APPLIED STATISTICS GROUP

SUBJECT N°22

Text mining and care pathway: what are the causes of mortality in heart failure patients?

Summary note

Keywords: Heart failure / GHM / Care pathway / Clustering / Pattern Mining / Cox Model / Survival Random Forest

Students :
Tristan AMADEI
Tristan KIRSCHER
Antoine KLEIN

Coordinator (CREST) :
Dr. Roxana FERNANDEZ

Support (AP-HP) :
Dr. Anne-Isabelle TROPEANO
Juliette MURRIS (PhD C)

July 5, 2023

1 Context and objective of the project

Heart failure is a disease of the cardiovascular system characterized by an inability of the heart to pump enough blood to meet the body's oxygen and nutrient needs. In France, more than 1.5 million people suffer from this disease, mainly people over 60 years old.

The objective of the study is to clarify the causes of mortality in heart failure patients, who are increasingly older. Knowing the main causes of mortality in these patients and their most frequent care pathways will have a major public health impact.

To answer our question, we first characterize the care pathways of patients through the study of sequential patterns: using GHM codes (*Groupes Homogènes de Malades*) defining hospitalizations, it is possible to find similarities in the care pathways, associated with a diagnosis.

Once these pathways are identified, a survival analysis will predict the survival trajectory after first hospitalization.

2 Description of the data

The data is extracted from the EGB (*Echantillon Généraliste des Bénéficiaires*, random sample representative of 1/97th of the population with a follow-up of at least 2 years, from the French health insurance databases). At the end of our data cleaning and selection phase, we were left with 10,051 patients and 85,594 hospitalizations.

As introduced in section 1, we will work with GHM codes which characterize the different hospitalizations patients undergo. The syntax of a GHM code is very insightful regarding one's hospitalization. A GHM code can be divided as follows:

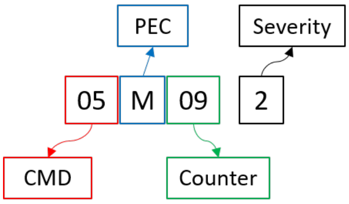

Figure 1: Dividing the GHM code into meaningful parts

3 Patients clustering

In order to obtain results more easily interpretable, and that would more easily respect the hypotheses of models that will be used later on, we decided to split the pool of patients into clusters. We wanted these clusters to closely fit the patients' hospital journeys, and thus we chose to use unsupervised learning to establish those clusters, instead of using information on the patients in order to create fixed clusters.

3.1 Clustering algorithm

The idea was to represent the patients by their care pathways. Based on the information lying in the GHM codes, we created our own distance metric to assess the distance between different care pathways. The goal behind clustering the data is to consider, within each cluster, patients that have followed similar care pathways, and on the contrary, consider patients from different clusters to have followed quite different ones. In this context, we want the frequency of GHM codes within a cluster

to be greater than the frequency of the same GHM code within the whole dataset.
In order to establish those clusters, we used the K-Medoids algorithm.

3.2 Interpretation of clusters

We were able to find optimal parameters to run our clustering algorithm, that highlighted the fact that 5 clusters were needed. Let's visualize those established clusters of patients.

