DM Project

Karla Otoude and Chloé Marquis

Link to our github project

Our Github Project

Research question

How does the level of social protection impact the general population health ?

We want to identify if a stronger social and health protection (quantified by different variables such as health expenditure, tax revenues, etc) has an impact on the avoidable mortality rate. In order to do so we will analyse data collected in OECD countries from 2000 to 2023.

There are two goals in our analysis : The first is to understand the global evolution of the data and the reasons behind its trends. The second is to study the Covid-19 crisis. We believe it revealed how necessary having a proper social protection is to limit the number of deaths during a pandemic.

Method

To conduct our study we will be using a **linear mixed model (LMM)** with lags. Such a model will allow us to properly study the OECD countries taking into account their differences. It also a fitting model since we are working with time-based data.

Considering our research question, we will use multiple datasets and variables.

For explanatory variables, in order to estimate the strength and quality of a country's health protection scheme, we will have : - Health expenditure - Social expenditure aggregates - Number of hospitals beds

We also have more general variables such as : - GDP per capita - Unemployment rate - Human development index - Life expectancy They will be used as control variables, since

richer countries tend to have better living conditions which may lead to a generally healthier population, without the need for a strong health protection.

Finally, the explained variable will be : - Avoidable mortality It will be used to estimate the population's general health.

Links to the sources and description

OECD

The Organisation for Economic Co-operation and Development (OECD) is an international organisation created in 1948. Its goal is to allow various countries to answer common problems and coordinate domestic and international policies. OECD members share data and receive analysis from experts on social, economic and environmental challenges.

Health expenditure (OECD)

Health expenditure (OECD)

This dataset details the annual health expenditure and financing as a percentage of GDP for 52 countries. The data is available for a period ranging from 2000 to 2023. It was computed using financial flows related to the consumption of healthcare goods and services from different health providers such as Hospitals or Residential long-term care facilities.

Social expenditure aggregate (OECD)

Social expenditure aggregate (OECD)

This dataset includes statistics on public and private social expenditure at program level (e.g. Old age, Survivors, Incapacity-related benefits, Health, etc) as a percentage of GDP. It covers 38 OECD countries for the period 1980-2021/23 and estimates for aggregates for 2022-24.

Avoidable mortality (OECD)

Avoidable mortality (OECD)

This dataset describes the number of "avoidable deaths" per 100,000 inhabitants annually for the period 2000-2024 for 46 countries. It contains both "preventable mortality" and "treatable (or amenable) mortality". The first refers to deaths that can be avoided through effective public health and primary prevention interventions. The second refers to timely and effective health

care interventions, including prevention and treatment.Both indicators refer to premature mortality (under age 75).

Number of hospital beds (OECD)

Hospital beds (OECD)

This dataset provides data on the number of total hospitals beds by 1,000 inhabitants by function of healthcare and by type of care (ie. somatic or psychiatric care) for the period 2000-2023 for 48 different countries. Total hospital beds are the sum of the following categories: Curative care (acute care) beds in hospitals, Rehabilitative care beds in hospitals, Long-term care beds in hospitals, All other beds in hospitals not elsewhere classified.

World bank

The world bank group, is a family of five international organizations created in 1944 which help developing countries through leveraged loans. These loans are given for the purpose of pursuing economic development only. 189 countries are currently members of the World bank group.

GDP per capita in 2015 $US (World Bank)

GDP per capita (World Bank)

The GDP per Capita dataset presents the GDP per capita (constant 2015 US$) from 1960 to 2023 for 70 countries. This indicator is expressed in constant prices, meaning the series has been adjusted to account for price changes over time. The reference year for this adjustment is 2015. This indicator is expressed in United States dollars.

Life expectancy (World Bank)

Life expectancy (World Bank)

Life expectancy at birth indicates the number of years a newborn baby would have to live if the general rules of mortality at the time of birth were to remain the same throughout its life. Here it is measured by the world bank from 2000 to 2023.

IMF

The International Monetary Fund (IMF) is an international financial institution and a specialized agency of the United Nations created in 1944. It is composed of 191 member countries. Its goal is to help its members achieve sustainable growth and to encourage the expansion of trade. The IMF supports certain economic policies that promote financial stability and monetary cooperation.

