups in the U.S compelled the researchers to analyse the leading causes of deaths amongst this ethic group. The data was obtained from the National Center for Health Statistics. The data had to be cleaned and processed as an estimate for denominator population counts for the study period had to be found out based upon the fact that the population of all the groups grew linearly. Age adjusted death rates were calculated using direct standardization. Joinpoint regression models and average annual percentage change was used in order perform statical analysis. 95% confidence intervals were used in order to determine the magnitude and direction of the trends. R version 3.1.0 was employed in order to manage and analyse all of the data. The research revealed that the Malignant neoplasms (cancer) were the leading cause of death in Asian American females (28.6% of all deaths) and Asian American males (27.4%). [4]

III. METHODOLOGY

The dataset used was collected by the Department of Health and Mental Hygiene (DOHMH). The file was available in the csv format at the NYC open data website. The file contains data records for years 2007 and onwards. The data covers four racial groups which includes Hispanic, White Non-Hispanic, Black Non-Hispanic, and Asian and Pacific Islanders. There were some rows where the ethnicities of the individuals were not known. The data containing such records had 'Not stated/Unknown' under the Racial_Ethnicity column. The unprocessed data has 1273 records.

The raw data contained a lot of blank entries, some of them were marked with a blank space while others were marked with a ".". Python (Spyder) was used to removed rows with entries that contained any of the two forms of Null values. The process removed 453 incomplete rows.

Some of the data entries in the Gender column used a mix of M,F,Male and Female in order to mark the gender. These data inconsistencies were corrected by making all the relevant values homogenous. The data in the Major_Cause column contained codes and numbers. Microsoft Excel find and replace feature was used to fix this so as to keep just the names of the diseases and causes in order to enhance the visualizations. Some of the data values in the Racial_Ethnicity column contained different strings to represent the same Race. For instance, White Non-Hispanic entries were sometimes represented by Non-Hispanic White. Microsoft Excel was used to all of these data inconsistencies. Finally, Python was used to find and remove all the rows that contain "Not Stated/Unknown" string in the Racial_Ethnicity column of the data frame. The processed data was saved as a csv file so that it could be accessed by other software used for the analysis to follow.

The data was then loaded into the R software and visualizations were created in order to answer the research questions. The data was also inserted into the Oracle Sql developer in order to execute some basic queries and create a table schema so as to demonstrate my understanding of the concepts learned in class.

IV. RESULTS AND ANALYSIS

A. *What cause had the highest contribution to the overall mortality count of the city of New York?*

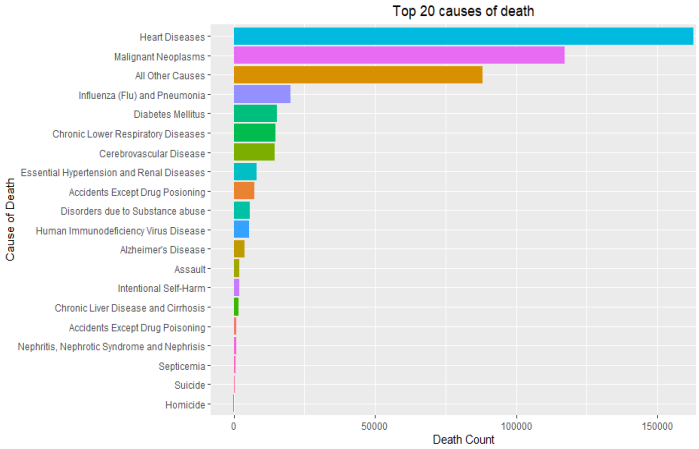

Fig 1.0: Graph displaying top 20 causes of death in the NYC

Data was sorted so as to plot top 20 Major causes of death along with the death count of each of the reason. R package dplyr [6] was used to prepare the data for the visualization and ggplot2 [5] was used to create the visualization. As it can be seen from the graph in Fig 1.0, heart disease was the cause of the greatest number of deaths (175,000) which was closely followed by Cancer (135,000) and All other causes (81,000). Influenzas, Diabetics, Respiratory diseases, and Cerebrovascular disease did contribute to the death count too. The mortality contribution of the other causes was not as significant as that of the top three diseases indicating that Heart Disease, Malignant Neoplasms (Cancer), and, All other Causes, were more common than any of the other diseases to have caused deaths in the city of New York.

