Kyra Zhu, Delia Carlino

December 14, 2023

MATH 7360-01

Final Report

**Introduction**

In investigating international health and disease trends, it is impossible to ignore the importance and the effects of non-communicable diseases (NCDs). Non-communicable diseases are those diseases that cannot be transmitted from to and from people and are chronic in nature. Understanding NCDs is incredibly important as they are the category of diseases with both the highest burden of disease and rates of mortality. As seen in Figure 1, there are two disease types (cardiovascular diseases and cancers) that contributed to a combined 644.5 million healthy life years lost in 2019 alone. This measurement of disability adjusted life years considers years of life lost to premature death as well as years lived with disability. Additionally, when looking at mortality statistics in 2019, the four most common classifications of non-communicable diseases (cardiovascular diseases, diabetes mellitus, chronic respiratory diseases, and cancers) were the combined causes of over 30 million total deaths per year with over 15 million of those deaths affecting people between the ages of 30 and 69. Of those diseases cardiovascular diseases has the highest mortality, followed by cancers, then chronic respiratory diseases, and then diabetes.

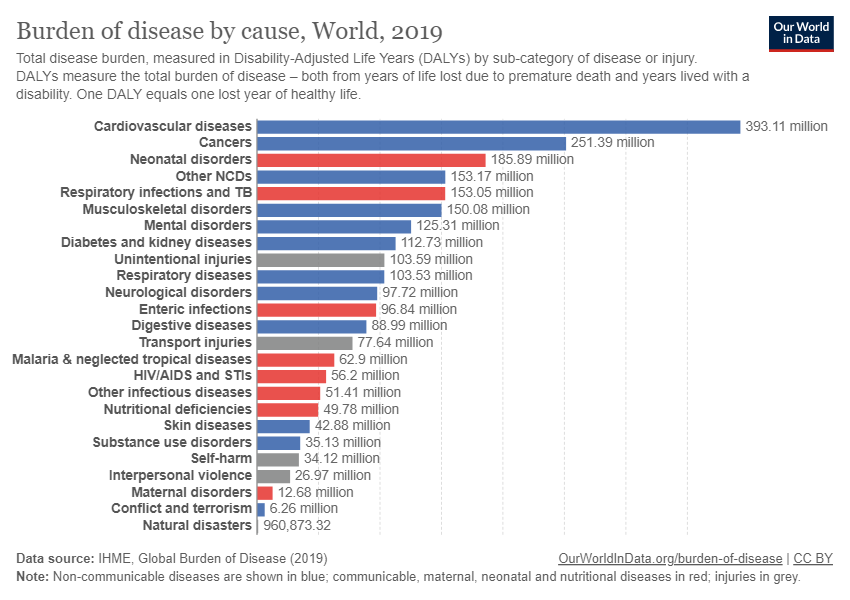


Figure 1: Burden of disease statistics in 2019 showing non-communicable diseases contribute heavily to disability

Therefore, non-communicable disease mortality is an important metric when studying global health. Non-communicable disease mortality can be investigated for patterns that can be used for a more focused response to health discrepancies and trends. The World Health Organization provides a useful and trustworthy database with NCD mortality data from the years 2000-2019, to help elucidate trends in NCD mortality relating to sexual dimorphism, regions, time, and potential risks.

**Analysis**

All of the analyses of this project were conducted using RStudio. The libraries that were necessary for these analyses include tidyverse, readxl, ez, countrycode, sf, rnaturalearth, rnaturalearthdata, wesanderson, gtsummary, and faraway.

Examining first trends related to sexual dimorphism, it was important to gather data on sexes and sex-related factors, specifically the causes in this context. The data had been aggregated by sexes and causes, focusing on the most recent ten-year period. These data were then categorized into three distinct tables based on both sexes, males, and females, along with four different classes of cause which are cardiovascular diseases, respiratory diseases, diabetes mellitus, and malignant neoplasms. For each year's data on sexes or causes as data points, line graphs were generated to observe differences and trends. This examination selectively excluded irrelevant factors, considering only variables related to sexes and causes. The extracted data was used to formulate NCD mortality rates across sexual factors and causes.