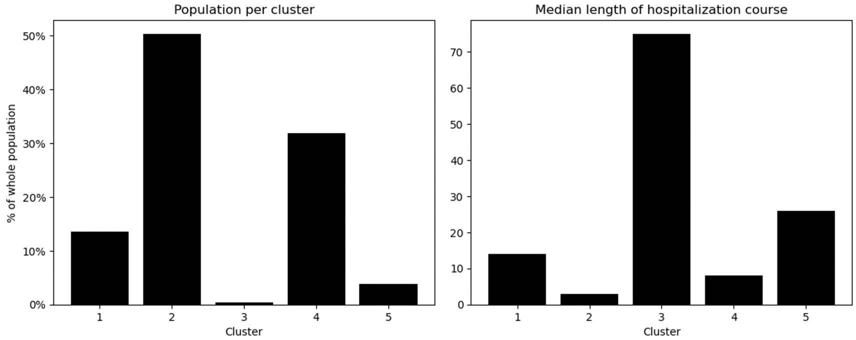

Figure 2: Population and median length of hospitalization course per cluster

What is interesting here is the cluster 3. There are only 35 patients in this cluster, and they seem to have been through considerably more hospitalizations than the other patients of the database. Thus, it is actually good news - and a testatement of sort that our clustering was successful - that these outliers were set aside in a cluster of their own.
If we take a look at other features, there does not seem to be any significant difference with the initial dataset; which shows that clusters, indeed, did get calculated from the hopitalization courses, and those courses only.

4 Health care pathways analysis

The aim of this section is to extract frequent health care pathway patterns, risky, or even potentially lethal trajectories. Their identification will improve future predictions, as well as trying to better understand the causes of death in heart failure patients. We will notably use the clusters defined previously in section 3 to see if patterns emerge according to patient groups.

4.1 Sequential Pattern Mining

Sequential Pattern Mining (SPM) is a data mining technique used to discover frequently occurring sequential patterns in a sequence database.

In order to analyze patients health care pathways, we built such a sequential database. In this dataset, which is divided into relative timestamps corresponding to the occurrence of hospitalization and its GHM type, time is perceived as a discontinuous variable. The example from Table 1 is taken from this dataset and corresponds to the health care pathway of patient P6. The corresponding sequence database is: ['02C05', '05M09', '04M13', '05C22', '23M10', '04M05', '04M24'].

Frequent patterns are being extracted from health care pathways using a *sequential pattern mining* algorithm. To do so, we used the *PrefixSpan* algorithm and computed frequencies for each of the patient clusters identified in Section 3.

We thus obtain a dataset containing, for each patients clusters, the most frequent patterns for different temporal stamp lengths. We notably found that hospitalization for heart failure ('05M09') is the last hospitalization before death in **33%** of care paths.

GHM	02C05	05M09	04M13	05C22	23M10	04M05	04M24
Hospitalization n°	0	1	2	3	4	5	6
Timestep	t_0	t_1	t_2	t_3	t_4	t_5	t_6

Table 1: Explanation of the sequence database for the patient P6 with corresponding timesteps and hospitalization numbers

4.2 Heart failure patients healthcare trajectory visualization

In this section, the aim is to highlight the trajectories of heart failure patients to visualize patient flows through the different care pathways. First, we did a visualization of frequent spatio-temporal patterns, the purpose of which is to give us an overview of the similarity of care pathways for heart failure patients.
Setting aside '05M09' heart failure and deaths events, the following GHMs can thus be highlighted:

- **04M05**: Pneumonias and common pleurisy.
- **04M13**: Pulmonary edema and respiratory distress.
- **05K10**: Diagnostic procedures using vascular route
- **23M20**: Other symptoms and reasons for seeking medical care under CMD 23
- **16M11**: Other disorders of erythrocyte lineage
- **05M08**: Arrhythmias and cardiac conduction disorders
- **04M20**: Chronic bronchopneumopathies with superinfection

Then, we created Sankey diagrams of patient flows. The interest of this visualization is that we can easily see the patterns that are upstream of the 'Décès' (death) event, which can give some insights to identify the causes of mortality of heart failure patients.