Unemployment rate (IMF)

Unemployment rate (IMF)

Instructions for the data : In the Data Explorer : - Time period : Custom : 01/01/2000 to 12/31/2024

This IMF dataset describes the percentage of the labor force that is unemployed and actively seeking employment. It is available from 1980 to 2025 and features 120 geographic regions.

UNDP

The United Nations Development Programme is a United Nations agency tasked, since 1965, with helping countries eliminate poverty and achieve sustainable economic growth and human development. They intervene by helping countries develop policies, leadership skills, partnerships and institutional capabilities.

Human development index (HDI) (UNDP)

HDI (UNDP)

Instructions for the data : - Filter by index : HDI - Filter by indicator : Human development index (value) - Filter by year : 2000 - 2023 - Filter by group/region/country/territory : Select all countries

This dataset describes the human development index, per country (195), annually (1990-2023). It is a summary measure of average achievement in 3 dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living. The HDI is between 0 (low human development) and 1 (very high human development).

Loading the packages

```
library(here)
here::i_am("Project-Data-Management.Rproj")
```

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(vroom)
library(readxl)
```

Loading the datasets and summary

Health expenditure

```
health_expenditure <- vroom(here("data", "OECD.ELS.HD,DSD_SHA@DF_SHA,1.0+.A.EXP_HEALTH.PT_B1
```

```
Warning: One or more parsing issues, call `problems()` on your data frame for details,
e.g.:
  dat <- vroom(...)
  problems(dat)
```

```
problems(health_expenditure)
```

```
# A tibble: 2 x 5
    row   col expected           actual            file
  <int> <int> <chr>              <chr>             <chr>
1  1116    39 1/0/T/F/TRUE/FALSE P                 /Users/chloemarquis/Library/~
2  1116    40 1/0/T/F/TRUE/FALSE Provisional value /Users/chloemarquis/Library/~
```

There are issues with columns 39/40. Opening the file in Excel tells us that these columns are mostly empty and "P" appears on row 1116. This is just a comment on the data that is not useful to us in the project and it doesn't impact the variables of interest. Columns 39 and 40 will just delete these columns during the data cleaning process. Therefore, we load the data without resolving this issue.

We compute summary statistics including the percentage of missing values. Although several datasets show a relatively high percentage of missing values, this does not significantly affect our final merged dataset as our analysis relies mostly on 2 columns (time period and observed value) of each dataset.

```
health_expenditure |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(health_expenditure),
            "Percentage of NA" = paste0(round(sum(is.na(health_expenditure)) / (n() * ncol(he
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
1246	46	25.62%

Avoidable mortality

```
avoidable_mortality <- vroom(here("data", "OECD.ELS.HD,DSD_HEALTH_STAT@DF_AM,1.1+.A.AVM.DT_1

avoidable_mortality |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(avoidable_mortality),
            "Percentage of NA" = paste0(round(sum(is.na(avoidable_mortality)) / (n() * ncol(a
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
997	44	18.18%

GDP per capita

When opening the file in Excel, we see that the first 4 rows are empty. We need to skip the first 4 rows and use row 5 as column headers.

```r
gdp_per_capita <- vroom(
  here("data/API_NY.GDP.PCAP.KD_DS2_en_csv_v2_242568", "API_NY.GDP.PCAP.KD_DS2_en_csv_v2_2425
  delim = ",",
  skip = 4,           # need to skip the first 4 rows
  col_names = TRUE,
)

gdp_per_capita |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(gdp_per_capita),
            "Percentage of NA" = paste0(round(sum(is.na(gdp_per_capita)) / (n() * ncol(gdp_pe
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
266	70	17.84%

Hospital beds

```r
hospital_beds <- vroom(here("data", "OECD.ELS.HD,DSD_HEALTH_REAC_HOSP@DF_BEDS_FUNC,1.0+..10P3

hospital_beds |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(hospital_beds),
            "Percentage of NA" = paste0(round(sum(is.na(hospital_beds)) / (n() * ncol(hospita
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
1034	38	36.41%

Life expectancy

The header is on row 5. We must delete the comments on the previous lines to properly load the dataset.