B. *How does the Mortality cause and counts differ by the gender of the residents of the city of New York?*

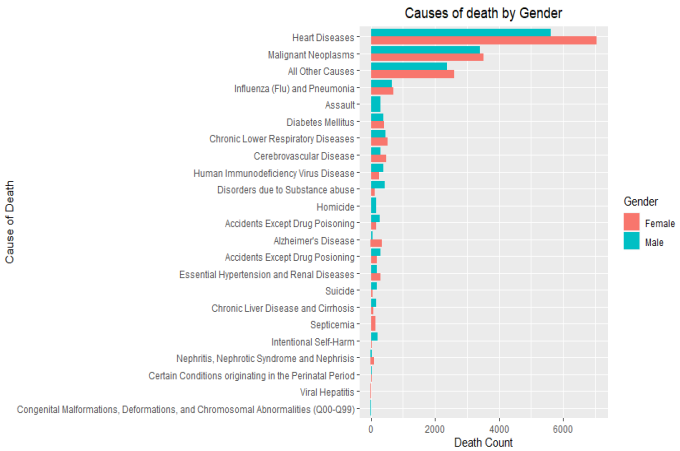

Fig 2.0: Grouped bar chart displaying causes of death by gender

The visualization in Fig 2.0 does a side-by-side gender comparison of the major causes of deaths. Female residents of the city were more prone to death by diseases as compared to their male counterparts. Between year 2007 and 2019, 7100 women died due to heart diseases alone while 5800 men died of the same disease. The men, on the other hand, had higher recorded deaths for assault crimes, deaths due to disorders caused by the substance abuse, Accidents, Suicides and Intentional self-harm as compared to women. As per the graph, chromosomal abnormalities, conditions originating in the perinatal period and, international self-harm had least effect on the female mortality count while men were better equipped to handle causes like the Alzeimer's disease, Nephrotic syndrome and viral hepatitis.

C. *Is there a disparity between the death reported data by different ethnicities of the city of New York?*

The third question is about investigating how complete the available mortality data for each of the ethnicity is. The graph in Fig 3.0 revealed that the Non-Hispanic White ethnicity had the Highest reported Death count in New York City. This was followed by the Non-Hispanic Black ethnicity which reported second highest death count with around 151,500 deaths reported. The results were a little surprising as the reported death count for Asian and Pacific Islanders is 26000 as compared to 81000 reported by the Hispanic communities of New York.

The table 1.0 showcases the percentage population of each ethnicity based in the city of New York. As it can be seen in the table, Hispanic or Latino make up 28.9 % of the city's total population while Asian and pacific islanders combined make up 14.1 %. The numbers suggest that the reported death count for both the ethnicities should differ by almost half considering the strength of both the populations.

Race and Hispanic Origin	
White alone, percent	41.3%
Black or African American alone, percent (a)	23.8%
American Indian and Alaska Native alone, percent (a)	0.4%
Asian alone, percent (a)	14.3%
Native Hawaiian and Other Pacific Islander alone, percent (a)	0.1%
Two or More Races, percent	5.6%
Hispanic or Latino, percent (b)	28.9%
White alone, not Hispanic or Latino, percent	31.9%

Table 1.0: Portraying racial population breakdown of NYC by the percentages. [7]

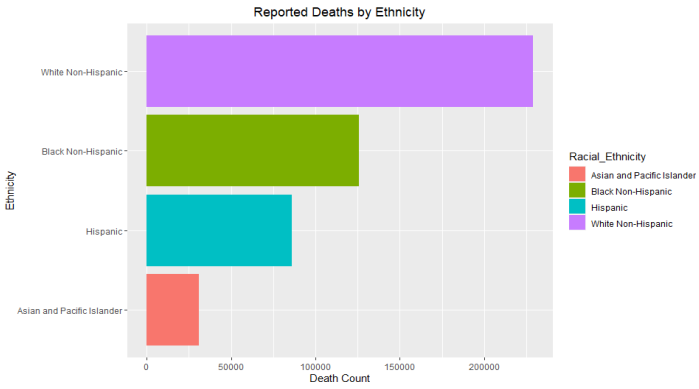

Fig 3.0: Reported deaths by Ethnicities

There could be two possible explanations for comparatively lower reported death count for the Asian Population. The first and the highly likely reason for this phenomenon would be the hesitance and obstacles faced by the Asian communities in having their deaths reported and registered by the authorities. If this is true, the city needs better collection of public health data. The second possible explanation for this could be healthier lifestyle choices of the Asian communities living in the city of New York.

