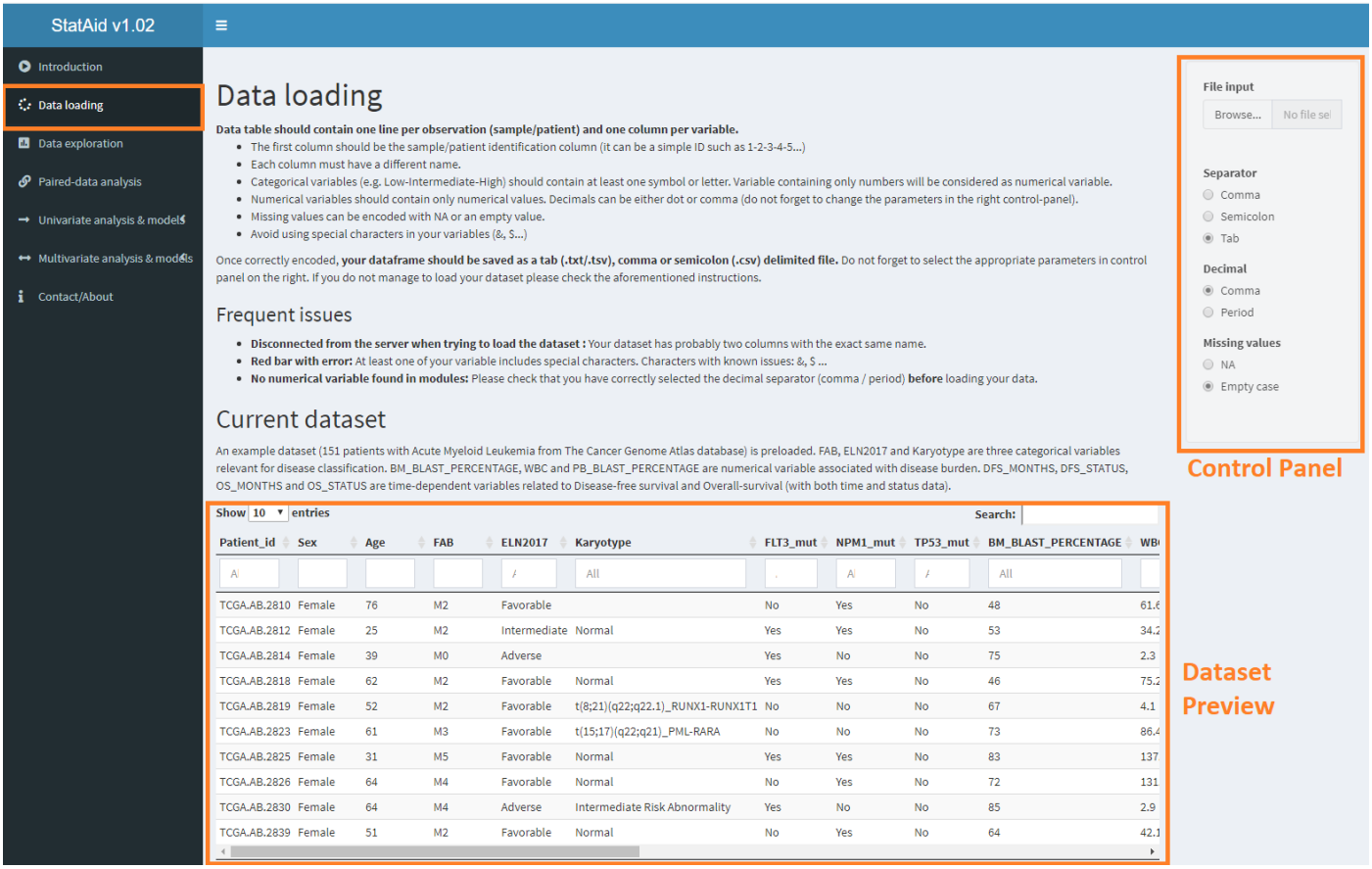
1. DATA LOADING

In this module, your dataset can be loaded and checked in a small table. StatAid expects dataset with one patient (or sample/observation) per row, and one variable per column. The first row corresponds to variables names. The first column of your dataframe will be used as a sample identifier so be sure to have an identification variable included here.

Important considerations:

- Each column must have a different name.
- Categorical variables (e.g. Low-Intermediate-High) should contain at least one symbol or letter. Variable containing only numbers will be considered as numerical variable.
- Numerical variables should contain only numerical values. Decimals can be else dot or comma (do not forget to change the parameters in the right control-panel).
- Missing values can be encoded with NA or an empty value.
- Avoid using special characters in your variables (&, $...)