```
life_expectancy <- vroom(
  here("data/API_SP.DYN.LE00.IN_DS2_fr_csv_v2_11258", "API_SP.DYN.LE00.IN_DS2_fr_csv_v2_1125
  delim = ",",
  skip = 4,           # need to skip the first 4 rows
  col_names = TRUE)

life_expectancy |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(life_expectancy),
            "Percentage of NA" = paste0(round(sum(is.na(life_expectancy)) / (n() * ncol(life_
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
266	70	3.39%

Social expenditure aggregates

```
social_expenditure <- vroom(here("data","OECD.ELS.SPD,DSD_SOCX_AGG@DF_SOCX_AGG,1.0PT_B1GQ.ES

social_expenditure |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(social_expenditure),
            "Percentage of NA" = paste0(round(sum(is.na(social_expenditure)) / (n() * ncol(s
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
1046	34	11.9%

Unemployment

```
unemployment <- vroom(here("data",
                           "dataset_2025-11-21T14_31_58.442440069Z_DEFAULT_INTEGRATION_IMF.R

unemployment |>
```

```
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(unemployment),
            "Percentage of NA" = paste0(round(sum(is.na(unemployment)) / (n() * ncol(unemploy
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
122	32	4.05%

HDI

```
hdi <- read_excel("data/hdr-data.xlsx")

hdi |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(hdi),
            "Percentage of NA" = paste0(round(sum(is.na(hdi)) / (n() * ncol(hdi)) * 100, 2),
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA
4484	10	20%

Some data cleaning before merging

We first need to keep only the interesting variables (observations, time period and country code) in all our datasets. We will later merge all data sets using the country code (almost the same in all datasets, minor exception in the unemployment dataset). We will also convert the GDP per capita, life expectancy and unemployment data sets to long formats, so that all datasets are in the same format before merging. We also check the class of all variables before merging.

Health expenditure

```r
# selecting only interesting variables
health_expenditure <- health_expenditure |>
  select(REF_AREA, TIME_PERIOD, OBS_VALUE)

# checking the class of these variables
sapply(health_expenditure, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

Social expenditure aggregates

```r
# We keep only the total social security contributions, and not the detail such as "paid by e
social_expenditure <- social_expenditure |>
  select(REF_AREA, TIME_PERIOD, OBS_VALUE)

# checking the class of these variables
sapply(social_expenditure, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

Avoidable mortality

```r
avoidable_mortality <- avoidable_mortality |>
  select(REF_AREA, TIME_PERIOD, OBS_VALUE)

sapply(avoidable_mortality, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

Hospital beds

```r
hospital_beds <- hospital_beds |>
  select(REF_AREA, TIME_PERIOD, OBS_VALUE)

sapply(hospital_beds, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

Life expectancy

```r
# we select only the period we are working on : 2000 to 2023
life_expectancy <- life_expectancy |>
  select(`Country Code`, as.character(2000:2023)) |>
  rename(REF_AREA = `Country Code`) # renaming to harmonise with the other datasets

# pivot this data set to a long format
life_expectancy_long <- pivot_longer(life_expectancy, cols=as.character(2000:2023), names_to=
                                     values_to="OBS_VALUE")

# checking the class of variables
sapply(life_expectancy_long, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character" "character"   "numeric"
```

```r
# Changing the TIME_PERIOD variable to numeric
life_expectancy_long$TIME_PERIOD = as.numeric(life_expectancy_long$TIME_PERIOD)
sapply(life_expectancy_long, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

GDP per capita

```
gdp_per_capita <- gdp_per_capita |>
  select(`Country Code`, as.character(2000:2023)) |>
  rename(REF_AREA = `Country Code`) # harmonizing the key variable for merging

# pivot to a long format
gdp_per_capita_long <- pivot_longer(gdp_per_capita, cols=as.character(2000:2023), names_to="
                                    values_to="OBS_VALUE")

sapply(gdp_per_capita_long, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character" "character"   "numeric"
```

```
# Changing the TIME_PERIOD variable to numeric
gdp_per_capita_long$TIME_PERIOD = as.numeric(gdp_per_capita_long$TIME_PERIOD)
sapply(gdp_per_capita_long, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

Unemployment

