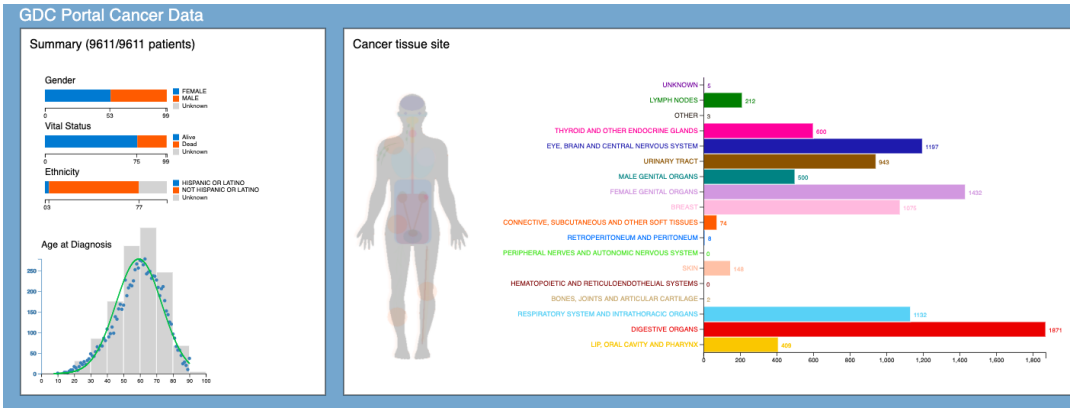
Project INF552

Capucine Leroux

December 2020



Contents

1 The Dataset 2
1.1 Presentation 2
1.2 Preprocessing 2
2 The visualisation 2
2.1 The summary widget 2
2.2 Description of the tumor, site and type widgets 4
2.3 Binding and correlation 5
2.4 Limits of the visualisation 5
3 Conclusion and insights 5

1 The Dataset

1.1 Presentation

The dataset I chose to visualize is the one available on the GDC Portal [1]. It covers anonymous data on more than 80k cancer patients, crossing several of the largest cancer open-source databases. Since there are many different types of information and data format, I decided to visualize the clinical data only from patients enrolled in the biggest program, the TCGA program [2].

The data I am visualising includes 9611 patients, hence 9611 files in the xml format. The information on each patient can vary, therefore I chose to represent information that was common to most patients, which is the age at diagnosis, the gender, the ethnicity, the vital status, and the tumor description (site of the tumor and its histological type).

1.2 Preprocessing

Each patient was represented by a separated xml file. I had to join all of the patients in one common json file. Since all the files did not follow the same key code, I had to standardize the key names, at least for the ones in common.

The main step of my preprocessing was to transform the tumor description as to get an understandable standardized name. All files had different tumor site and tumor type names, which were impossible to represent in a common design, but they also had a site code and a type code following the same nomenclature : ICD_O_3 [3]. I had to download the code correspondence and understand it to classify the tumors correctly, with a unified name correspondence across patients.

2 The visualisation

I separated the information into three different widgets :
- A summary of the patient-related information (gender, ethnicity and live status proportions, and a summary of the age at dagnosis)
- Tumor sites
- Tumor types

For each one of those categories, I represented the number of patients per feature.

2.1 The summary widget

For the binary categories (gender = male/female, vital status = alive/dead, ethnicity = hispanic or latino/not hispanic or latino), I chose a bar representation because it is an intuitive way to compare categorical data. Moreover, I did not represent it with two bars

but one global bar to get a better sense of the proportions (not only did I want to show which one was the biggest value, but I also wanted to emphasize each global percentage). I used 2 distinct color hues (orange and blue) so that the spectator does not have to struggle to differentiate both categories. The "unknown" proportion is always represented in gray to be less visible.

I added a label (on mouse over) with the exact number of patients and proportion for each bar. I wanted a simple design to get the most important information but I also wanted the visualization to be more precise if needed.

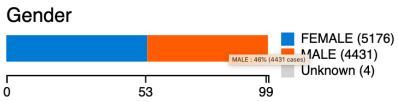


Figure 1: Bar with proportions of categorical data

For the age at diagnosis, there were too many categories to represent it all by a bar chart with one bar per year. But doing a grouped bar (one bar for one interval) tends to interpret too much. Indeed, the chart will be completely different if we represent ages 5 by 5 or 20 by 20. That is why I combined both solutions. I represented the number of patients for each age with circle points, and I added age categories with bars (10 by 10) behind to have a more general idea.
The bars were a lot higher than the dots so I had to scale the bars down. I did not want to have two y-axis so I only put an axis for the dots and preferred to let the details of the bars visible only on mouse over.
To visualise an age category of patients, the interval bars can also be clicked on to update the rest of the visualisation to represent only this category. I will go more into details about correlations and clicks at the end of the section.

Last , when I printed the chart, I noticed that the age distribution was closed to a gaussian distribution and I wanted to emphasize that by adding the gaussian function line calculated with the mean and the deviation of the age distribution.

