DATA WRANGLING PROJECT

AIR POLLUTION AND DISEASE

By Ahmet Karabas, Sungmin Kim, Kieran Nicholl-Morris Isabelle Ruiz and Tuan Minh Dinh Van

INTRODUCTION

We all need to take a break once in a while from our day to day life. That may be reading a book, meeting some friends or even going outside for - what some may call - fresh air. But how fresh is that air? Breathing is something essential that we take for granted yet it can be profoundly impacted by the world around us. From our current lifestyle, whether it be the way we travel or build industrial facilities, to situations we are unable to control such as wildfires or volcanic eruptions. All these factors create the environment in which we breathe. But what does this mean for our health?

Among the diverse pathologies that can affect our lungs, one such group are the so-called chronic lower-respiratory diseases or CLRDs. These encompass conditions such as asthma, emphysema and constructive obstructive pulmonary disease (COPD) and are often diseases where the standard of care is minimising their effect on the lives of the individual rather than curing the condition (Prezant et al. 2008). According to the WHO, smoking is considered to be one of the major players in developing LRTDs. That being said, it also lists pollutants as an important player in the prevalence of these diseases (World Health Organisation, 2012).

We live in a world where it is inevitable to experience a polluting atmosphere. To counter this ever-increasing pollution, the European Union enacted the “Effort Sharing Decision” initiative to keep track of the emissions of the member states. However, has this policy made an impact on our daily lives?

To answer our research question, we will first look at which European countries have the highest and lowest disease-related death rates and the highest and lowest pollution rates. Furthermore, by also taking into consideration age and sex, we can evaluate if there is indeed a difference across different age groups (Age 0-64 and Age 65+) and different sexes. After gathering the data separately, we will determine if there is a relationship between disease and pollution.

RESEARCH QUESTION

How does the level of air pollution impact the death rate for lower respiratory disease?

SUBQUESTIONS

1. Which European Union countries have the highest/lowest death rate in chronic lower respiratory disease from 2005 to 2019?
2. Which European Union countries have the highest/lowest levels of air pollution from 2005 to 2019?
3. In two age categories (0-64 and 65+), are respiratory disorders more frequent in men or women?
4. Is there an overlap between the countries that have the highest/lowest incidence of disease and the highest/lowest levels of pollution?

DATA SOURCES

To investigate our research question, we obtained data from several sources.
Firstly, we collected air pollution data from the European Environment Agency (EEA), which provides independent environmental information on the European Union. We assumed that the degree of air pollution is proportional to the degree of greenhouse gas emission; therefore, we obtained the data on greenhouse gas emission.

Subsequently, data was collected to examine the level of Chronic lower respiratory diseases. This data was obtained through the World Health Organization (WHO) which measures the standard death rate (SDR) of CLRD per 100,000 inhabitants. *The standard death rate* is the death rate adjusted to standard age distribution ("Glossary: Standardised death rate (SDR) - Statistics Explained", 2022). This transformation offsets the variation between people's age and sex, improving comparability over time and between countries. This data had been separated into two age groups, 0-64 and 65. For this research, both age groups were merged, a process which will be detailed in the subsequent section.

The data was acquired from the on the following sources:

1. The data for ***greenhouse gas mission***: https://www.eea.europa.eu/data-and-maps/data/esd-3
2. The data for ***SDR of Chronic lower respiratory diseases***:
    a. Age (0-64) https://gateway.euro.who.int/en/indicators/hfamdb_162-sdr-0-64-chronic-lower-respiratory-diseases-per-100-000/
    b. Age (65+) https://gateway.euro.who.int/en/indicators/hfamdb_168-sdr-65plus-chronic-lower-respiratory-diseases-per-100-000/

DATA WRANGLING METHODS

DATA ACQUISITION

To respond to the given question, we decided to look at data from health and wellness open data sites such as WHO, CDC medical, Europe open data. This data needed to satisfy the following conditions: suitable format (ie: csv, excel, etc.), up to date, containing the necessary variables in a sufficiently large quantity. Said search resulted in 3 datasets which would allow us to determine the impact of air pollution on the death rate for CLRD. In the datasets about CLRD, we have the following columns: Measure code, AGE_GRP, SEX, COUNTRY_REGION and years from 1968 to 2019. These are CountryLong, year (from 2005 to 2019), ValueNumeric, Unit and Data_source in Greenhouse gas emissions dataset.

DATA CLEANING

Raw dataset usually contains missing data, duplicated records and the wrong data format. Therefore, checking, cleaning and formatting values is required to make the datasets usable for the research.

