# 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 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 underlying reasons are still undefined. This paper aims to add to the work of Jasilionis et al. (2023) focusing on cardiovascular diseases. In 2019, they 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, 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. 

Firstly, we will analyze cardiovascular diseases in Germany from 1990 to 2019 and compare them to the rest of the world (Section 2.1). One of the biggest obstacles in our analysis was the lack of data that matches our data for diseases in terms of time period and countries (Section 2.2). We present our model for the death rate of ischemic heart disease and its interpretation (Section 3). Finally, we discuss our results and limitations of our analysis (**??**).

## 2. Data and Methods

There is no single dataset that contains all the information we need. Thus, we had to combine data from multiple sources. In this section, we will describe the data we used and show exploratory analysis of the data. We will also explain the methods we used to analyze the data.

### 2.1. Cardiovascular data

The first step in our analysis was to find out whether cardiovascular diseases are a problem in Germany. We used the data from the Global Burden of Disease study (Institute for Health Metrics and Evaluation (IHME), 2020) to compare the incidence and death rate of cardiovascular diseases in Germany to other high-income countries and the world. The data also contains the incidence and death rate for specific age groups which we used to compare the age distribution of cardiovascular diseases in Germany to other countries. In Figure 1 we can see that Germany has very high incidence and death rate of cardiovascular diseases. We performed a permutation test to check whether the incidence rate in Germany is statistically significant. In the test, we compared the average incidence rate from 1990 to 2019 of Germany to the average incidence rate of the world. Over a million permutations, we found that the p-value is $\approx 0.045$. In the plot on the right we can also see that the ratio of death rate to incidence rate is much more similar. There appears to be no major difference between the global average and the average of high-income countries, which should have a better healthcare system. This suggests that the effect of the quality of healthcare on the ratio is not very significant.

Since cardiovascular diseases are just a group of diseases, we also looked at the most common cardiovascular disease in Germany. The top two are ischemic heart disease and stroke (Figure 2). We found out that the incidence rate of

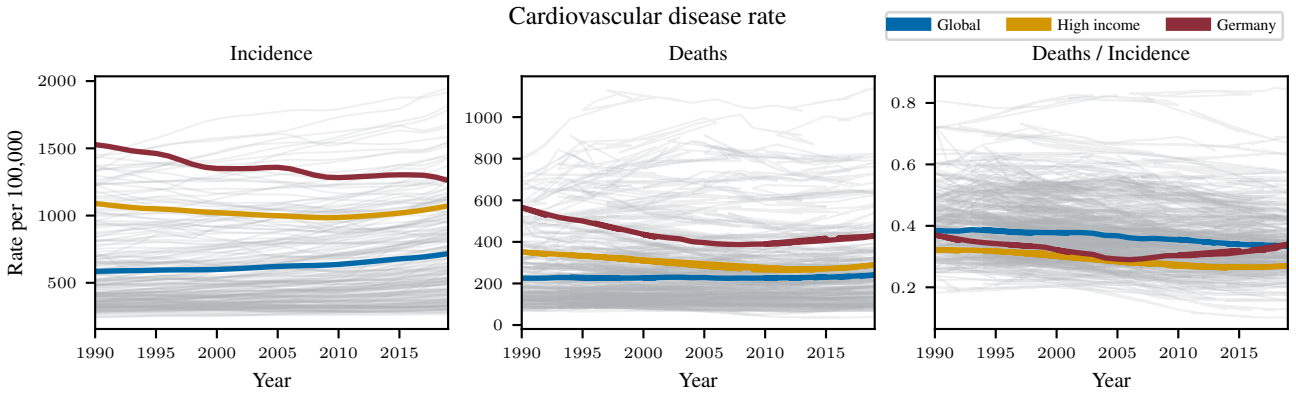


*Figure 1.* Cardiovascular diseases in the world over time. From left to right: incidence rate, death rate, and the ratio of death rate to incidence rate.

ischemic heart disease in Germany is also very high, as shown in INSERTURL. The incidence rate of stroke in Germany is way more similar to the global average, as shown in INSERTURL. This suggests that ischemic heart disease is the main cause of the high incidence rate of cardiovascular diseases in Germany.

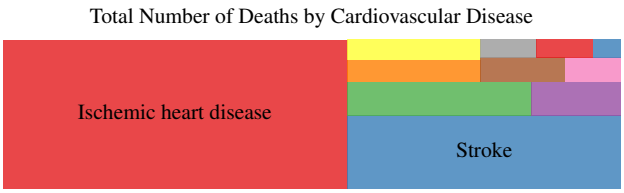


*Figure 2.* Total number of deaths by specific cardiovascular diseases in Germany. Only the top 2 are labeled because the rest are much less common. The most common cardiovascular disease in Germany is by far the ischemic heart disease, followed by stroke.