```
unemployment <- unemployment |>
  select(SERIES_CODE, as.character(2000:2023)) |>
  rename(REF_AREA = SERIES_CODE) # harmonizing the country code

# country code is in the format "ZAF.LUR.A". We only keep "ZAF", and drop the rest for all r
unemployment$REF_AREA <- sub("\\..*$", "", unemployment$REF_AREA)

# pivot to a long format
unemployment_long <- pivot_longer(unemployment, cols=as.character(2000:2023), names_to="TIME_
                                  values_to="OBS_VALUE")

sapply(unemployment_long, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character" "character"   "numeric"
```

```
# Changing the TIME_PERIOD variable to numeric
unemployment_long$TIME_PERIOD = as.numeric(unemployment_long$TIME_PERIOD)
sapply(unemployment_long, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

HDI

```
hdi <- hdi |>
  select(countryIsoCode, year, value) |>
  rename(REF_AREA = countryIsoCode, TIME_PERIOD = year, OBS_VALUE = value)

# Checking variable class
sapply(hdi, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character" "character" "character"
```

```
# TIME_PERIOD and OBS_VALUE need to be converted to a numeric format
hdi$TIME_PERIOD = as.numeric(hdi$TIME_PERIOD)
hdi$OBS_VALUE = as.numeric(hdi$OBS_VALUE)
sapply(hdi, class)
```

```
   REF_AREA TIME_PERIOD   OBS_VALUE
"character"   "numeric"   "numeric"
```

Merging the data sets

In order to have one unique database to work on, we need to merge all the datasets. To do so, we will use the variables `REF_AREA` and `TIME_PERIOD` together as our key. Our aim is to analyse the impact of the selected variables on `avoidable_mortality`, our dependent variable. Therefore, we use `left_join` to merge all data, relative to the data available in the `avoidable_mortality` dataset.

```
database <- avoidable_mortality |>
  left_join(gdp_per_capita_long, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_gdp")) |>
  left_join(hdi, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_hdi")) |>
  left_join(health_expenditure, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_health_ex
  left_join(hospital_beds, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_hosp_beds")) |>
  left_join(life_expectancy_long, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_life_ex
  left_join(social_expenditure, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_social_ex
  left_join(unemployment_long, by = c("REF_AREA", "TIME_PERIOD"), suffix = c("", "_unemp"))


# Giving a better name to the avoidable mortality column
database <- database  |>
  rename(OBS_VALUE_mortality = OBS_VALUE)

# short overview of our merged database
head(database, n=5) |>
  knitr::kable()
```

REF_AREA	TIME_PERIOD	OBS_VALUE_mortality	OBS_VALUE_gdp	OBS_VALUE_hdi	OBS_VALUE_health_exp	OBS_VALUE_hosp_beds	OBS_VALUE_life_exp	OBS_VALUE_social_expenditure	OBS_VALUE_unemp
AUS	2000	237	45859.52	0.897	7.589	4.04	79.23415	18.184	6.292
AUS	2001	227	46191.31	0.900	7.674	3.95	79.63415	17.488	6.775
AUS	2002	218	47486.28	0.902	7.872	3.93	79.93659	17.301	6.358
AUS	2003	207	48394.13	0.906	7.877	3.97	80.23902	17.484	5.942
AUS	2004	199	49902.02	0.910	8.083	4.00	80.49024	17.181	5.392

Summary statistics of the merged dataset

```
database |>
  summarise("Number of rows" = n(),
            "Number of columns" = ncol(database),
            "Percentage of NA" = paste0(round(sum(is.na(database)) / (n() * ncol(database)) *
            "Number of countries studied" = n_distinct(REF_AREA),
            "Time period" = paste0(min(TIME_PERIOD), "-", max(TIME_PERIOD))) |>
  as.data.frame() |>
  knitr::kable()
```

Number of rows	Number of columns	Percentage of NA	Number of countries studied	Time period
997	10	2.72%	46	2000-2023

We should note that, as a consequence of the `left_join` and the gaps in the `avoidable_mortality` dataset, there is missing data for several years for some countries : Turkey (2000 to 2008) and Greece (2000 to 2013). Although this represents a limitation of the dataset used, we decide to keep both countries, as we cannot afford to lose more data. We therefore keep all 46 countries, over the period 2000-2023.