D. Are specific ethnicities more prone to a certain cause of mortality in the New York city?

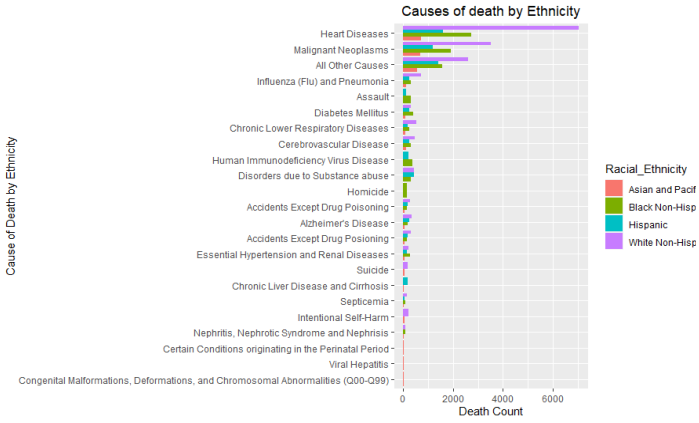

Fig 4.0: Causes of death by Ethnicity

The grouped bar chart in Fig 4.0 does a side-by-side comparison of top causes of mortality along with the death counts for each of the ethnicity providing a detailed overview. The graph shows that White Non-Hispanic population reported the highest number of deaths due to heart diseases followed by Black Non-Hispanic, Hispanic, and Asian & Pacific Islanders respectively. Asian and Pacific Islander communities reported roughly the same number of deaths for both the heart disease and Cancer which is also their top cause of mortality recording their highest death numbers.

Listed below are the 4 leadings causes of mortality for each of the ethnicities part of the analysis (Highest to lowest).

Asian and Pacific Islanders - Cancer, Heart diseases, All other causes, and Influenzas.
Black Non-Hispanics – Heart diseases, Cancer, All other causes, and Human Immunodeficiency Virus Disease.
Hispanics - Heart disease, All other causes, Cancer, and disorders due to substance abuse.
White Non-Hispanics – Heart diseases, Cancer, All other causes, and Influenzas.

Black Non-Hispanics were most likely to die of Assault incidents than that of any other ethnicity while Hispanic community also recorded significant number of deaths by Assault. White Non-Hispanics and Asian & Pacific Islanders recorded very low (almost insignificant) number of deaths where the prime cause was Assault. Similarly Black Non-Hispanic communities were found extremely vulnerable to the incidents of homicide as they were the only community to have recorded significant number of deaths to such violent incidents.

As per the bar chart, White Non- Hispanic communities were more likely to die of Intentional self-harm and suicide than any other racial groups. Asian and Pacific Islanders communities also reported significant number of deaths due to Intentional self-harm and suicide. Hispanics and Black Non-Hispanic communities reported very few deaths due to any of these unfortunate incidents.

The graph also suggests that the Hispanic communities had the highest likelihood of being affected by Chronic liver diseases and disorders due to substance abuse. Asian & Pacific Islanders were the only other racial groups that had significant reported deaths due to chronic liver disease. Asian & Pacific Islanders reported the least number of deaths for incidents related to substance abuse.

Asian & Pacific Islanders was the only community to have reported considerable number of deaths for certain conditions originating in the Perinatal Period, Viral Hepatitis, and Congenital Malformations.

E. How has time impacted the major cause of mortality?

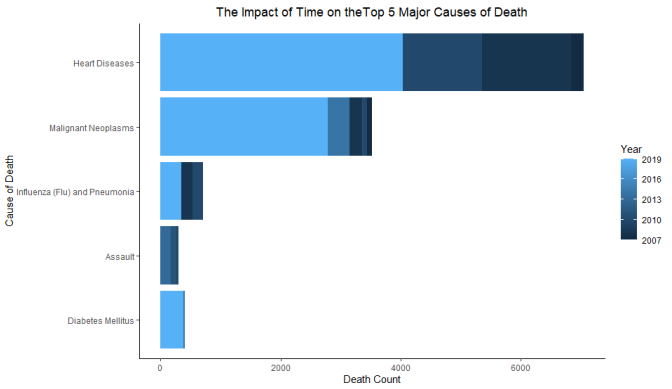

Fig 5.0: Time impact on major cause of mortality

The stacked bar chart in the fig 5.0 enables the visualization of the impact of time on the 5 major causes of death in an attempt to answer my final research question. The lighter shades of blue represent latest years while the darkest shade represents the initial years the data began to be recorded (2007).

The graph suggests that there has been a significant drop in the number of total deaths due to heart diseases. The year 2007 recorded the highest number of deaths due to the disease followed by the 2010, and 2013. The year 2019 logged the lowest number of deaths due to heart disease. Although 2019 saw the least number of deaths reported due to heart disease, heart disease was still the major cause of mortality amongst the New York City residents that year. One of the major causes of such a positive shift could be the residents of the city being more aware of the consequences of poor lifestyle habits. The residents of the city might have collectively decided to adopt healthier lifestyles.

Malignant Neoplasms followed similar pattern when it came to the death counts by the year. 2007 showcased the highest number of reported deaths due to cancer while the count gradually dropped for the years to follow. Deaths due to Influenzas and Pneumonia also dropped as the years went by. This hints to the health care systems of the city being enhanced and updated so as to better serve its residents.

The graph also made a surprising revelation that the deaths due to assaults saw a massive decline. While the years 2007 to 2013 saw a lot of deaths related to the incidents of assault, the death count almost dropped to 0 post year 2013. The data goes on to show that the city's law and order situation was greatly improved post 2013.

