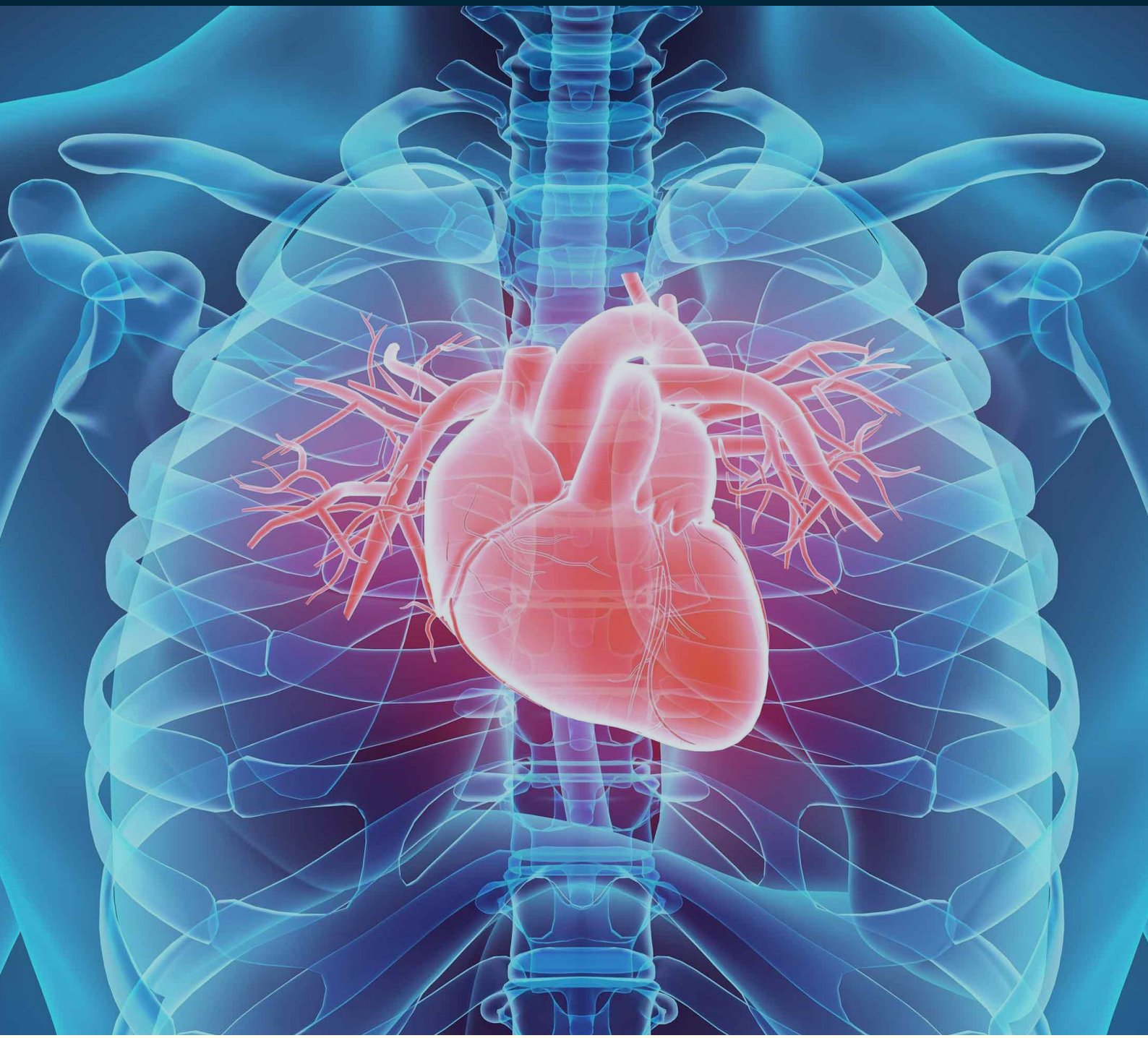
By Juliette Challot, Pauline Conti & Clémence Barsi

HOW CAN WE PREVENT CARDIOVASCULAR DISEASE ?



