# Evolution of the number of deaths by pathologies over thirty years using GHDX Health Database

António Santos
up202008004@up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

José Osório
up202004653@up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

Mariana Teixeira
up201905705@up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

Pedro Silva
up202004985@up.pt
Faculty of Engineering of University of Porto
Porto, Portugal

## ABSTRACT

Studying population health has become more significant in recent years with the prevalence of historical and current medical conditions worldwide. It is essential to visualize and comprehend historical and geographical data, as it helps in predicting emerging medical conditions and preparing customized approaches for each country's population.

This project aims to compile information on the incidence of medical conditions and epidemics globally, with data supplied by the GHDX Health Database[4] and Wikipedia[7]. We seek to fulfill prospective search tasks such as showing incidence rates of given conditions in different parts of the world, which, in tandem with the aid of other economic and demographic indicators, can help to paint a picture of a country's development and history.

## CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; *Data cleaning*; *Geographic information systems*;

## KEYWORDS

Information systems, Digital libraries, Digital archives, World health

## 1 INTRODUCTION

During the last years, there have been more and more medical research and technology advances, allowing the availability of a substantial growing volume of medical data about each patient and its context. However, some of it is not so easily found, and when it is, the format in which the data is presented can be sub-optimal. We will use this report to document our process of retrieving and processing information about pathologies and their evolution in terms of mortality, across several countries and years.

This topic is highly relevant in current times, since there is a tendency to focus only on the most recent pandemics, while there is much that history can teach us to avoid future tragedies. For these reasons, the group decided to tackle the issue.

## 2 DATASET

Data for this project is an aggregation of the GBD results over thirty years, between 1990 and 2019, sourced from the GHDX Health Database, mixed with descriptions of the medical conditions pulled from Wikipedia. GBD stands for Global Burden of Disease, a study providing "a comprehensive estimate of mortality and disability across countries, time and age. It quantifies health loss from hundreds of diseases, injuries, and risk factors"[6]. The data is available free of charge for everyone as comma-separated values (*.csv*), containing millions of rows of data.

### 2.1 Collection & Preparation

The data retrieval pipeline is shown in Appendix I[3].

As with any pipeline, the first step is to collect the data. We download the dataset from the GHDx database, which arrived in a group of distinct .csv files and, with the use of a Python script and the libraries Matplotlib and Pandas, aggregate them into a single **global.csv** file. From there, we extract causes, obtaining a brief description of them through the Wikipedia API, storing them into a **cause_description.csv** file, as well as countries, pulling the ISO country, or *CCA3*, codes from the REST Countries API and using it to obtain the geographic information of said country in the *.geojson* format from InMagik's GitHub page. The countries and their respective codes are stored into a **countries.csv** file, while the *.geojson* files are kept separate.

It is important to note that the original dataset is extremely large, and since we are only interested in some specific attributes of the data, we don't extract information regarding, percentage and rate values, as we only focus on the "raw" estimate number of deaths.

After retrieving, joining and filtering the data in the multiple *.csvs* into a single global file we gather information that helps us characterize and describe the dataset and its contents. Furthermore, we aggregate data by country, year and cause.

Structuring the data is relatively simple seeing as all of the information is closely related. Essentially, each data entry is an *Occurrence*, which is the value of the number of deaths, alongside the upper and lower confidence bounds, of a certain *Cause* in a *Country* during an *Year*, the following domain model[**??**] illustrates these relationships.

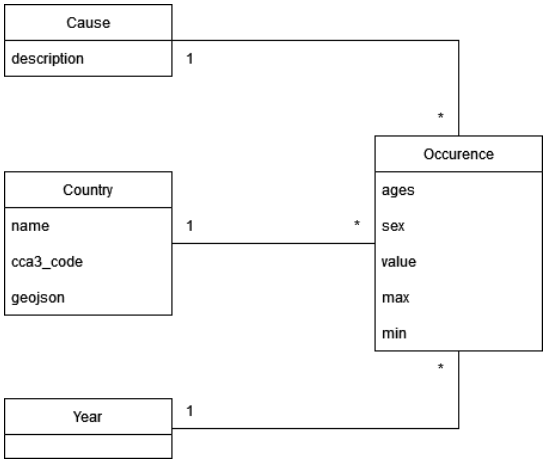



**Figure 1: Domain Model.**

With the information retrieved from Wikipedia, we do some textual analysis to better characterize the information, which allows us to evaluate word frequency and help us define what may be some interesting textual search tasks, as well as present this information in a more concise and human readable way, through bar plots and word clouds. We use several Python libraries to aid in building these visualizations, namely, NLTK[1], for natural language processing of the descriptions of each medical condition, seaborn[2] and WordCloud[3], to create plots and word clouds.

## 2.2 Characterization

The dataset contains 1481040 rows, with 16 columns[1].