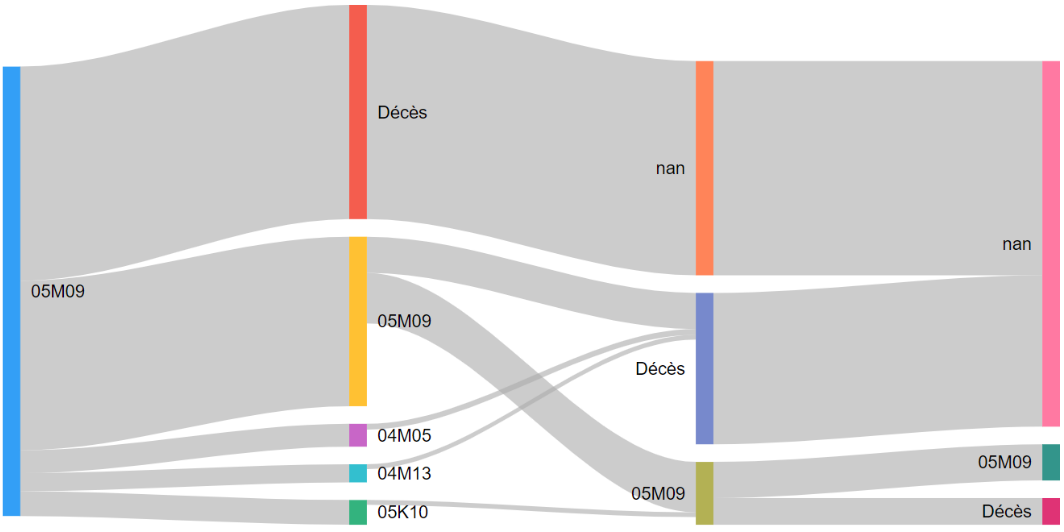

Figure 3: Sankey Diagram of frequent patient flows

By visualizing the frequent trajectories for the entire heart failure patient population, we can identify several frequent trajectories leading to death: a succession of hospitalizations for heart failure ('05M09'), hospitalization for *pleurisy* ('04M05') or for *Pulmonary edema and respiratory distress* ('04M13') after the first heart failure. The same is done at a cluster level in the full report.

5 Survival analysis

5.1 Goal of the survival analysis

As we now have a clustering for the patients and insights of their health care pathway, we implement a survival analysis to obtain prediction.

Definition 1 (Survival Model)**.** A model is called survival if it contains censored data.

The survival time is then **right-censored** by the study window. The objective is to estimate, for a cluster of patients, the **probability of survival over time**. The model must be **interpretable** with statistical guarantees since its predictions will have clinical implications. The exact model and methodology can be found in our final report. The latter allows us to obtain p-values and confidence intervals. Wanting to see the marginal impact of having or not a cardiac shock in one's care pathway, we decided to **stratify** our study by the binary variable CHOC.