Summarising the number of NA per variable

```
colSums(is.na(database)) |>
knitr::kable(caption = "Number of NA per column")
```

Table 11: Number of NA per column

	x
REF_AREA	0
TIME_PERIOD	0
OBS_VALUE_mortality	0
OBS_VALUE_gdp	0
OBS_VALUE_hdi	0
OBS_VALUE_health_exp	12
OBS_VALUE_hosp_beds	126
OBS_VALUE_life_exp	0
OBS_VALUE_social_expenditure	133
OBS_VALUE_unemp	0

Identifying missing values per country and per variable

```
database |>
  group_by(REF_AREA) |>
  summarise(across(everything(), ~ sum(is.na(.)), .names = "{col}")) |> # number of NA
  filter(rowSums(across(!REF_AREA)) > 0) |> # keep countries with at least 1 NA
  select(REF_AREA, where(~ any(. > 0))) |> # keep variables with at least 1 country with 1 NA
  knitr::kable(caption = "Missing values by country and variable")
```

Table 12: Missing values by country and variable

REF_AREA	OBS_VALUE_health_exp	OBS_VALUE_hosp_beds	OBS_VALUE_social_expenditure
ARG	0	18	23
AUS	0	8	1
BGR	0	0	1
BRA	0	1	22
COL	0	10	10
CRI	4	0	11
DNK	0	3	0
HRV	3	0	8
IRL	0	2	0
ISL	0	7	0
LUX	0	4	0
NOR	0	2	0
NZL	0	9	0
PER	2	23	12
POL	0	3	0
ROU	3	0	4
THA	0	20	20
ZAF	0	16	21

```
# removing countries with too many NA
database <- database |>
  filter(!REF_AREA %in% c("ZAF", "ARG", "PER", "THA", "BRA"))

dim(database)
```

```
[1] 888  10
```

```
# we save this cleaned version of the database so that it is directly reusable in phase 3
vroom_write(database, here("data", "cleaned_database.csv"), delim = ",")
```

There are 133 missing values in `OBS_VALUE_social_expenditure`. However, we cannot deduce that there are no social contributions in these countries, only that they have not been reported. As the missing values span over several years at a time, and often at the beginning of the time period, replacing them with a 0 or imputing them would result in a risk distorting our results by forcing an artificial trend. We therefore retain the missing values. However, we decide to remove 5 countries of our analysis : South Africa (ZAF), Argentina (ARG), Peru

(PER), Thailand (THA) and Brazil (BRA) as there are too many values missing to analyse them further during the modelling phase.

There are 12 missing values in `OBS_VALUE_health_exp`. They also correspond to missing values at the beginning of the time period. For instance, the health expenditure data is missing for Romania (ROU) from 2000 to 2002. Given both the small number of missing values, we also choose not to impute this variable.

Finally, 126 values are missing in `OBS_VALUE_hosp_beds` also in the beginning of the time periods available (e.g. 2007-2008 for Ireland). As with the previous variables, imputing these values would require imposing information that we do not have, and thus we retain them as missing.

Limitations of the merged database

The datasets used in this project present several limitations that must be considered when interpreting the results.

First, some variables contain an important proportion of missing values (e.g. social expenditure and hospital beds). The NA are primarily concentrated at the beginning of the time period studied (2000-2023) and likely reflect data reporting issues rather than an absence of social policies such as health contributions, usually marked as 0 in the rest of the dataset. This will naturally impact our results, especially since the missing data often comes from developing countries, thus biasing the results towards the trends observed in developed countries.