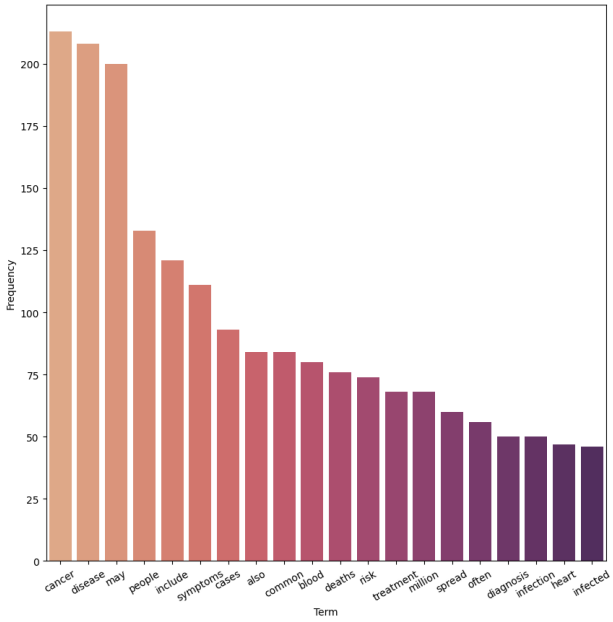
**Table 1: Dataset rows and data types.**

| | |
|---|---|
| integer | measure_id, location_id, sex_id, age_id, cause_id, metric_id, year |
| float | val, upper, lower |
| string | measure_name, location_name, sex_name, age_name, cause_name, metric_name |

There are no missing values in the dataset and all values are in the same format, as such there is no need to clean data or replace values. Whilst the *IDs* may seem redundant we are currently keeping them as they could prove to be useful in a later stage. In total there are: 1 measure, the number of deaths, 204 locations, 1 sex, as there is no distinct values for sexes, 3 age groups, 85 causes/diseases and 30 years.

Multiple data measures are included, with the number of deaths being the most relevant. Alongside this, useful data to determine factors of discrimination of a given illness, such as age groups, are also present. The plots presented below aim to demonstrate some of the patterns within the data.

**Figure 2: Diagram illustrating the most used terms in the database.**




In figure 2, cancer is a prevalent term due to the existence of various cancer types. Additionally, keywords related to pathologies frequently appeared, such as disease, symptoms, cases, blood, and treatment. The word deaths was not a surprise, given that we were only dealing with conditions that caused fatalities.

The most lethal medical conditions of our dataset are represented in the table 2 below:

**Table 2: Top 10 deadliest medical conditions in the dataset.**

| Cause Name | Estimated Deaths |
|---|---|
| Ischemic heart disease | 9,132,782.32 |
| Stroke | 6,549,283.15 |
| Chronic respiratory diseases | 3,972,676.69 |
| Diabetes and kidney diseases | 2,986,391.52 |
| Digestive diseases | 2,556,207.69 |
| Lower respiratory infections | 2,491,470.60 |
| Neurological disorders | 2,220,132.38 |
| Maternal and neonatal disorders | 2,077,672.60 |
| Tracheal, bronchus, and lung cancer | 2,041,697.54 |
| Diarrheal diseases | 1,533,773.22 |

The prevalence of general medical conditions is higher, such as chronic respiratory diseases and maternal and neonatal diseases, which encompass multiple specific conditions. We can also observe that the most fatal medical condition is Ischemic Heart Disease.

**Table 3: Most dominant conditions in a given country.**

| Country | Cause Name |
|---|---|
| China | Stroke |
| India | Ischemic heart disease |
| Russian Federation | Ischemic heart disease |
| United States of America | Ischemic heart disease |
| Ukraine | Ischemic heart disease |
| Pakistan | Maternal and neonatal disorders |
| Indonesia | Stroke |
| Germany | Ischemic heart disease |
| South Africa | HIV/AIDS |

The table 3 displays the deadliest illness in different countries. Although Ischemic Heart Disease is the most common cause of death, some underdeveloped countries experience a higher number of fatalities due to HIV/AIDS and maternal and neonatal disorders as a result of inadequate healthcare facilities.

**Table 4: Country with most deaths by a given condition.**

| Cause Name | Country |
|---|---|
| Brain and central nervous system cancer | China |
| Breast cancer | China |
| Malaria | Nigeria |
| Mental disorders | Japan |
| Stroke | China |
| Tuberculosis | India |
| Varicella and herpes zoster | India |
| Yellow fever | Nigeria |
| Zika virus | Brazil |

We concluded some countries have a higher number of deaths, as a result of larger populations, such as China and India. One particular case might be Japan, which has a higher prevalence of mental disorders, possibly related to a strict work culture.

## 3 PROSPECTIVE SEARCH TASKS

After completing our information search system, some prospective search scenarios can be fulfilled, in a way that is fast and condenses a great amount of relevant information together. These scenarios can be as follows:

- I want to know which medical condition took the most amount of lives in Portugal, in 2018.
- I want to know the 10 deadliest medical conditions in the United States, throughout the available time period.
- I want to discover the year in which a certain condition took the most amount of lives.
- I want to find out which country suffered the most with cancer of any type after 2005.
- I want to find out which medical condition, that relates to "vaccines", caused the most deaths in individuals aged 55 years or above.