For Com-480 Data visualization

Table of contents

Sommaire

1.Path

A. SUBJECT

We wanted to explore the subject of Cardiovascular diseases as it is the first cause of death worldwide and identify what might impact its occurence. Hence we looked for datasets related to cardiovascular diseases to see under which angle we could study this subject.

We found two interesting datasets. The first dataset we found one contained 70'000 medical records about patients that were or weren't suffering from a cardiovascular disease. It also contained features such as whether the patient smoked, drank, had regular physical activity,... The second dataset contained all deaths registered by participating countries in the World Health Organization from 1988 to 2019 (even though 2019 records were incomplete as the world health organization needs time to process and verify all this data). It contains the year, country and cause of death, as well as the sex and the age of deceased people.

The datasets gave us the idea to focus on three main axes. The first one being the individual factors directly linked to the person, his habits like smoking, his age, his sex, his physical activity,... How frequent are these parameters found in the sick or in the healthy population ? We used the first dataset for this axis. The second axis is geographical and comes from the WHO dataset. How does living in a specific region of the world affect our risk of dying from a cardiovascular disease? This is a really interesting angle as we often fail to see the impact of our environment on our health. Finally we focus on the time factor : over the last 30 years, how has the rate of cardiovascular disease evolved ? We can answer this question using the last dataset.

B. MAKING THE SITE

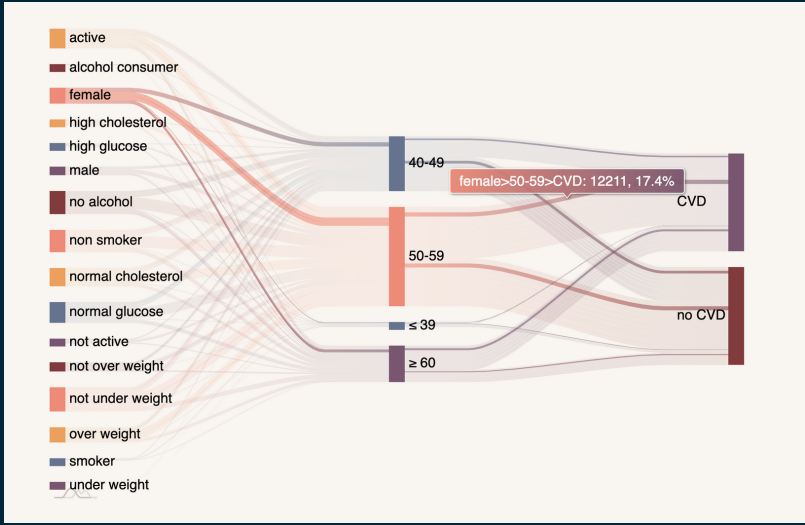
We started by implementing the skeleton of our website only missing the three visualizations. Then as we had three axes, we had a clear layout of how we could break down the work. We each had an axis and a visualization to focus on. This allowed us to work at our own pace.

Finally we merged all our visualizations into the site and added explanations of how to use our visualizations and interpretations. We also added explanations on our thought process so that the reader could understand the red line of our project and that it was smoothly understandable.

C. FINAL STRUCTURE

I. Individual factors

To identify the impact of each factor, we used a Sankey chart. On the left are all the aggravating factors and their opposite to see if the factor really has an impact on the risk of contracting a cardiovascular disease.