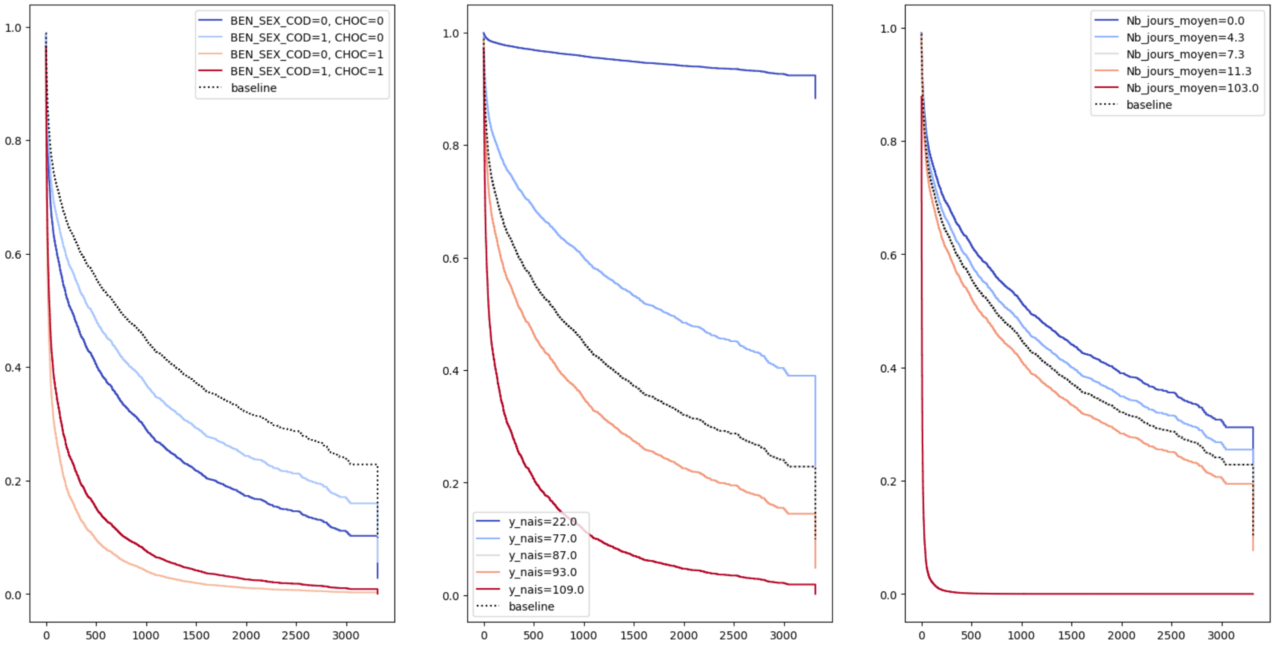

Figure 4: Survival Trajectory for **Cluster 2** stratified by :
Gender and Cardiac Shock | Age | Duration stayed at the hospital

This is the predicted survival trajectory given for the cluster 2. Others plot are located in our final report.

From these results, several observations emerge :

- Old age is an **aggravating factor** in mortality for all clusters. What is interesting is to see that this impact is **not of the same order** between clusters. In particular, the elderly in cluster 2 have a much more pessimistic trajectory than those in cluster 1.
- Time spent in hospital is again a sign of a pessimistic trajectory. However, we should not see this as causal: it is those with the most serious pathologies who stay the longest (**selection bias**). Nevertheless, for cluster 3, we notice that the people who stayed for a long time have a better survival trajectory: is this because the care provided allows them to remedy their pathology?

In terms of predictive power, we obtain for a test_set (20% of the dataset):

Metrics on the test_set		
Clusters	Concordance Index	Wish
Cluster **1**	0.612	Near 1
Cluster **2**	0.64	Near 1
Cluster **3**	0.887	Near 1
Cluster **4**	0.621	Near 1
Cluster **5**	0.543	Near 1

We observe **good predictive capabilities** with a mean concordance index of **0.66**, especially since our models are of the same order of power regardless of the cluster.

5.2 Survival Random Forest

We also wanted to compare the predictions of the chosen model with **non-parametric approaches**: the Survival Random Forest and the Survival Gradient Boosting. They are an extension of the well-known Random Forest and Gradient Boosting for censored data. The plots are given in the final report.

We obtain **similar trajectories** with the most pessimistic faith for cluster 2 and cluster 4. In term of predictive power, we recover the performance of Cox model : approximately **0.68**.

The same performance goes for Survival Gradient Boosting.

Conclusion of this part : The relaxation of the parametric model **loses explicability** without bringing any predictive power. We will then keep a parametric Cox approach.

6 Conclusion

This study is part of the search for causes of mortality in heart failure patients. We had at our disposal a medical cohort of 10,000 patients followed over several years with their successive GHM. From this dataset, we elaborated a metric adapted to our situation to carry out a relevant clustering. We then retained 5 clusters.

The identification of frequent hospitalization patterns and the visualization of the care pathways of heart failure patients allowed us to highlight frequent hospitalization causes and trajectories leading to patient death.

Finally, a survival analysis was conducted to establish predictions of death. The selected cohort did not meet all the conditions for a parametric analysis (too great a diversity of survival trajectories). However, it provides information on the dynamics according to age, sex, the presence of a cardiac shock, and the average length of stay in hospital. A nonparametric analysis by Survival Random Forest was conducted. The latter did not perform any better than the parametric one, while leading to a loss of interpretability. We will therefore retain the use of Cox models stratified on each cluster.

Thus, this study, which is more methodological than operational, paves the way to the use of GHM Big Data to better understand heart failure.