Once correctly encoded, **your dataframe should be saved as a tab (.txt/.tsv), comma or semicolon (.csv) delimited file**. Do not forget to select the appropriate parameters in control panel on the right.



Example dataset

An example dataset consisting of 151 Acute Myeloid Leukemia (AML) patients is provided. This public dataset is extracted from The Cancer Genome Atlas (TCGA) database with modified data for StatAid feature illustration. This dataset includes:

- **Numerical variables:**
  - Age: patients' age at disease diagnosis

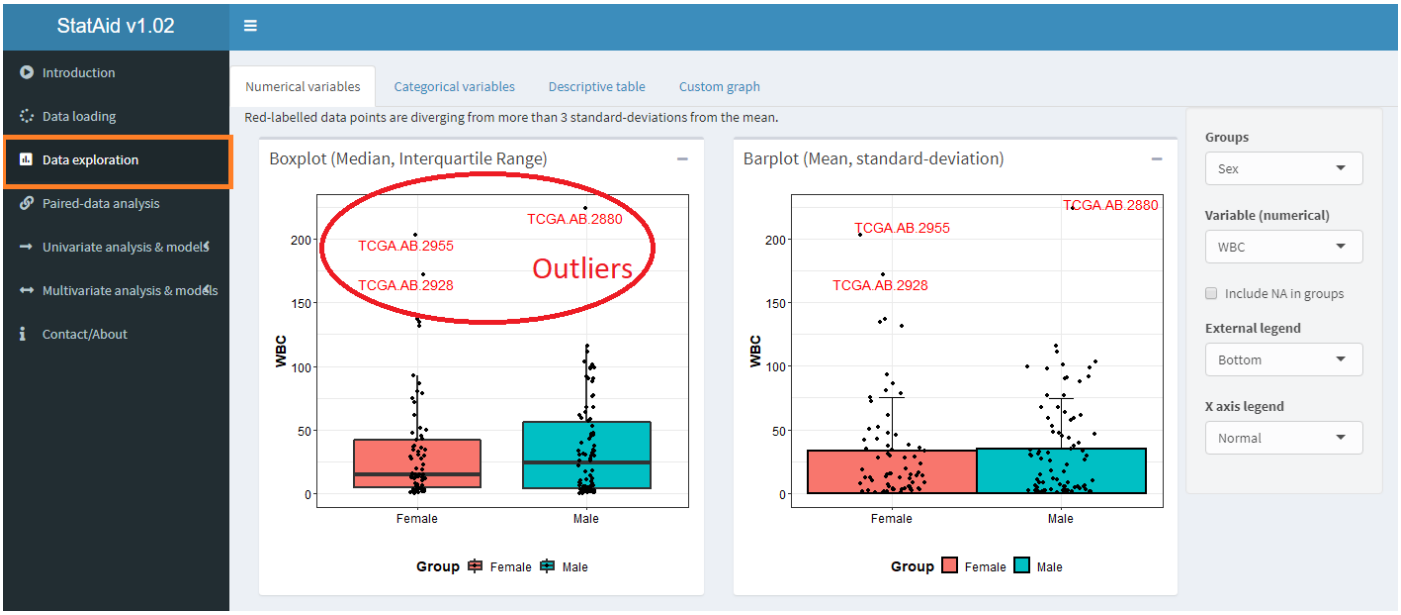
    - BM_BLAST_PERCENTAGE and PB_BLAST_PERCENTAGE: Bone Marrow and Peripheral Blood blast percentage, which can reflect the disease burden
    - WBC: white blood count, representing the total number of white cells in the peripheral blood, here again correlated with disease burden
    - Concentration_day0, 15 and 30: Simulated data of residual blood concentration level of a therapy. Added for paired-data analysis.
- **Categorical variables:**
    - FAB: French American British disease classification, from M0 to M7
    - ELN2017: Disease prognosis classification (Adverse – intermediate – Favorable)
    - Karyotype: Disease prognosis classification – genetic subtypes
    - TT_type: Type of therapy (intensive – HMA = hypomethylating agent (not intensive) – Supportive care – Other)
- **Time-dependent variable:** (NB: for STATUS variable, 0 = no event, 1 = event)
    - DFS MONTHS and Status: Disease free survival (time between diagnosis and disease relapse or censor)
    - OS MONTHS and STATUS: Overall survival (time between diagnosis and death or censor)
    - DFS_COMPRISK_STATUS: DFS Status with OS as a competing event (2).