In looking at region and cause specific trends, it was necessary to not include sex as a variable, for ease of analysis. Therefore, the next sections of analysis only considered the following variables: year, region, cause, and cumulative mortality. The total mortality rates over time based on regions, then the total mortality rates over time based on regions for each specific disease category, and then the total mortality rates over time for each region with the diseases represented were examined. These analyses were completed by separating out the data into numerous variables before aggregating the data into appropriate data frames. This section of analysis required the creation of several sets of variables to allow for the specific plot outputs. Each variable filtered out data that was not relevant and aggregated mortality values for each year. The cumulative mortality variables were used to create comparisons of each region. The disease related mortality variables were used to create both disease specific and region specific graphs.

In addition to these visual and qualitative measures to assess the trends in NCD mortality across the specified time period and regions, ANOVA tests were utilized to establish that the differences within the regions across time were significant. This was done using the ezANOVA function from the ez package. This allowed for the performance of six repeated measures ANOVA tests, one for each region. In each of the tests, the dependent variable was the number of NCD deaths, the case identifier was the year, and the cause of death was the observed variable. The resulting F-values, p-values, and values are presented in the next section.

The next step in analyzing the region trends was to create a NCD mortality map that showed the cumulative mortality rates of all four disease categories on one global map. This section included the same categories of variables as the last sections, but also included country specific data. The idea was to create a map with a gradient showing which countries have higher cumulative mortalities due to non-communicable diseases, but the initial map was heavily biased due to population differences and, thus, did not show any interesting trends. Therefore, cumulative population data (from the World Bank) was incorporated to create a ratio based map showing which countries had higher rates of NCD related deaths:cumulative population.

For the map, the cumulative NCD related mortality for each country was imported for the most recent year (2019). This required filtering the data to only the data for both sexes, aggregating the separate data for each cause of death, and inserting ISO 3166-1 alpha-3 codes for each country to allow for superimposition onto the map. The population data from the World Bank was then read and merged. The two datasets were joined and then a column for the ratio of deaths:total population was included. A data table of the world map was then joined with the other data. Antarctica was removed from the map.

Two regression models were then created: one that included all regions and one that excluded Europe in the training to see if the model could accurately predict the values of NCD mortality in that region. The reason Europe was chosen as the testing region was that Europe was the only region in which NCD mortality has decreased from 2000-2019. The process for creating the two models was identical. The data was imported and filtered to include only data from 2016, as this was the only year within the period that included data on all of the relevant variables. The variables that were considered were sex, cause of death, obesity rates, blood pressure, cholesterol rates, human development index, and life expectancy. The models were binomial functions.

