BINF-F401 Project

Analysis of the morphological clusters from heart and there associations with clinical and transcriptomic data.

Draguet Simon Godin Maximilien Guyot Léopold

June 4, 2024

1 Introduction

R (**?**)

2 Exploration of clinical variables

Description des variables .....

2.1 Distribution of the variables

Associated script : .R

By examining the clinical variables through histograms, we can visualize their distribution. Most of the continuous variables do not seem to follow a normal distribution. For instance, the variable AGE appears to be asymmetric, exhibiting a right skew (fig X). To now if the continuous variable were normally distributed, we used a Shapiro-Wilk test. The results showed that AGE, HGHT, BMI, TRISCHD all had p-value under 5%, leading us to reject the null hypothesis, which say that the sample is drawn form a normally distributed population. However, for WGHT, it was higher than 5 , so we didn't reject the null hypothesis (table X) .

To do further analysis, we need them to be normally distributed. We tried to apply various transformations like log, square root, square. Only the square transformation successfully normalized the AGE sample. Because we needed it to work on all the continuous variable, we finally chose to use the rank-based inverse normal transformation (INT). that first convert the variable into ranks, then map it to a normal distribution.

$Y_i^t = \Phi^{-1}(r - C/N - 2C + 1)$

where r_i is the ordinary rank of the $_ith$ case among the N observations and Φ^{-1} denotes the standard normal quantile (or probit) function.For the value of C we use C=3/8 (**?**).

If we look at the discrete variable, we can see that we don't have a balanced distribution. For example, for the variable AGE, there are more than twice as many males as females (fig x). For DTHHRDY, we can observe that most of the deaths occurred in ventilator cases (fig x). When conducting analysis, it's essential to keep these observations in mind as they can significantly influence the interpretation.

2.2 Correlations between the clinical variables

Principal Component Analysis (PCA) is a method of determining individual profiles and linear relationships between variables, based on correlation coefficients. The various graphs produced by a PCA, notably the correlation circle, are therefore a good way of illustrating the links between different quantitative variables, and getting a general idea of these links. Nevertheless, the data we are considering here are also partly qualitative. It is therefore preferable to turn to a Factor Analysis of Mixed Data, i.e. an analysis that applies a PCA to quantitative variables and a Multiple Correspondence Analysis to qualitative variables. The various dimensions defined by this method can be used to characterize all the variables. The FAMD() function in the FactoMineR package was used, with the argu-

ment allowing the creation of a series of graphs.

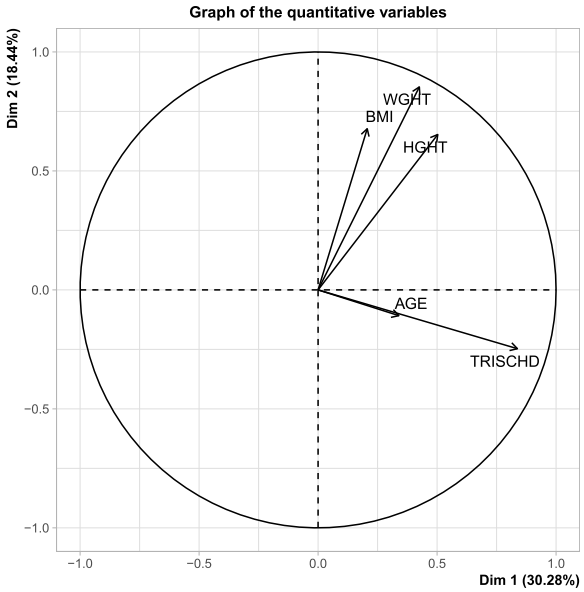

Figure 1: Correlation circle of the continuous variables

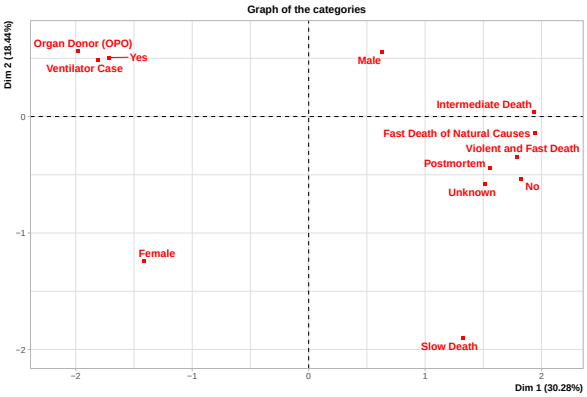

Figure 2: Level maps of all the different categories in the qualitative variables