Firstly, the two CLRD death rate datasets from the two different age groups have the same shape and columns, because they are both from the same source. Therefore, they can be cleaned and processed at the same time by merging them into one dataset. After merging them as a huge dataset, the next step was to check and fill any missing values. This was done by using the *fillna()* function with methods forward fill (ffill) and backward fill (bfill) on all the row values. Although it can cause bias, the missing data are mainly concentrated in the pre-80s period, which is outside the scope of our research. There were no duplicated records. However, the dataset contained data from lots of countries and groups. As our research focuses on the EU countries, all groups and countries that fall outside of this scope were eliminated from the dataset. Then, country codes were translated (“COUNTRY_REGION”) to country names (“Country_long”) by using *country_converter()* function. Finally, the index needed to be reset, due to the merge step which caused repetition of several indice values. Column names were left unchanged as they provide a clear description.

The air pollution dataset required minimal data cleaning efforts.adjustments which were made to this dataset were the removal of cumulative data for the ‘EU-27’ and ‘EU-28’ groups. Country codes were added by translating these from the country names (“CountryLong”).

DATA MERGING

As mentioned in the previous section, we decided to merge the 2 datasets before cleaning due to the similarity in shape and columns. After concatenating and cleaning the two datasets, a master dataset was obtained, which contains no redundant columns, NaN values, or duplicated records. The reason for merging 2 datasets is to increase the amount of data and to append data from the 2 age groups: 0-64 and 65+. From there, we were able to compare mortality rates from chronic lower respiratory disorders across age groups and investigate the effects of air pollution. The third dataset already has the pollution values of EU countries by year, from 2005 to 2019. Thus, we can apply describe functions for analysis and start conducting data aggregation and visualisation.

DATA AGGREGATION AND VISUALISATION

To answer the sub-question about EU countries having the highest/lowest death rate in chronic lower respiratory disease, the average death rate values of EU countries in the period 2005-2019 were extracted from the 'sdr_master' dataset. Here, we considered 3 groups: age group 0-64, age group 65-more and both. Calculating the mean values by grouping, allowed us to determine the countries that have the highest/lowest average death rate. To visualise this, 6 bar plots were created. The results indicated that Denmark had the top death rate in general. Hungary had the highest mortality rate of respiratory disease for the age group of 0-64. Cyprus had the lowest value for the age group 0-64, while France showed the lowest rate in the age group 65+. When we compare these bar charts against the two age groups (0-64 and 65+), we see that the death rate for the 65+ group is much higher than the 0-64 group. For example, the highest death rate of the younger group is just around 12 deaths per 100,000, while for the 65+ group it exceeds 283 deaths per 100,000.

The second sub-question addresses the air pollution levels of EU countries from 2005 to 2019. To compare air pollution levels across European countries, we averaged the above values using *the groupby* function and plotted a bar plot for visualisation. By using *nlargest* and *nsmallest* functions, we determined that Germany, France and the United Kingdom have the highest levels of air pollution, while Malta, Cyprus and Estonia have the lowest values.

To provide insights into the third sub-question, we aimed to identify if there was a gender group in which CLRDs were more common, based on both age groups (0-64 and 65+). As the dataset has units of death per 100,000, we summed the mortality values for each gender and age group for the countries that were considered. These results were portrayed on pie charts to visualise the difference in death rate between sexes among the 2 age groups. These charts were constructed for all EU countries as well as both the top 3 countries with the highest mortality. For all EU countries, the rate of people dying from respiratory disorders was twice as high in men as in women for the 0-64 age group, 66.2 and 33.8%, respectively. Similarly, in the 65+ age group, women make up about 30%, while this number is 70% in men. On the other hand, for the top 3 death rate countries, about 40% of people who die of chronic lower respiratory tract disorders are female across all age groups.

Although a slight downward trend in average pollution values and average disease death rate can be observed throughout the years, when considering the time period for our research (2005 to 2019), there appears to be no relationship between air pollution and the standardised death rate of respiratory diseases. The analysis illustrates that countries with high levels of air pollution do not appear in the list of countries with the highest mortality rates and vice versa. Similarly, France, Latvia and Estonia are the top 3 countries with the lowest death rate, while the top 3 countries with the lowest levels of pollution are Malta, Cyprus and Estonia.