**Results**

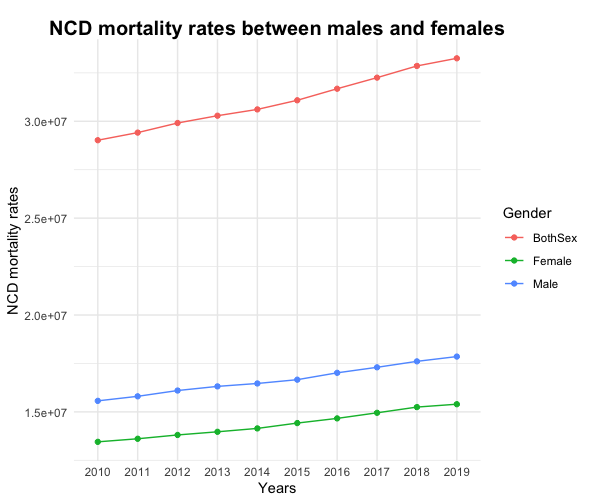


Figure 2. NCD mortality rates between males and females from 2010 to 2019

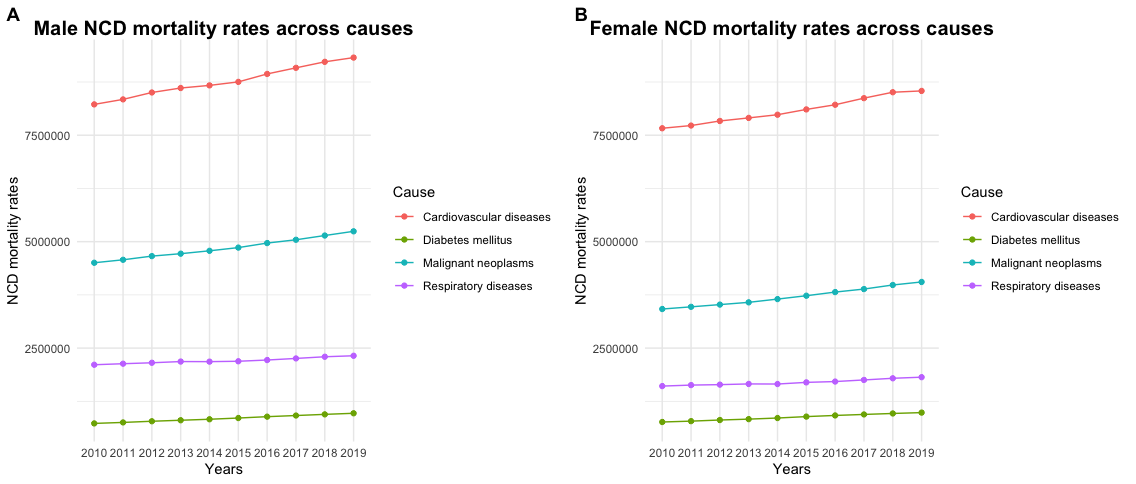


Figure 3. NCD mortality rates across different causes between male and female

NCD mortality rates have been consistently increasing in recent years. According to Fig. 2, there is a significant discrepancy between different sexes. Males have experienced notably higher rates of NCD mortality than females from 2010 to 2019, and both genders have shown an increasing trend. The rates of NCD mortality between males and females across distinctive causes are investigated, including cardiovascular diseases, diabetes mellitus, malignant neoplasms, and respiratory diseases. By presenting all the lines for each cause in one graph, it is evident that cardiovascular diseases caused the highest mortality rates, surpassing other causes by more than two times (Fig. 3). The second-highest was malignant neoplasms, followed by respiratory diseases and diabetes mellitus with the lowest rates. Therefore, interventions targeting cardiovascular diseases and malignant neoplasms are crucial to reducing NCD mortality rates. Meanwhile, males still experienced higher mortality rates than females for all causes except diabetes mellitus. This figure also indicates a consistent increase from 2010 to 2019 for both sexes and all causes.

Region trends show an overall increase in NCD mortality from 2000 to 2019 across causes, with the exception of the European region (Fig. 4). While this region has relatively high levels of NCD mortality, the rates are decreasing. Additionally, since this data is not corrected for population sizes, it is possible that the high levels are an artifact of higher populations. The other interesting observations from this figure is the steady increase in NCD mortality in South East Asia starting in 2005 and the continuous high rates and increases in mortality in Western Pacific.