The correlation circle allows us to represent quantitative variables according to the relationships they might have with each other, and the degree of explanation of these variables provided by the chosen dimensions. In this case, the two dimensions represented together explain only 48.72% of the data, suggesting a certain complexity. This graph shows some correlation between the variables height (HGHT), weight (WGHT) and body mass index (BMI), which is not surprising given that taller people tend to be heavier, and also given that BMI is a function of height and weight. There also appears to be a correlation between age (AGE) and ischemic time (TRISCHD), i.e. the time between an individual's death and organ removal. This last observation is more difficult to explain intuitively. We can, however, point out that the variable AGE is weakly explained by the first two dimensions chosen, since its vector is close to the origin. This indicates that this possible correlation may not be significant, since by considering other dimensions, the vectors could be significantly far apart.

In contrast to the correlation circle, the level map represents the different categories or levels of categorical variables. We can already see that gender is close to different clusters. It would seem that there is a correlation between gender and type of death in these data, given that women seem to be more prone to ventilatory problems than men, which would explain the presence of a respirator before the person's death (Yes). It also appears that women are more likely than men to be organ donors. This graph therefore shows a certain correlation between gender and organ donation, type of death, but also between type of death and the presence or absence of a respiratory system prior to death.

Although these graphs illustrate the presence of possible correlations, they do not give any indication of their intensity, as they are not quantified. This quantification of correlation is particularly complex, as there is no correlation coefficient that can be applied to either quantitative or qualitative variables, or between these two types of variable. It was therefore decided to use different coefficients depending on the type of comparison, although this choice does not allow all correlations to be compared with each other. Correlations between quantitative variables were established using the cor() function with Spearman's method. This correlation was chosen because it does not require normally distributed variables, which is not the case for most of the variables considered. This coefficient is the only one to consider negative correlations. Cramer's V was used to determine correlations between categorical variables, a coefficient based on the Chi-square statistic, and which is non-parametric. This correlation was calculated using the cramerV() function in the rcompanion package, with a bias correction given that this test tends to overestimate the

relationships between categorical variables. Logistic regression is used to determine the correlation between a quantitative variable and a binomial categorical variable, in this case gender and cohort. Finally, Cramer's V is applied to the H statistic of the Kruskall-Willis test between categorical and quantitative variables, as this statistic is the Kruskall-Willis chi-square. This application simply involves calculating the square root of the H statistic divided by the number of observations, the H statistic coming from the kruskall.willis() function.

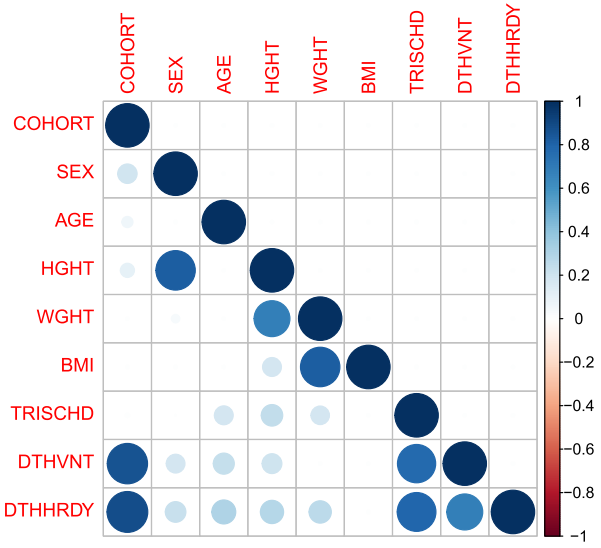

Figure 3: Correlation matrix of all the variables

Looking at all the results together, we can see that there is a significant correlation between gender and height, which can simply be explained by the fact that men are generally taller than women. The correlation between height and weight, as well as between weight and BMI, can be explained as described above. Nevertheless, the absence of any significant correlation between height and BMI is rather odd given that the BMI calculation takes into account both height and weight, and that the correlation circle has positioned the corresponding vectors relatively close to each other. There is also a correlation between cohort and both type of death (DTHHRDY) and use of artificial ventilation before death (DTHVNT). There is a certain logic behind these correlations: the cohort variable divides people into organ donors and postmortem donors, i.e. victims of brain or cardiac death, thus depending on the type of death, but we can also consider that people with brain death are generally assisted with artificial ventilation before death. There also seems to be a correlation between the time between the individual's death and the removal of his or her organs and the type of death, but also the presence of a ventilatory apparatus. Finally, DTHVNT and DTHHRDY are also correlated.