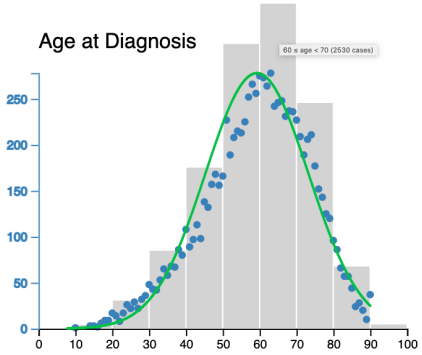


Figure 2: Age at diagnosis chart

2.2 Description of the tumor, site and type widgets

For the descriptions of the tumors, there were many different categories. I represented 3 level of details for both site and type. The site code looks like this "C40.8". The first part "C40" corresponds to a body region and the second part ".8" to an organ or a smaller part within the region. I chose to represent intervals of regions "C40-C44" in a bar chart to get a global look at the proportions, and then the two next levels "C40" and "C40.8" in a tree. The nodes of the tree are initially all closed, so that the visualisation can show a personalized level of details if needed. Besides, the whole opened tree would be too big to be shown at first. I used the same representation for the tumor type code since the code can be similarly decomposed.

As for the colors, because the sites/types were categorical, it would have been inappropriate to choose one color with several saturations, so I chose one color hue per interval, choosing them to be as distinct as possible.

Again, the bars are detailed on mouse over. I printed the names and not the code of the tumor description to have something easily understandable.

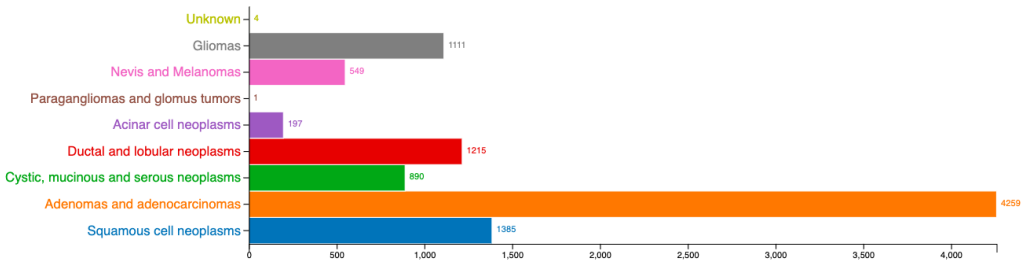


Figure 3: Bar Chart for tumor type

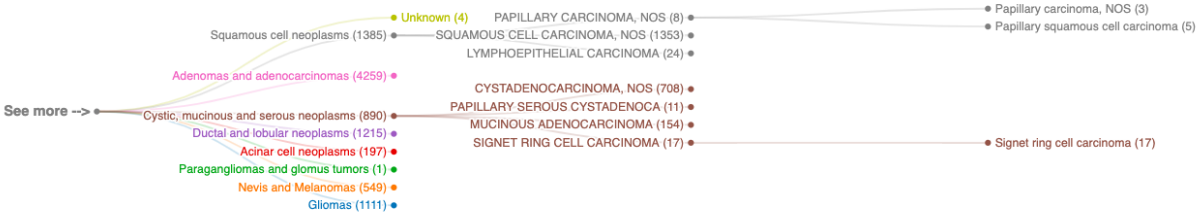


Figure 4: Tree for tumor type

For the tumor site, one important feature that is not represented by the tree or the chart is the physical location of the site in the body. That is why I added a body icon with very simplified shapes showing the region of the body of each category. Each time a site category is selected on the chart, the corresponding region shows on the body.

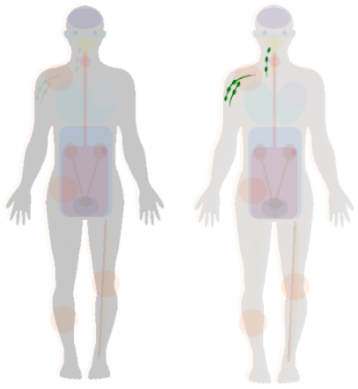


Figure 5: Body icon

2.3 Binding and correlation

I wanted to bind all the charts and trees. Every time a bar is clicked on, all the visualisation around is updated to show only the selected category. The selection event is shown to the user by a color fading of the selected bar.

2.4 Limits of the visualisation

One limit to this visualisation is that it cannot show one patient only, but groups of patients with common features.
Moreover, a problem occurring with the bar charts is that when a category is very small compared with the others, it is not clearly visible on the chart, therefore cannot be clicked on.

3 Conclusion and insights

Thanks to this visualisation, we can have a good insight at the general content of the dataset, which cannot be done by reading each patient file. We get a good sense of the proportions of each feature categories.

Besides, the binding obtained by clicking on individual categories enables us to observe many correlations within the data set. For example, we can see that far more patients have died from a brain/eye cancer than from a cancer of male genital organs. Young adults (between age 20 and 30) don't have lung cancer. Or Gliomas are only touching brain. To conclude, the knowledge that can be extracted from the visualisation is huge, and with few features compared to what the data set originally contained.

Bibliography

[1] *Genomic Data Commons Portal*
https://portal.gdc.cancer.gov/
National Cancer Institute

[2] *The Cancer Genome Atlas Program*
https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
National Cancer Institute

[3] *International Classification of Diseases-Oncology-3*
https://cancercenter.ai/icd_0_3_pathology/