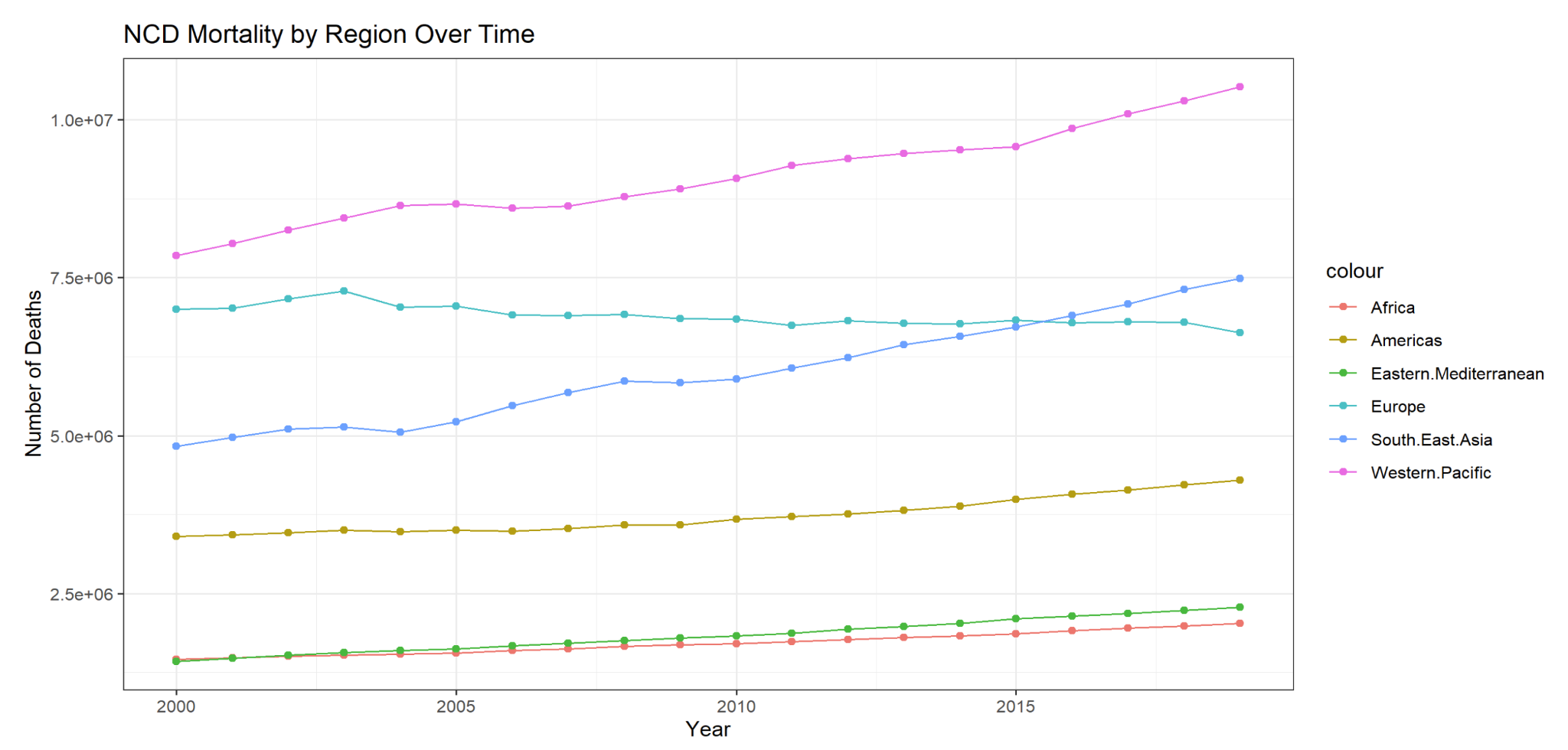


Figure 4. Cumulative NCD mortality from 2000 to 2019

Analysis of the trends of each specific disease were consistent with the region trends. As seen in Figure 5, the rates of mortality from each disease increased in the examined time span. Additionally, Figure 5C shows a steep increase in diabetes mellitus related mortality from 2000 to 2019. Currently, diabetes is the disease category that contributes the least to cumulative NCD mortality, but this steep upward trend across all regions may predict that diabetes related mortality will continue to increase to a point of major concern.