### 2.2. Data sources for factors

Our next step was to use different data sources to find out the underlying factors that have an effect on the incidence rate of ischemic heart disease. We looked at some of the most common risk factors: alcohol consumption and fat consumption. We also included health expenditure as a proxy for the quality of healthcare. Because of the distribution of incidence rate depending on age, as shown in Figure 3, we also included the median age of the population as a factor.

The biggest obstacle was the lack of data that is available for all countries over a long time period. For different factors we had to use different data sources. The data for health expenditure (per capita) and alcohol consumption (liters of pure alcohol per person aged 15 and older) is from OECD (Organisation for Economic Co-operation and

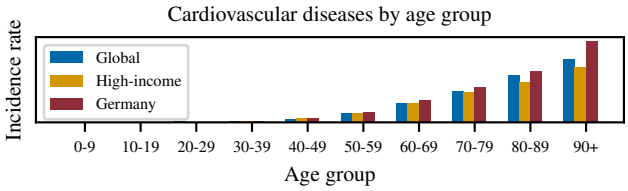


*Figure 3.* Average incidence rate of cardiovascular diseases for 10 year age groups. The incidence rate is much higher for older age groups, but always the highest in Germany.

Development). The data for population age is from United Nations, World Population Prospects (2022) (2022), while the data for fat consumption (grams per person per day) is from Food and Agriculture Organization of the United Nations (2020). The data varies in the time period it covers, the number of countries it contains, and the number of missing values, as shown in Table 1. We used the data from 1990 to 2019, which is the time period that is available for the diseases.

*Table 1.* Overview of the data sources. The missing data is calculated for the time period from 1990 to 2019.

| Data source | No. of countries | Missing data |
|---|---|---|
| Ischemic heart disease | 206 | 0% |
| Health expenditure | 266 | 41% |
| Fat consumption | 194 | 0% |
| Alcohol consumption | 187 | a lot% |
| Population age | 238 | 0% |

We made animations and visualizations of the data with enough data points as part of the exploratory analysis. The animations are available at INSERTURL. For the rest of the project we merged the data sources into one dataset,

with the goal of modeling the death rate of ischemic heart disease. Because the data sources did not only have different countries, but also different country names we had to do some data preprocessing. The main part was generating ISO-3 country codes for every dataset. We decided on not averaging the missing values, but instead using the data as it is.

## 3. Results

To model the death rate of ischemic heart disease we tried using a linear model with different transformations of the data, but the results were not satisfactory. Because of this, we decided to use random forest regression. We used the `RandomForestRegressor` from the `scikit-learn` library (Pedregosa et al., 2011) with grid search to find the best parameters. Our best model had the $R^2$ score of $\approx 0.82$ on the test set, while the $R^2$ score of the linear model was $\approx 0.39$. The value represents the proportion of the variance that is explained by the model. We interpret the model using the SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017) (see Figure 4). SHAP values show how the impact of a feature on the model's output. The color gradient from pink to blue indicates the feature's value from low to high, showing how each feature contributes differently based on its value.

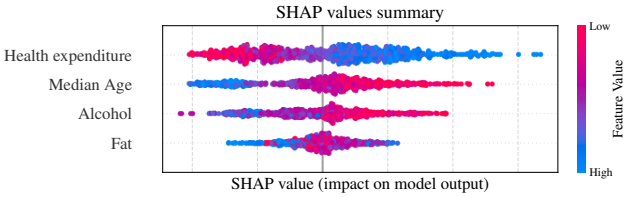


*Figure 4.* Summary of SHAP values for a predictive model of disease death rate. Each dot's color represents the feature's value (red high, blue low), and its position represents the impact on the model's output. The impact is measured by the SHAP value, which represents the feature's contribution to a change in the model's output.

We can see that the effect of fat consumption is mixed and not very easy to interpret. The effect of the other three features (Healthcare spending, alcohol consumption, and median age) is more clear. For that reason, we decided to use only those three features in our final plot (see Figure 5) in which we show the combined effect of the three features on the death rate. The median age is divided into three groups, under 19, 19-38, and over 38. The size of the bubble represents the age group. The position of the bubble represents the healthcare spending and the alcohol consumption. The color represents the death rate from ischemic heart diseases. Counries with higher alcohol consumption (on the right) appear to have higher death rates. It is clear that Germany is among the highest investors in healthcare, but it is also among the countries with the highest alcohol consumption. The median age is in the third group.