2.3 Impact of technical variables on clinical variables

In this section, we investigate the influence of technical variables on clinical data and describe the methodology employed to isolate their effects. To decouple the impact of technical variables from the data of interest, linear regression analyses were conducted. Prior to regression, the data underwent transformation using inverse data normalization for numerical variables (cf. section 1.1).

From the preceding section where correlations between variables were examined, certain expectations were established. Age was anticipated to be associated with TRISCHD, DTHHRDY, and COHORT, while height and weight were expected to be linked to the same variables. Additionally, sex appeared correlated with DTHHRDY and COHORT. Surprisingly, BMI seemed uncorrelated with any technical variables.

For the technical variable DTHHRDY, a decision was made to group certain categories into three new categories: slow, intermediate, and fast. This choice was made since more observations by categories would enhance statistical power.

Upon conducting linear regression (generalized linear regression in the case of the categorical variable sex), it was observed that age was dependent on cohort, sex was dependent on DTHHRDY, height was dependent on DTHHRDY, weight was dependent on DTHHRDY, and, surprisingly, BMI was dependent on both DTHHRDY and COHORT.

Subsequently, linear regression (or generalize d linear regression) was rerun for each variable of interest, utilizing only the significant technical variables from the overall models. The residuals of these models were stored for further analysis. These residuals represent the data of interest with the unwanted effects

of the technical variables removed, providing a clearer understanding of the clinical data.

3 Association between clinical variables and morphological data

3.1 Method

To investigate the relationship between clinical and morphological data, we treated this problem as a differential expression analysis. In this case, the clinical data represents the samples metadata and the morphological data the features. This approach allow us to take advantage of the numerous tools and methods developed for differential expression analysis. Here we used the DEseq2 package (**?**). The strength of DESeq2 is that it does correct the morphological count matrix to take into account potential bias. These bias are for instance the difference in the total count between samples and the difference in library composition that can lead to bad normalisation if not taken into account. To execute this normalisation, DESeq2 will compute geometric mean for each morphological cluster. Then, it will use these values and create a new matrix that consists of the counts divided by their associated geometric mean. After that it will use this new matrix to build scaling factors for each sample, and finally applying this scaling factor to the original counts.

For the question 2.1 and 2.2 we used the each clinical variable and each technical variable as the contrast for the design of the differential analysis. For the question 2.3 we used a contrast composed of each clinical variable plus the technical variables that were significantly associated with the clinical variable in the previous section. By doing that we ensure that we adjust the analysis to take into account the confonding effect of the technical variables. Note that for numerical variables, we used the rank-based inverse normal transformated values obtained in the Question 1.1.

We also decided to make extra analysis by converting the numeric clinical data to binary data. By doing this we assume to obtain more significant results. This can be explained by the fact that the DESeq2 method will in the case of continous variable report the log2 fold change per unit of change of that variable. Therefore for variables with an high range of value, the fold change will be very low. By using binary variable we can increase the fold change and therefore the significance of the results. The binary conversion was done by taking the median of the variable as the threshold to separate the two categories.

Some other changes have been made for this analysis, they are listed in the following:

- The Hardy scale variable was converted to an 3 categories variable, by grouping the intermediate and slow categories together (cf. Q1)
- Samples with DTHVNT unknown were discarded
- A pseudo-count of 1 was added to the count matrix to avoid log transformation of 0 values

As for the interpretation of the results, we decided to fix the threshold values as follow, alpha at 0.05 and log2 fold change at 1.

3.2 Clinical data vs. morphological data

3.3 Technical data vs. morphological data

3.4 Adjusted clinical data vs. morphological data

4 Association between morphological data and gene expression

4.1 ...

4.2 Gene Set Enrichment Analysis

GSEA (Gene Set Enrichment Analysis) is a preranked method used to analyze gene expression data, allowing us to determine if genes within a given gene set exhibit non-random behavior. However, this method can be slow because the analytical form of the null distribution for the Enrichment Score (ES) statistic is not known, necessitating the calculation of

an empirical null distribution. The process involves calculating an ES value for each pathway, then generating random gene sets of the same size and calculating an ES for each of these random sets. .The P-value is estimated as the number of gene sets that have an ES value equals or more extreme and divided by the total number of generated random gene sets. But to have a good statistical power it needs a large number of gene set samples.