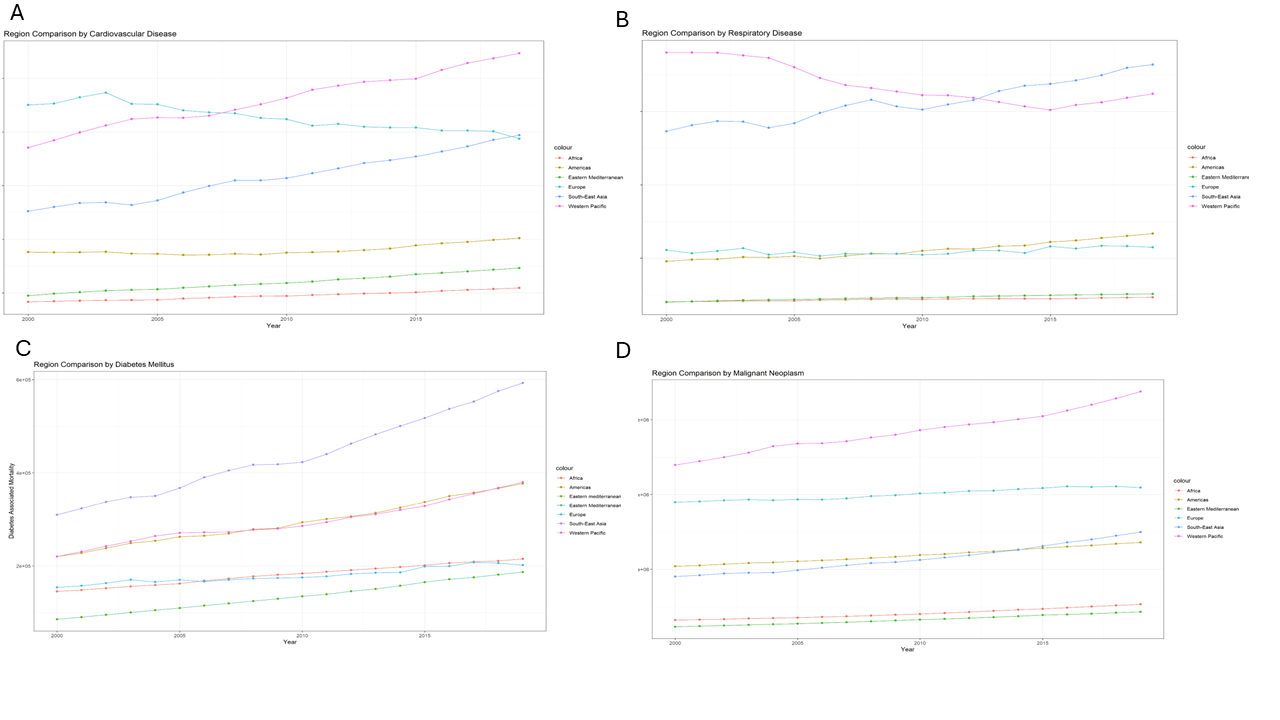


Figure 5. Relative rates of mortality for each of the main non-communicable diseases

The ANOVA analyses showed that in each region, there is a statistical significant change in NCD mortality due to year. Additionally, due to the effect size values, it is shown that it is a large effect. This means that, in all regions, year has a large statistical significant relationship to NCD mortality.

Table 1: Values from ANOVA tests of significance across years

|  | Africa | Americas | East Med. | Europe | SE Asia | W Pacific |
| --- | --- | --- | --- | --- | --- | --- |
| F-Value | 2106.349 | 7187.747 | 1228.121 | 3614.302 | 1007.348 | 958.3625 |
| P-Value | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |
|  | 0.9730286 | 0.9846346 | 0.9634546 | 0.9942187 | 0.9446147 | 0.9687332 |

The map of NCD mortality, as seen in Figure 6, shows that countries with higher populations have higher rates of mortality, even when adjusted for population. As seen in the map, China and India have the highest rates of NCD mortality. Other countries that stand out are the United States, Russia, Brazil, and Indonesia.