## 4. Discussion & Conclusion

In this project, we analyzed cardiovascular diseases, in Germany, but also in general. We showed that the incidence rate of cardiovascular diseases in Germany is statistically significantly higher than the world average. We analyzed some of the possible factors that could influence the death rate of ischemic heart disease, which is the most common cardiovascular disease in Germany. Our model showed that the death rate of ischemic heart disease is influenced by healthcare spending, alcohol consumption, and median age, while the effect of fat consumption is mixed and not very easy to interpret. The fact that the effect of fat consumption is not clear is surprising, because it is a well-known risk factor for cardiovascular diseases. We found that lower median age, lower alcohol consumption, and higher healthcare spending all lead to lower death rates, which was to be expected. Comparing the German healthcare spending, as well as the ratio of death and incidence rates, to the world average, we didn't find any fault in the German healthcare system. We also found that the German alcohol consumption is higher than the world average (nearly three times as high as the world average), which could be a reason for the higher death rate. The limitation of our analysis is that we didn't analyze all the possible factors that could influence the death rate of ischemic heart disease. Such factors could be smoking, vegetable consumption, or physical activity. We focused only on cardiovascular diseases, but it would be interesting to analyze less prevalent diseases as well. Another possible extension of this project would be to analyze if there is any genetic predisposition for cardiovascular diseases in Germany, using a genome-wide association study (GWAS).

## Contribution Statement

This project was done in a group of three. Because of the amount of different data sources, there is no clear division of work. All of us worked on the data collection and analysis. Building on that work, each of us also produced visualizations for their part of the analysis. Writing of the report was split between the group members. Vojtěch Sýkora wrote the abstract and the introduction. Nam Nguyen wrote the data and methods section. Denis Kovačević wrote the results and the discussion section.

## References

Food and Agriculture Organization of the United Nations. Daily per capita fat supply, 2020.

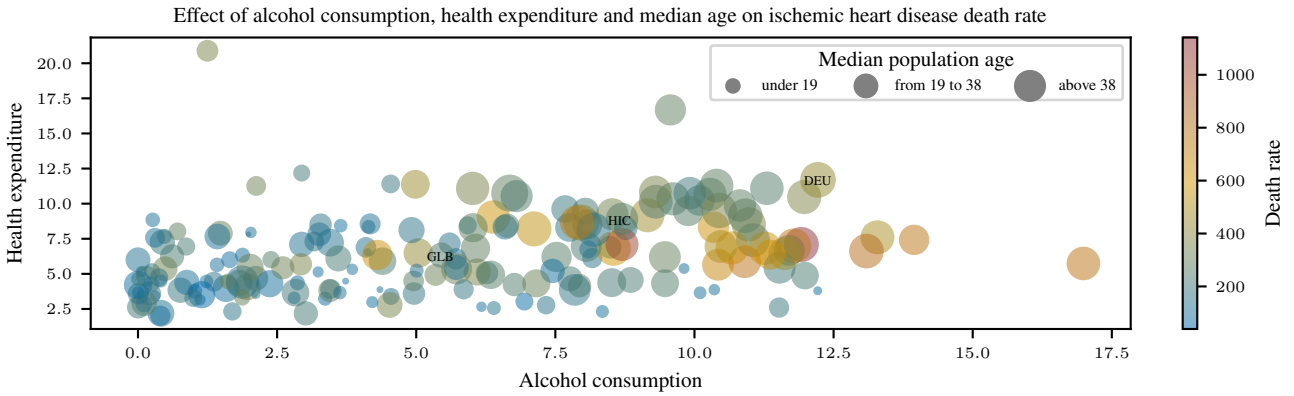


*Figure 5.* The combined effect of healthcare spending, alcohol consumption, and median age on the death rate of ischemic heart disease. Special emphasis to the comparison between Germany, high income countries, and the world.

https://ourworldindata.org/grapher/daily-per-capita-fat-supply, 2020. Accessed: Jan 2024.

Institute for Health Metrics and Evaluation (IHME). Global Burden of Disease Study 2019 (GBD 2019) Data Resources. http://ghdx.healthdata.org/gbd-2019, 2020. Accessed: Jan 2024.

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.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

OECD (Organisation for Economic Co-operation and Development).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

United Nations, World Population Prospects (2022). Median age. https://ourworldindata.org/grapher/median-age?tab=table&time=earliest..2020, 2022. Accessed: Jan 2024.