Heatlh concerns in Germany

Vojtěch Sýkora [*1] **Denis Kovačević** [*2] **Nam Nguyen** [*3]

Abstract

Even though medicine advances further each day, health risks are still present. In this paper, we will analyze the most common health concerns from 1990 to 2019 in Germany in comparison to the rest of the world, with a focus on more developed countries. We used data with multiple features, such as the number of incidences of diseases and indicators of lifestyle and healthcare systems. We will show which diseases are the most common in Germany and their statistical significance using a permutation test. We analyze the difference in lifestyle and healthcare indicators and their effect on the incidence rate of ischemic heart disease using a random forest model. In the end, we will show the importance of each feature. Code is available at https://github.com/sykoravojtech/IHD_germany_2024.

1. Introduction

Germany, one of the largest economies in the world with its advanced healthcare system, faces an intriguing paradox: its life expectancy lags behind other high-income countries. This discrepancy, as highlighted in the analysis by Jasilionis et al. (2023) in "*The underwhelming German life expectancy*," poses critical questions about the underlying factors contributing to this phenomenon.

The reasons behind this phenomenon are still undefined. This paper aims to add to the work of Jasilionis et al. (2023) focusing on cardiovascular diseases. In 2019, CVDs were the leading cause of death in Germany, accounting for 38% of all deaths. We will also put a special emphasis on ischemic heart disease (IHD), which is the most common cardiovascular disease in Germany.

[*]Equal contribution [1]Matrikelnummer 6636502, vojtech.sykora@student.uni-tuebingen.de, MSc Machine Learning [2]Matrikelnummer 6707752, denis.kovacevic@student.uni-tuebingen.de, MSc Machine Learning [3]Matrikelnummer 6608479, nam.nguyen-the@student.uni-tuebingen.de, MSc Machine Learning.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

2. Data and Methods

In this section, describe *what you did*. Roughly speaking, explain what data you worked with, how or from where it was collected, it's structure and size. Explain your analysis, and any specific choices you made in it. Depending on the nature of your project, you may focus more or less on certain aspects. If you collected data yourself, explain the collection process in detail. If you downloaded data from the net, show an exploratory analysis that builds intuition for the data, and shows that you know the data well. If you are doing a custom analysis, explain how it works and why it is the right choice. If you are using a standard tool, it may still help to briefly outline it. Cite relevant works. You can use the `\citep` (whole citation in parenthesis) and `\citet` (only year in parenthesis) commands for this purpose (MacKay, 2003).

2.1. Data

Obtaining non-sparse data for all countries in the world over a long time horizon is quite challenging.

We began by exploring ourselves if cardiovascular diseases are in Germany truly abnormal with comparison to the rest of the world. In all of our analysis we wanted to include a closer comparison with high-income countries since Germany is one of them and comparing countries with similar healthcare and gdp could lead to more concrete results. There is not a wide variety of enormous datasets including data about all countries over a long time period and the one that we decided to use for Figure 1 was the Global Burden of Disease study (for Health Metrics & , IHME). The filtering on their website makes it easy to download only the data we truly needed. We obtained information about the death rate and incidence rate of CVDs for all countries divided into 5-year age groups for years 1990-2019. The data also included a lower and upper bound for uncertainty. The only change we did outselves was leaving only the rows corresponging to the sum of all age groups.

Furthermore, CVDs are a group of diseases and to get to the root of the issue we needed to investigate which specific diseases from this group take up the majority of cases. For this we went back to (for Health Metrics & , IHME) and requested data about specific CVDs. There was too much data for one file which meant that we had to combine files

and filter out unwanted data as to leave us only with the number of deaths grouped by disease only for Germany.

Since a disease death rate and incidence rate may largely rely on the healthcare system of a country. Thus, we used healthcare expenditure as the percentage of GDP from (Bank, 2023) as to indicate whether a country has a high quality medical system.

When it comes to causes, CVDs are likely influenced by a range of factors. These include dietary habits, especially fat consumption as detailed in (Food & of the United Nations, 2020), where the data is measured in daily fat intake per person. Lifestyle choices such as alcohol consumption, referenced from (Data, 2022a), are quantified by annual sales of pure alcohol in liters per person aged 15 and older. The role of smoking in CVDs prevalence is also considered, with data sourced from (Data, 2022b), highlighting the percentage of the population over 15 years old who smoke daily. Additionally, the impact of an aging population on CVDs is a significant factor, as indicated by demographic data from (United Nations, 2022), suggesting potential correlations.

A key challenge on using data from various sources lies in consistency. They might differ in the number of countries, years of study, and some even contains missing data for different years. See table 1 for details. As a result, we ended up choosing a mutual set of 178 countries and a date range from 1990 to 2019 to perform analysis on, which balance between data quantity and the number of missing values. The joining process is done using the ISO-3 country code.