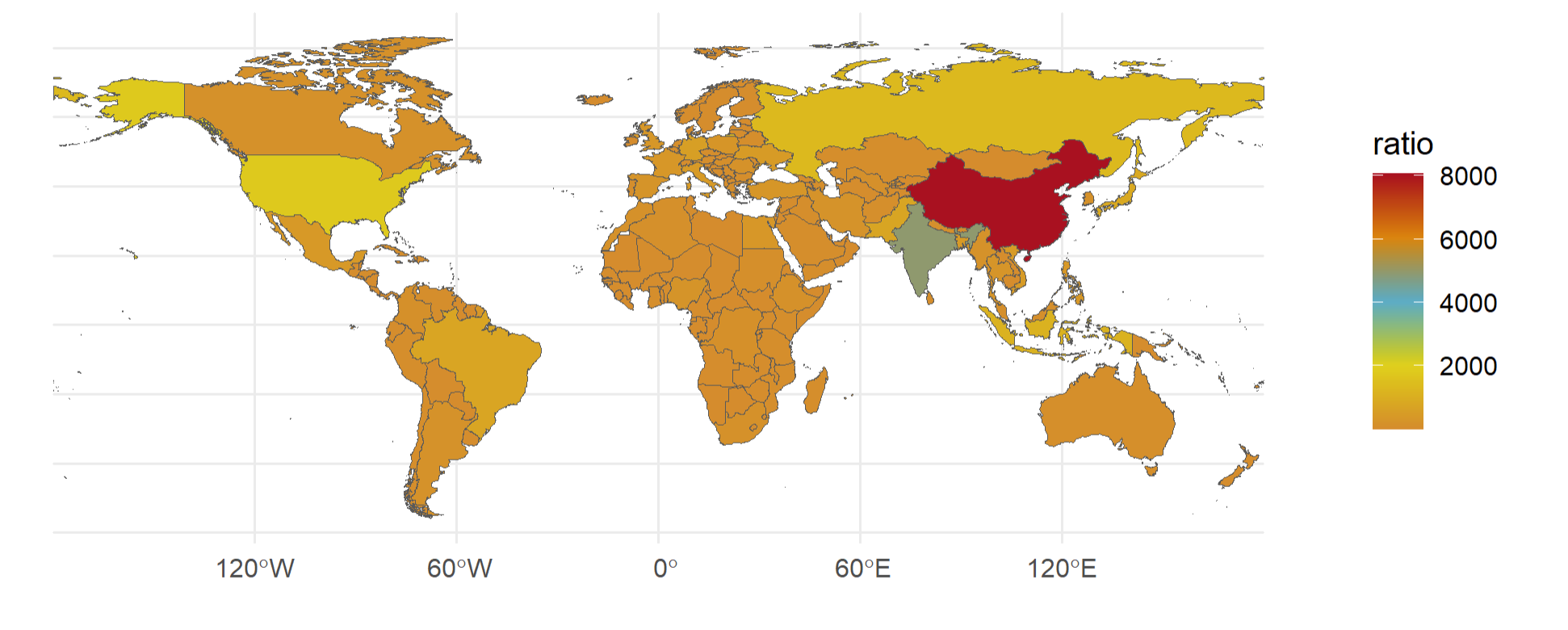


Figure 6. World map of NCD mortality when adjusted for population

For the regression models, neither model was evaluated to be very accurate. The equation for the cumulative regression model was:

The equation for the model that excluded Europe in training was:

The values for both models were 0. Residual plots also showed significant deviation from the expected lines.The accuracy of prediction of the second model on the European countries is shown in Figure 7.