## 4 FUTURE WORK

Having established our processed data with the aforementioned pipeline, we aim to implement an information search system that can fulfill every search task described above, and provide proper visualizations of everything obtained with this system, creating an interesting and user-friendly experience.

## REFERENCES


[1] 2003. NLTK. https://www.nltk.org/
[2] 2003. seaborn. https://seaborn.pydata.org/
[3] 2003. WordCloud. http://amueller.github.io/word_cloud/
[4] Global Burden of Disease Collaborative Network. 2020. Global Burden of Disease Study 2019 (GBD 2019) Disease and Injury Burden 1990-2019. https://ghdx.healthdata.org/record/ihme-data/gbd-2019-disease-and-injury-burden-1990-2019
[5] Max Roser, Hannah Ritchie, and Fiona Spooner. 2021. Burden of disease. https://ourworldindata.org/burden-of-disease
[6] University of Washington. 2019. Global Burden of Disease (GBD). https://www.healthdata.org/research-analysis/gbd [Online; accessed 11-October-2023].
[7] Wikipedia contributors. 2004. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/ [Online; accessed 11-October-2023].

# APPENDIX I

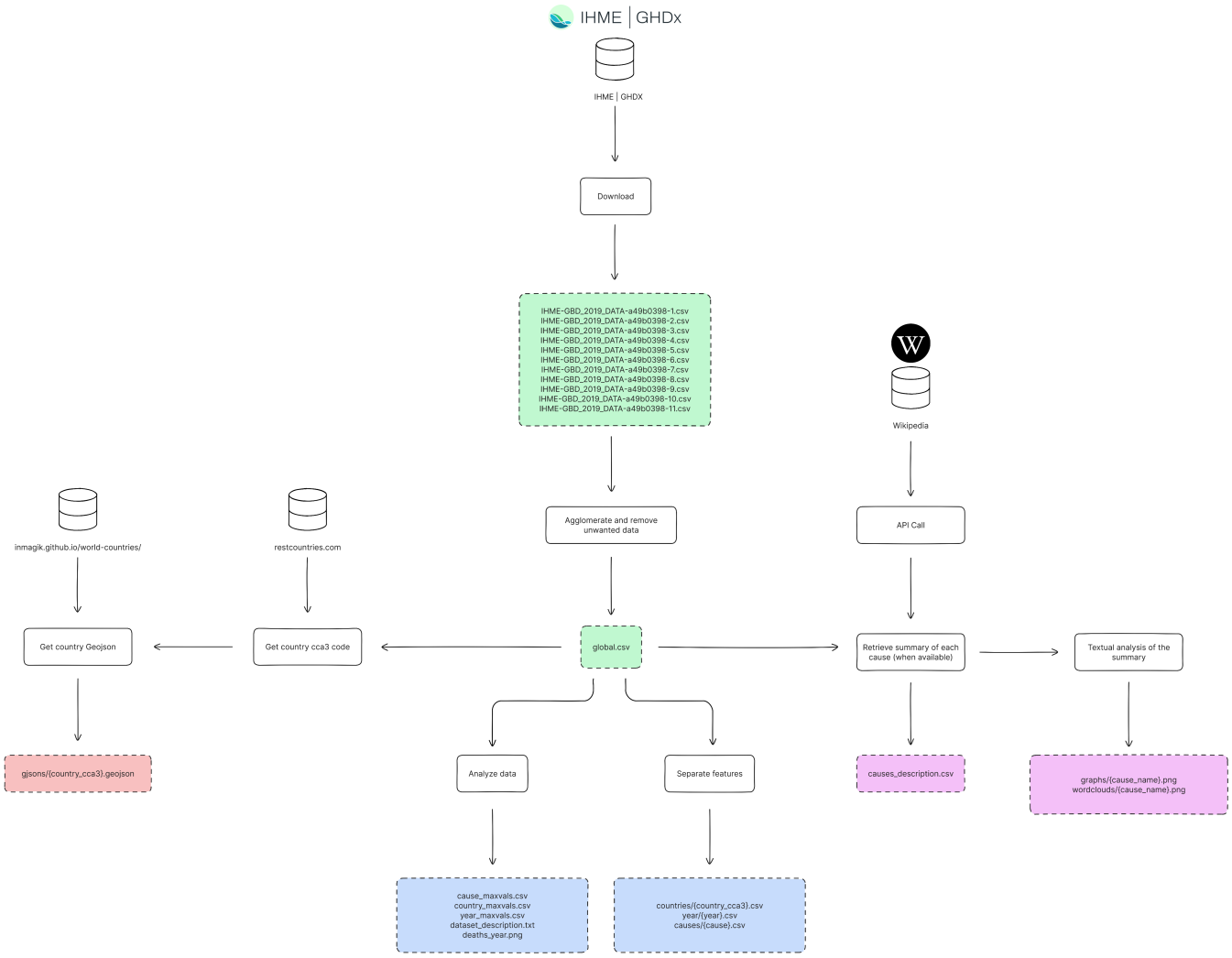



**Figure 3: Diagram illustrating the pipeline.**