Data source	Number of countries	Year range	Percentage of missing data from 1990 - 2021
Global Burden Disease IHC	206	1990 - 2019	0%
Health expenditure	266	1960 - 2022	41%
Fat consumption	194	1961 - 2020	0%
Alcohol consumption	187	2000 - 2019	0%
Population age	238	1950 - 2100	0%

Table 1

3. Results

In this section outline your results. At this point, you are just stating the outcome of your analysis. You can highlight important aspects ("we observe a significantly higher value of x over y"), but leave interpretation and opinion to the next section. This section absoultely *has* to include at least two figures.

Figure (for Health Metrics & , IHME) with how it shows the real issue CVDs are in Germany compared to HIC and GLO. Talk about the permutation test.

CVDs are way more prominent for older people (perhaps a plot for inc rate in age groups).

so comparing with all of the world countries will not give a good reasoning for life expectancy. limit ourselves to some countries with high income and high avg age (we could make a bubble plot of some sort for it) (also helps that we don't have data for all the factors for all the countries).

4. Discussion & Conclusion

Use this section to briefly summarize the entire text. Highlight limitations and problems, but also make clear statements where they are possible and supported by the analysis.

Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

Notes

Your entire report has a **hard page limit of 4 pages** excluding references. (I.e. any pages beyond page 4 must only contain references). Appendices are *not* possible. But you can put additional material, like interactive visualizations or videos, on a githunb repo (use links in your pdf to refer to them). Each report has to contain **at least three plots or visualizations**, and **cite at least two references**. More details about how to prepare the report, inclucing how to produce plots, cite correctly, and how to ideally structure your github repo, will be discussed in the lecture, where a rubric for the evaluation will also be provided.

References

Bank, T. W. Current health expenditure (https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS/, 2023. Accessed: Jan 2024.

Data, O. Alcohol consumption. https://data.oecd.org/healthrisk/alcohol-consumption.htm, 2022a. Accessed: Jan 2024.

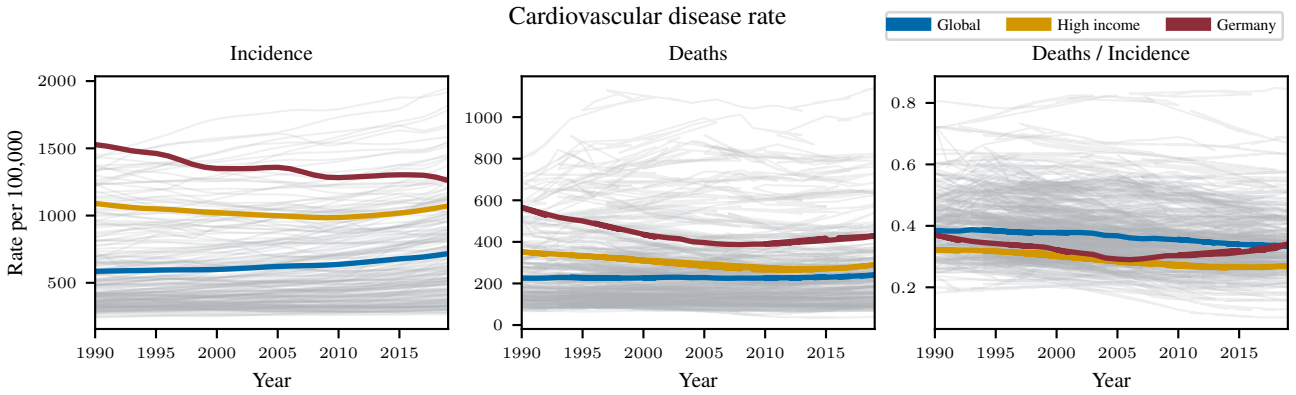

Figure 1. Effect of the cardiovascular diseases on the world over time. From left to right: incidence rate, death rate, and the ratio of death rate to incidence rate. The data is taken from the Global Burden of Disease study (for Health Metrics & , IHME). We specifically focus on the values for Germany in comparison to other high income countries and the world.

Data, O. Daily smokers. https://data.oecd.org/healthrisk/daily-smokers.htm, 2022b. Accessed: Jan 2024.

Food and of the United Nations, A. O. Daily per capita fat supply, 2020. https://ourworldindata.org/grapher/daily-per-capita-fat-supply, 2020. Accessed: Jan 2024.

for Health Metrics, I. and (IHME), E. Global burden of disease study 2019 (gbd 2019) data resources. http://ghdx.healthdata.org/gbd-2019, 2020. Accessed: [Insert date here].

Jasilionis, D., van Raalte, A. A., Klüsener, S., and Grigoriev, P. The underwhelming german life expectancy. *European Journal of Epidemiology*, 38(8): 839–850, 2023. ISSN 1573-7284. doi: 10.1007/s10654-023-00995-5. URL https://doi.org/10.1007/s10654-023-00995-5.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

United Nations, W. P. P. . Median age. https://ourworldindata.org/grapher/median-age?tab=table&time=earliest..2020, 2022. Accessed: Jan 2024.