2. DATA EXPLORATION

Once your dataset is loaded, the data exploration module can be launch. In this module, you can check the quality of your data, the distribution of your variables, detect outliers and variables with too much missing values.

a) Numerical variables

Numerical variables are variables encoded with only numeric values (they should not contain any character or symbol such as "<"). Variable distribution can be checked in the first two box (boxplot on the left with median and interquartile range, and barplot with mean / standard deviation on the right). Any value diverging from more than 3 standard-deviation from the mean is identified in red as a potential outlier. Missing values can be considered as an independent group by ticking the box in the control panel on the right.



/!\ A value identified as outlier is not necessarily a wrong / bad value, but rather a value that should be checked and require your attention. In the present example, the White Blood Count (WBC) is represented on the Y axis, according to the different sex (Female/Male). Patients identified as outliers are in fact presenting a really high WBC, which is associated with a highly active disease / poor prognosis.

The two next boxes are different way of representing data (histogram and density), according to your preferences. In the last box, you will find a table summarizing the total number of observations per group category, the number of missing value and the p-value of the Shapiro-Wilk test.

Normality assessment

Despite aberrant value detection and data quality checking, plotting numerical data is an important step to appreciate data distribution. Among the several possible distributions (gaussian, bimodal, negative binomial…), the gaussian or normal distribution is important as it is a prerequisite for numerous parametric statistical test. To assess the normality of a distribution, the first thing is to look at the scatter and boxplot and appreciate the gaussian/curve distribution. In addition, the Shapiro-Wilk test can be used. Remember that in this test, a p-value > 0.05 is in favor of a normal distribution. The Shapiro-Wilk test is only a supplementary tool to help you in your decision and should not be taken as a gold-standard. Normality assumption is not always easy and parametric test use can be discussed in case of a good-looking gaussian distribution with a Shapiro-wilk test < 0.05.

If your data shows a non-normal distribution, you can else use non-parametric statistical test or transform your variable (e.g. log-transform for classic numeric values or square root transform for percentages) to approach normality.

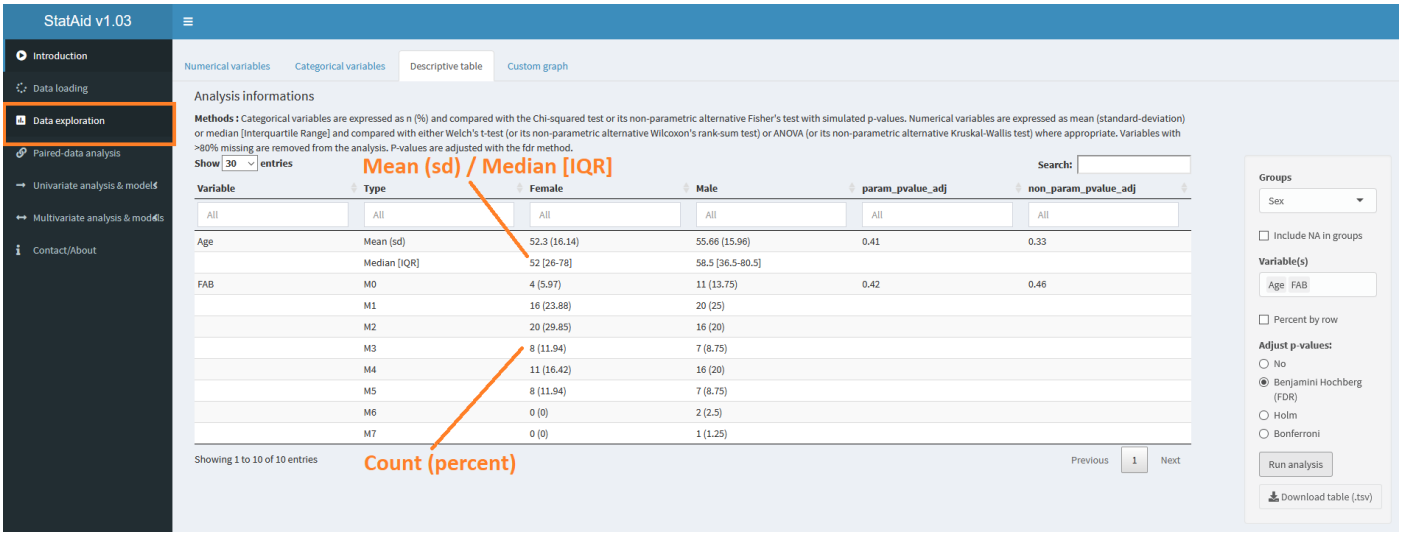
b) Categorical variables

Categorical variables can be controlled in this module. The first barchart plot the count per category while the second on the right plot the percent per category. You can check the boxes in the control panel on the right to include the missing values in the categories (groups) and/or the count (variable). A summary box is showing the total count in a table format.