Although all of the top major causes of mortalities saw some sort of decline in the number of deaths, the deaths occurring as a result of Diabetics remain almost the same. New York City's health administration should be alarmed by the relentlessness of this number and should work to lower it.

V. LIMITATIONS

I did not have any age specific mortality counts for each of the ethnicities. It would have made for a great analysis if I could incorporate those into the discussion. The knowledge of which age group is the most vulnerable to what type of mortality cause could greatly help the health and rescue organizations working around the city to lower the numbers of preventable deaths every year.

After the removal of null/missing values as the part of data processing, all of the values for the year 2017 were unfortunately removed. The presence of data from 2017 would have added to the accuracy of the analysis. 476 rows of the data had to be removed due to the incompleteness of data. The presence of all the values for data would have further added to the reliability of the analysis.

The literature on the mortality analysis of the New York City is extremely scarce. It would have provided me with a great starting point to build the research on had I found some similar previously done work in the similar niche.

Another limitation that I ran into was the generalization of the diseases in the Major_Cause column. Although it did ease the analysis, the presence of subcategories for each disease/cause of death would have made a positive impact on the analysis as the whole. The presence of each individual disease/major cause of death instead of "All Other Causes" would have added to the credibility of the research.

VI. CONCLUSION

In conclusion, this research work was successfully able to examine data and provide a thorough analysis on the leading causes of death across years for the New York City's residents belonging to each of the racial groups and genders.

The analysis revealed that Heat Diseases is the leading cause of death across all genders and races in the city of New York. Although the death count due to this disease has decrease in the recent years, the number is still huge. The female residents are a lot more vulnerable to death by heart diseases than their men counterparts. Residents need to make tiny adjustments to their lifestyle and take time out to exercise more and lower the consumption of junk/high fat food in order to lower this number.

The mortality cause that claimed the second highest number of deaths as per my analysis was Malignant Neoplasm (Cancer). Female residents of the city were more likely to die of cancer than the men. The government should impose mandatory cancer screening for all of its residents once every year. The government should apply increased taxes on all the tobacco products so as to discourage their consumption which could in turn prevent a lot of deaths caused by cancer.

Black Non-Hispanic and Hispanic communities have been victims of violent crimes. The law enforcement agencies around the city need to work with these communities so as to lower the

number of these preventable deaths. Similarly, White Non-Hispanics and Asian & Pacific Islander communities around the city were more prone to death due to Intentional Self Harm, and Suicides. The government could work to subsidize therapy for these ethnic groups so as to prevent/lower the number of mortalities occurring due to this cause.

VII. References

[1] Centers for Disease Control and Prevention. (2018, April 10). Stats of the State of New York. Centers for Disease Control and Prevention. Retrieved May 1, 2022, from https://www.cdc.gov/nchs/pressroom/states/newyork/newyork.htm

[2] Moy, E., Garcia, M. C., Bastian, B., Rossen, L. M., Ingram, D. D., Faul, M., Massetti, G. M., Thomas, C. C., Hong, Y., Yoon, P. W., & Iademarco, M. F. (2017). Leading causes of death in nonmetropolitan and metropolitan areas— United States, 1999–2014. MMWR. Surveillance Summaries, 66(1), 1–8. https://doi.org/10.15585/mmwr.ss6601a1

[3] Espey, D. K., Jim, M. A., Cobb, N., Bartholomew, M., Becker, T., Haverkamp, D., & Plescia, M. (2014). Leading causes of death and all-cause mortality in American Indians and Alaska natives. American Journal of Public Health, 104(S3). https://doi.org/10.2105/ajph.2013.301798

[4] Hastings, K. G., Jose, P. O., Kapphahn, K. I., Frank, A. T., Goldstein, B. A., Thompson, C. A., Eggleston, K., Cullen, M. R., & Palaniappan, L. P. (2015). Leading causes of death among Asian American subgroups (2003–2011). PLOS ONE, 10(4). https://doi.org/10.1371/journal.pone.0124341

[5] Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis (2nd ed.) [PDF]. Springer International Publishing.

[6] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

[7] U.S. Census Bureau quickfacts: New York City, New York. (n.d.). Retrieved May 2, 2022, from https://www.census.gov/quickfacts/newyorkcitynewyork

VIII. Glossary

Malignant Neoplasm – Cancer

Chronic Lower Respiratory Disease – Lung Diseases

CDC – Centre for Disease Control

DOHMH - Department of Health and Mental Hygiene

Mortality Cause – Cause of death

Joinpoint regression - A Windows-based statistical software package used to perform regression analysis

CSV – Comma Separated Values

Poisson distribution - A statistical tool that helps to predict the probability of certain events happening when we know how often the event has occurred in the past

Correlation - Relationship