Second, the temporal coverage varies between countries. Indeed, there is no data for Turkey before 2009 or for Greece before 2014. This unevenness reduces the length of the time series for some countries and will influence the statistical analysis later on.

Presentation of the main variables

- `REF_AREA` : first 3 letters of the country whose data is shown.
- `TIME_PERIOD` : year for which the data was collected.
- `OBS_VALUE_mortality` : avoidable mortality, the number of "avoidable deaths" per 100000 inhabitants annually. It contains contains both "preventable mortality" and "treatable (or amenable) mortality". The first refers to deaths that can be avoided through effective public health and primary prevention interventions. The second refers to timely and effective health care interventions, including prevention and treatment.
- `OBS_VALUE_hdi` : human development index, per country, annually. It is between 0 and 1, 1 meaning the country in question is well developed.

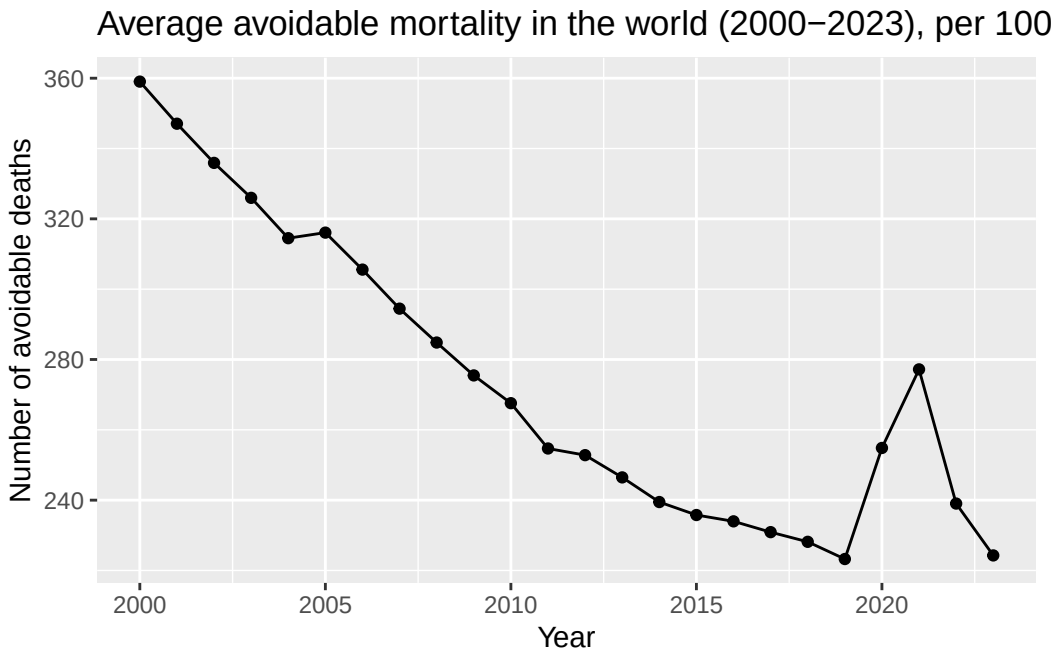
- `OBS_VALUE_gdp` : GDP per capita, per country, annually. It is measured in constant US dollars from 2015. Gross domestic product is the total income earned through the production of goods and services in an economic territory during an accounting period.
- `OBS_VALUE_hosp_beds` : number of hospital beds available per country, annually.
- `OBS_VALUE_life_exp` : life expectancy of a country's inhabitants, in years. It indicates the number of years a newborn baby would have to live if the general rules of mortality at the time of birth were to remain the same throughout its life
- `OBS_VALUE_social_expenditure` : social expenditure aggregate including : Old age, Survivors, Incapacity-related benefits, Health, Family, Active labor market programmes, Unemployment, Housing, and Other social policy areas. It is measured in percentage per GDP for all countries.
- `OBS_VALUE_unemp` : unemployment rate per country, annually.
- `OBS_VALUE_health_exp` : health expenditure observed annually. It is measured in percentage of the country's GDP. It was computed using financial flows related to the consumption of healthcare goods and services. Health expenditure can be filtered by "healthcare provider" such as Hospitals, Residential long-term care facilities, etc. The total of these categories was selected in the downloaded dataset.