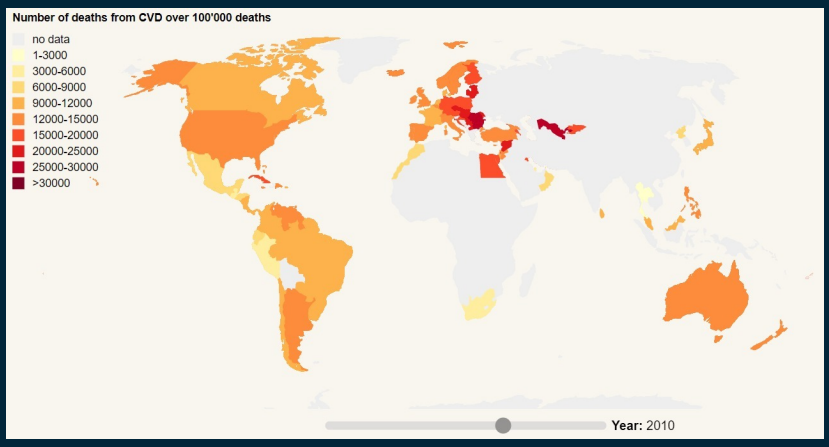


Each factor is linked in the middle to the range of ages to see which demographic is more touched by a specific aggravating factor. Then finally the age is linked on the right to the development or not of a cardiovascular disease. You can easily disable or reorganize nodes to better visualize the impact of a factor or combination of factors on the risk of having a cardiovascular disease. Finally, you can see the number and percentage over all records in the dataset of people represented by a specific path through the columns. In comparison to milestone 2, we added the opposites for all the factors to better visualise how frequent each factor actually is.

II. Geographical factor

To see the impact of the living area on the percentage of deaths caused by cardiovascular diseases, we chose to use a choropleth map. We wanted a map of the world where countries were displayed darker as their ratio of deaths from cardiovascular diseases over all deaths was getting greater.
This is to see if a country and more generally a continent is more affected by cardiovascular related deaths than others. For more precise information on a particular country; we added an information box showing the exact percentage of deaths from cardiovascular diseases in the country over which the mouse is placed. We also created a year slider to see if the tendency was repetitive over the years or if it had changed since the early 90's. This is important as stating that a particular place is more severely touched by cardiovascular related deaths based only on one year only would be a bit quick especially considering the lack of data for certain countries in the dataset.

Due to lack of data over the years 1988 to 1994 (only 1 country in each of these years) we decided to remove these years from the visualization altogether and focus on the rest of the data.

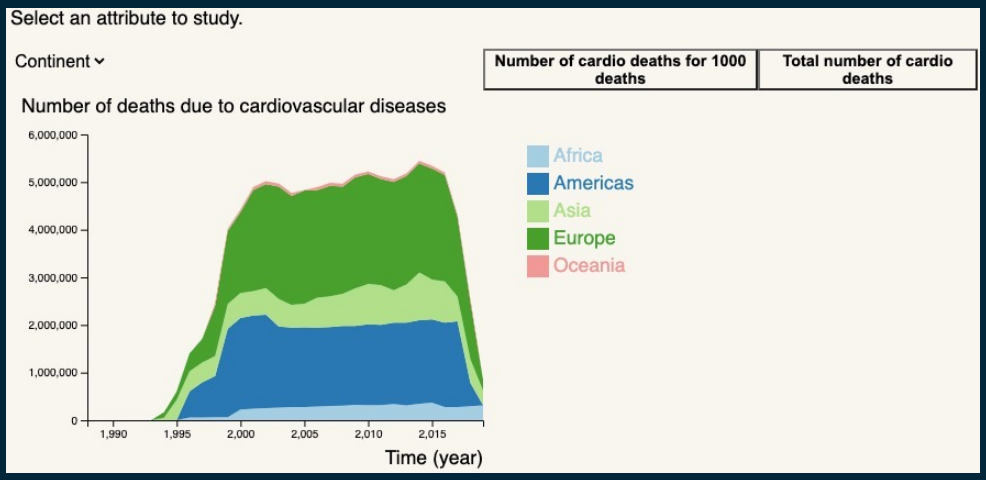


III. Time factor

We chose to use stacked area charts to see the impact of time to see if there was an overall rise in the number of cardiovasculare induced deaths during the last 30 years. We also wanted to study if this raise was particular to certain factors like the sex, the continent or the age. So we plotted the number of deaths in function of the year where the colors are factors that the user can choose between sex, age or continent. We also wanted to have a visualization where the bias of the population is erased, meaning that for example the population of Asia is much larger than the population of Oceania.