***All the visualisations can be found in the Jupyter file*

REASONING BEHIND VISUALISATION

The plotting library "Matplotlib" and the data visualisation module "Seaborn" were used to visualise the data, answer the sub-questions and, finally, the primary research topic. Seaborn is a useful tool that enhances the graphical attractiveness of Matplotlib plots and expands the Matplotlib library with built-in themes. It has a high-level interface for creating visually appealing and instructive statistics visuals. Together, these two strong technologies allow us to view the data. Bar and pie charts were primarily used to illustrate the outcomes of the research. The reason behind this is that the datasets' values mostly consisted of both numerical and category variables.

CONCLUSION

While smoking has been considered a leading contributor to CLRD, more and more evidence has been researched that exposure to outdoor air pollution affects the occurrence of CLRD. Using this as the basis of our hypothesis, this research was conducted to define a possible relationship between chronic lower respiratory disorders and air pollution. By examining the trend in chronic lower respiratory disorders from 1960 to the early 2000s within Europe, we tried to investigate whether the trend of those two aligned.

The data shows that the disease incidence has experienced a gradual decrease over time. Specifically, when compared between 1968 and 2019, a decrease from almost half the rate it was can be observed. Upon examination of the rates of CLRD on a country specific basis from the period 2005 to 2019, Denmark, Ireland, Hungary, the United Kingdom, and the Netherlands recorded the highest disease rate in order. In contrast, Cyprus, Greece, Italy, France, and Slovenia showed the lowest disease rate in order. Furthermore, when we examine the pollution rates among EU countries, Germany, France, United Kingdom, Italy, and Spain were ranked as high pollution rate countries. Comparing pollution to CLRD rates, there was no observable overlap between these two. As an illustrative example, Italy and France possess the lowest rate of CLRD, while possessing some of the highest pollution levels. One country did show an overlap between both the rate of pollution and the rate of CLRD, namely the United Kingdom. However, considering that the relationship does not hold for any of the other EU countries it is likely that this overlap is due to chance rather than a true relationship. As such, there is insufficient evidence to support our hypothesis.
After analysing the data, we found no apparent connection between the death of CLRD and pollution rate. Thus, with the data we gathered, we conclude that there is no effect of pollution on the incidence of disease.

That being said, there are some limitations to our research. The first one, namely, is that the research was not conducted systematically. This refers to the fact that no substantial statistical techniques were employed to draw conclusions. As such, this research cannot make any comments on significant differences. Therefore, it is difficult to define the relationship between CLRD and air pollution by simply comparing only two variables. Furthermore, this research defined pollution as a measure driven by the amount of greenhouse gasses. It is possible that if a different definition, which accounts for other compounds, is taken that a relationship might emerge. Finally, the time frame for the research might also play a role. This research considered the 2005 - 2019 period as it covered under the Effort Sharing Decision plan from the EEA. Obtaining pollution data before this period that is uniformly reported would not be feasible option given the scope of this research, That being said, it is possible that if such data could be acquired, and thus broaden the time horizon of the analysis, that there might be an observable relationship between pollution and the rate of CLRD.

REFERENCES

1. Forum of International Respiratory Societies. The Global Impact of Respiratory Disease – Second Edition. Sheffield, European Respiratory Society, 2017.
2. *Glossary:standardised death rate (SDR)*. Glossary:Standardised death rate (SDR) - Statistics Explained. (n.d.). Retrieved February 4, 2022, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary%3AStandardised_death_rate_%28SDR%29#:~:text=The%20standardised%20death%20rate%2C%20abbreviated,age%20distribution%20of%20that%20population
3. *Greenhouse gas emissions under the Effort Sharing Decision (ESD)*. European Environment Agency. (2021, December 16). Retrieved February 4, 2022, from https://www.eea.europa.eu/data-and-maps/data/esd-3
4. Prezant, D. J., Levin, S., Kelly, K. J., & Aldrich, T. K. (2008). Upper and lower respiratory diseases after occupational and environmental disasters. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, *75*(2), 89–100. https://doi.org/10.1002/msj.20028
5. World Health Organization. (n.d.). *Who European Health Information at your fingertips.* World Health Organization. Retrieved February 4, 2022, from https://gateway.euro.who.int/en/indicators/hfamdb_162-sdr-0-64-chronic-lower-respiratory-diseases-per-100-000/
6. World Health Organization. (n.d.). *Who European Health Information at your fingertips.* World Health Organization. Retrieved February 4, 2022, from https://gateway.euro.who.int/en/indicators/hfamdb_168-sdr-65plus-chronic-lower-respiratory-diseases-per-100-000/