Here, we are going to use FGSEA (Fast Gene Set Enrichment Analysis), which efficiently estimates the GSEA P-value for a collection of pathways. It uses FGSEA-simple and FGSEA-multilevel on the pathways that have a low P-value.

FGSEA-simple calculates the estimated P-values efficiently and simultaneously for the whole collection of gene sets but with limited accuracy. It is based on the idea that random gene set samples can be shared between different input pathways. It estimates the P-value of a pathway with the largest gene set size "M" as a proportion of a sample of random gene sets of size KM, having the same or more extreme ES value as the pathway. For other pathways "j", it constructs a set of independent samples of size Kj (Kj $<$ KM) by considering the prefixes in each sample of the larger random gene set and then estimates the P-value using the set of independent samples. It also uses the idea that using the larger random gene sets, the ES values for all the prefixes can be calculated efficiently using a square root heuristic. It uses a variant of the enrichment curve where the enrichment curve is constructed starting with the gene that is the most upregulated to the most downregulated, with the curve going right if the gene is not present in the pathway and up if it is present. Then the enrichment score can be determined using the point that is furthest from the diagonal.

FGSEA-multilevel can estimate low P-values accurately but for individual gene sets. It is based on a multilevel split Monte Carlo scheme. It calculates the probability of a random gene set of size K to have an enrichment score no less than γ, where γ is an ES value > 0 given as input. Successively, the method finds ES levels (l_i) where the probability of ES being greater than the ES levels (l_i) is equal to 2^{-i}. Finally, the method stops when $l_i$becomes greater than γ and when the P-value approaches 2^{-i}.

fgsea take as input a ranked list of or genes differentially expressed, we ranked them by significance and by the direction of the change using the formula:

$RANK = sign(FC) * -log_{10}(p - value)$

Where FC is the log2(fold change).We used this formula instead of the log2fold to take into account genes with a large fold change but non-significant. At the top, we have the most significant and up-regulated genes, and at the bottom, the most significant down-regulated genes. Then it needs a list of gene sets. We used "c2.cp.reactome.v7.5.1.symbols.gmt". This list of gene sets has been filtered to only contain genes that are present in the ranked list of differentially expressed genes and with at least 5 genes present. Finaly, we run fgsea and exclude pathways that have gene set with a length under 10 and pathways with a gene set of over 1000 genes

Cluster	AGE		SEX	HGHT		WGHT		BMI		COHORT	TRISCHD		DTHVNT	DTHHRDY	
	Normalized	Categorical (56+ vs -56)	Female vs Male	Normalized	Categorical (71+ vs -71)	Normalized	Categorical (181+ vs -181)	Normalized	Categorical (28+ vs -28)	PostMortem vs OrganDonor	Normalized	Categorical (50+ vs -50)	Yes vs No	Intermediate vs Slow	Fast vs Slow
0		1,22	-1,26							1,98		1,78	-1,49	2,10	2,00
1															
2		1,87	-1,99				1,15		1,70	2,46		2,36	-2,38	1,71	2,85
3										1,24					
4															
5										1,55		1,19	-1,41		
6			1,70	-1,02	-2,32					-5,33		-5,91	6,27	-6,34	-5,90
7										-1,10					
8			1,03							-2,31		-2,02	2,17	-2,25	-2,02
9										1,17					
10		-1,53	1,07		-2,69		-2,60			-6,72	-1,12	-6,90	7,35	-6,41	-7,12
11															
12										-1,32			1,03		
13															
14										-1,84		-1,72	1,78	-1,62	-1,58
15										1,25		1,04	-1,03		
16										1,48		1,19	-1,07	1,41	
17										-1,42					-1,20
18						4,40		3,12			-5,25				
19															
20			1,02							-2,38		-2,07	2,44	-2,43	-2,26
21			-1,09							1,83		1,60	-1,69	1,68	1,73
22										1,47		1,22	-1,26	1,23	1,02
23			1,28		-1,56					-4,53		-4,22	5,19	-5,27	-4,91
24										2,19		1,93	-1,61	2,22	2,09
25			1,59		-1,42					-5,75		-5,21	5,58	-6,35	-5,25
26										1,72		1,17	-1,46	1,28	1,57
27					-1,14					-3,09		-2,62	2,92	-3,27	-2,85
28															
29															
30															
31		-1,57			-1,47		-1,72		-1,10	-4,27	-1,38	-4,64	5,01	-5,77	-4,42

Table 1: Significant fold changes for each variable considered in this project, we also considered to categorize the continous data as a possible way to have more significant results.