So we added a button that you can click on, where instead of plotting the total number of deaths, we plotted the number of cardiovascular deaths of this category for 1000 deaths of this category this year.

For example, for the year 2000 for the category female, we divided the number of cardiovascular related deaths of women by the total number of deaths of women this year.



2.Challenges

A. CONCEPTUAL SIDE

We are missing data from some regions of the world and some years in the second dataset. For example for the first six years we only have data from America so in our visualization we just didn't plot it or put it as no data.
As we have had disparities in numbers of records in the WHO dataset (like between Europe and Asia), we decided to normalize in some cases the number of cardio deaths by the total number of deaths to fight the predominance of some countries in our dataset.

Also, while preparing the data for the first visualization, we had to separate the factors in binary categories, in order to limit the number of nodes or the first columns. To do this, we had to combine multiple categories of weight of cholesterol levels for example. We combined them in "normal/not normal" categories using known standards (e.g. a BMI over 25 means that someone is overweight, while on below 18.5 means they are underweight).

While those choices were justified and explained in our preprocessing, but it was still challenging to choose adequate categories for our visualisation.

B. TECHNICAL SIDE

We faced multiple challenges on the technical side. The first one being how to format correctly our datasets to comply with the necessity of each visualization. Particularly for the WHO datasets where our initial columns were not very usable.

For instance age was stored in a different column depending on the age of death (e.g. D_at_0 contains the number of deaths in the first year of life, and D_at_15-19 sums all the deaths occurring between ages 15 and 19).

Another drawback we experienced was the "perfusion gauge" that we wanted to make for the first axis. We wanted to make a gauge that would fill proportionally to the number of aggravating factors we added. But we realized that it would require an unfeasible number of cases to compute the percentages of people with CVD for each combination of factors possible.

We also faced issues when trying to construct the first visualization with d3. The parsets library that we initially wanted to use required having disjoint sets in each category. This did not coincide with the goal of the visualization, as we would have been forced to add a column for each factor, or display only one factor at a time. We then found the d3 library for sankey charts, but it did not allow for the interactivity that we were looking for. We found the sankey Amcharts library, which allowed us to add multiple interactive features directly into the visualization (such as the disabling of nodes or the option to show the values when clicking on a link), which is why we finally opted for it. However, using the amcharts library also implied some challenges, as we were not acquainted with it.

For the second visualization the first challenge was due to the dataset : the code for countries was not the same as the one used by the world geomap and the name of the countries was a bit different as well (e.g. Iran was called "Iran (Islamic Republic of)" in the WHO dataset and simply "Iran" in the second one.

Thus we had to rename quite a few countries to be able to merge this dataset with the country codes used in our map (ISO 3166-1 alpha 3). Then a second challenge was adding a year slider and making the data update. As some countries have no data in certain years we had situations where the data from the previous year was still displayed just because the new year didn't contain any rows with the ID of this country.

To avoid this problem it was decided to set all countries' number of deaths from cardiovascular diseases to 0 and then update the rows present in the data from the selected year. As we noticed quite a few data were missing we dropped the idea of plotting the number of deaths from CVD over 10'000 per continent as we estimated that the few countries we had data over on each continent were not representative enough of the full continent.

3.Peer Assessment

The process of finding a subject and relevant datasets was a collaborative work from the team, we met several times to share our findings and ideas. For the design of the visualizations, we started by exchanging ideas and decided on specific communications goals for each visualization.

As said before we agreed on three different axes and each worked on one of them : Pauline focused on the aggravating factors (first visualization), Juliette on the geographical factor (second visualization) and Clémence on the time factor (third visualization). We kept in touch regularly to help each other with our visualizations and be sure that we were relevant to the answer we were trying to answer.

Pauline and Clémence both worked on the skeleton of the website, we each integrated our own visualization on the website. The process book was a collaboration from all group members.