Graphical representation of our main variable : avoidable mortality (OBS_VALUE_mortality)

Global avoidable mortality

First we can take a look at the number of avoidable deaths globally

```
database |>
  group_by(TIME_PERIOD) |>
  summarise(mean_mortality = mean(OBS_VALUE_mortality, na.rm = TRUE)) |>
  ggplot(aes(x = TIME_PERIOD, y = mean_mortality)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Average avoidable mortality in the world (2000-2023), per 100,000 inhabitants a
    y = "Number of avoidable deaths",
    x = "Year")
```

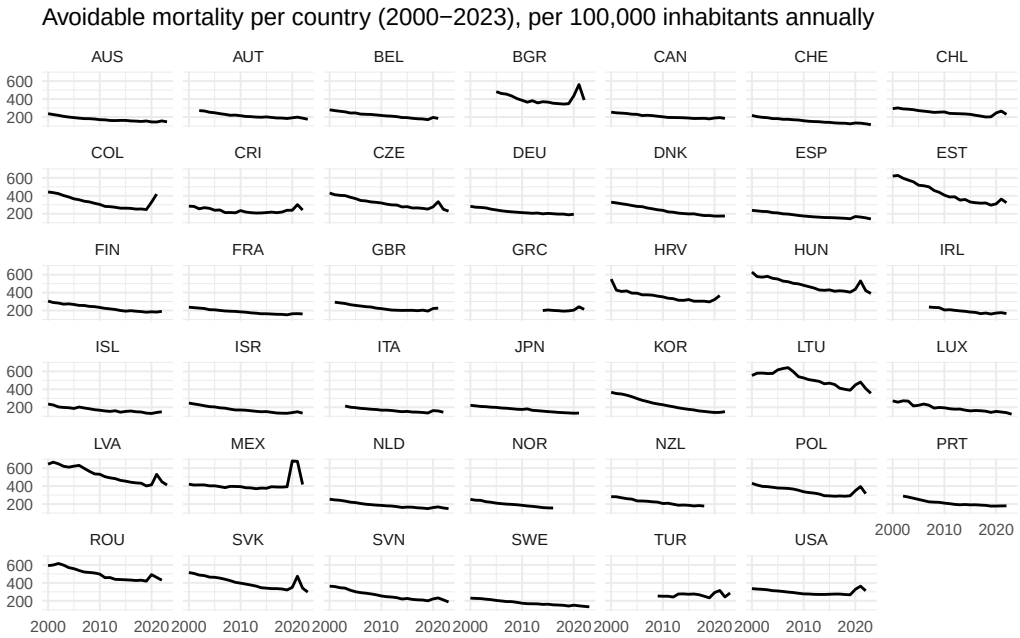


This graph reveals a downward trend in avoidable mortality from 2000 to 2019 cut by a sudden increase in 2020 and 2021 during Covid-19. However it seems that, afterward, the variable goes back to its previous trend.

Avoidable mortality per country

We also check the `avoidable_mortality` variable per country.

```
database |>
  ggplot(aes(x = TIME_PERIOD, y = OBS_VALUE_mortality)) +
  geom_line() +
  facet_wrap(~ REF_AREA) + # we do not use free_y so that we can compare rates between countr
  theme_minimal(base_size = 7.5) +
  scale_x_continuous(breaks = c(2000, 2010, 2020)) +
  labs(
    x = "",
    y = "",
    title = "Avoidable mortality per country (2000-2023), per 100,000 inhabitants annually")
```



Most countries share the same downward sloping trend in avoidable mortality. We can only observe a long term increase in the avoidable mortality rate in Thailand (THA) and South Africa (ZAF) before 2010. In addition, we can see the impact of the Covid-19 crisis clearly in some countries : Mexico (MEX), Peru (PER). Moreover, the high mortality rate in South Africa (1,000 deaths per 100,000 inhabitants around 2005) makes it more difficult to observe the trend and shocks in avoidable mortality in other countries, particularly the most developed ones.