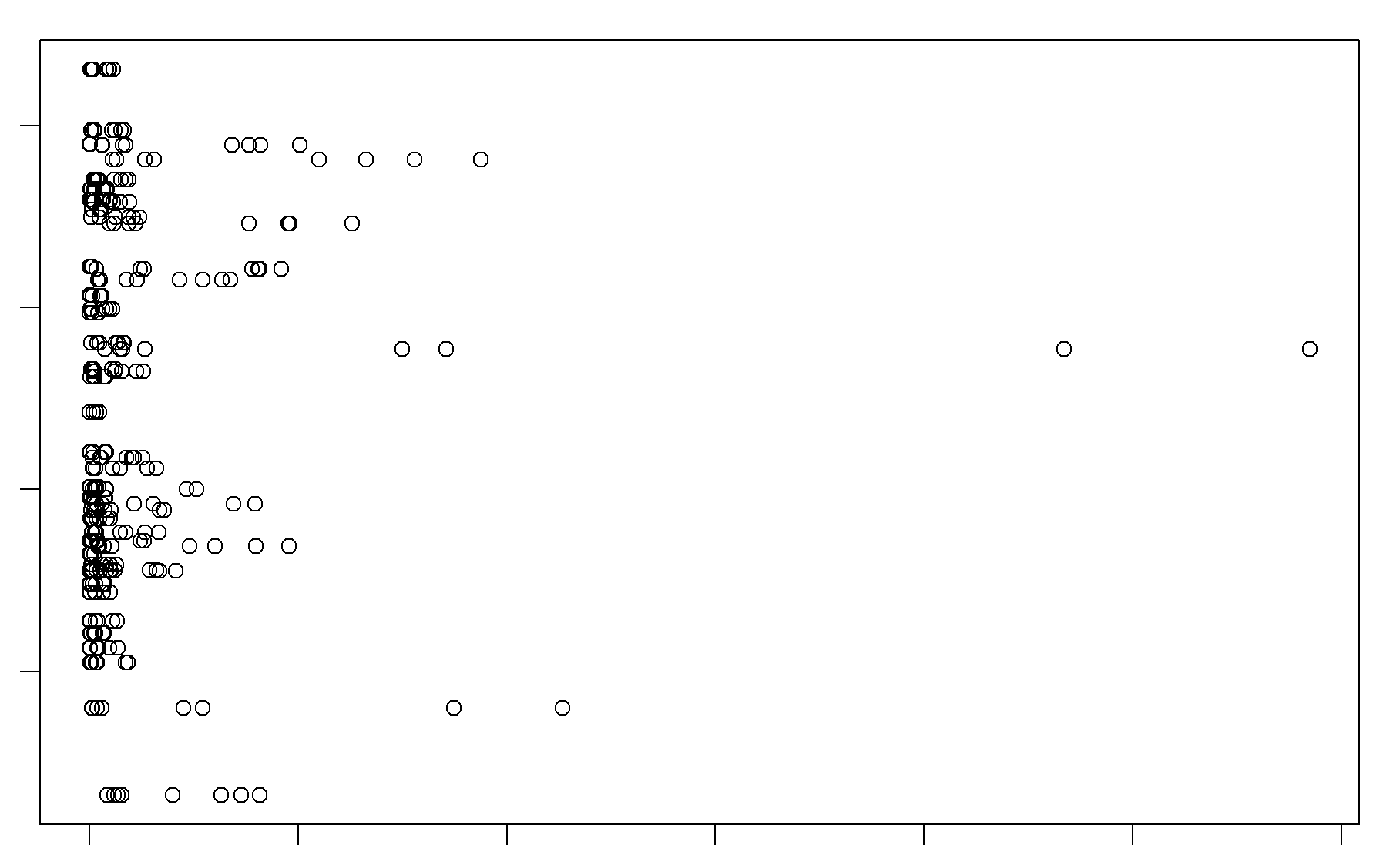


Figure 7: Graph showing predicted mortality from the regression model (x-axis) vs actual mortality (y-axis) for all countries in Europe

**Discussion**

This analysis is limited in scope, which creates some distinct limitations in applicability of the results. However, the findings of this analysis remain interesting. The sex related trends show higher rates of NCD mortality in men, although this could be an artifact of population sizes, but merits further investigation due to heightened interest in sex as a biological variable. The region related trends show that Europe has decreasing levels of NCD mortality, which shows the effectiveness of their health initiatives. Europe, therefore, could stand as a model for public health interventions. Conversely, the Western Pacific and Southeast Asian regions show high and increasing rates of NCD mortality and pose as potential sites for increased public health attention and interventions. These findings also merit further study. Further attention could be placed on investigating the correlation between regions, years, and causes of mortality as well as increased use of population when comparing rates. Additionally, the regression model shows the difficulties associated with creating good models of NCD mortality due to the large numbers of variables contributing to etiology and advancement of the diseases and lack of linearity in the data. Overall, despite some difficulties, particularly with the regression model, this project has been useful in learning about NCD mortality and examining data using R.

**Supplemental Information: Github Repository**

<https://github.com/delia-c/NCDMortality>

**Supplemental Information: Data Sources**

Alcohol -

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-consumers-past-12-months-(-)>

Blood Pressure -

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/mean-systolic-blood-pressure-(age-standardized-estimate)>

Cholesterol -

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/mean-total-cholesterol-(crude-estimate)>

Global Population -

<https://data.worldbank.org/indicator/SP.POP.TOTL>

Human Development Index -

<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI>

Number of deaths attributed to non-communicable diseases, by type of disease and sex -

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/number-of-deaths-attributed-to-non-communicable-diseases-by-type-of-disease-and-sex>

Obesity -

<https://www.who.int/data/gho/data/indicators/indicator-details/GHO/prevalence-of-obesity-among-adults-bmi-=-30-(crude-estimate)-(-)>