c) Descriptive table

This module can be used to quickly check data parameters, make comparison or output a publication-ready descriptive table ("Table 1). In the control panel on the right, you can select the groups (categorical variable) you want to compare (or select "Whole cohort" if you do not want to compare groups). You then select the variables you want to compare (or "All" if you want to automatically output all your variables). Run analysis then launch the analysis and output the table.

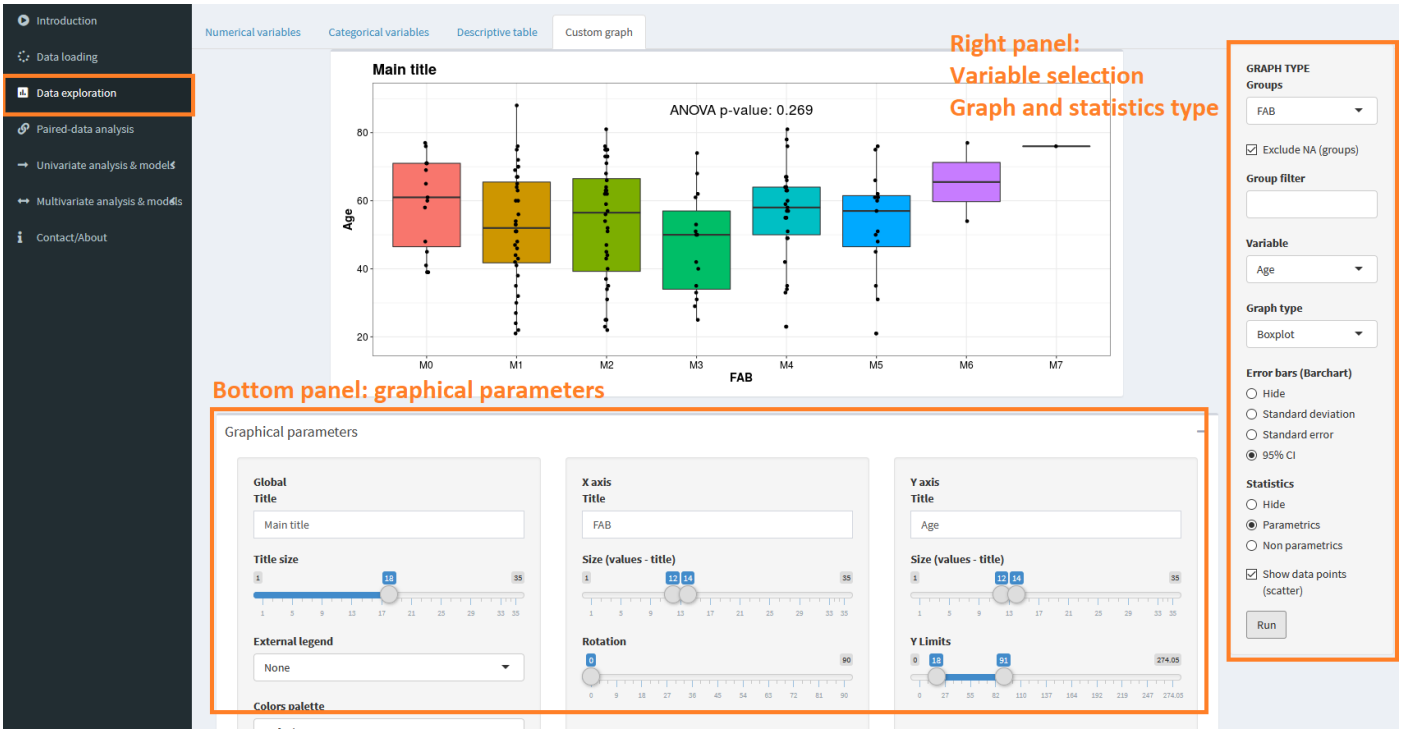
Mean (standard-deviation) and median [interquartile range] are shown for numerical variables, while count (percent) is shown for categorical variable. You can switch percentage from column (= percent of group) to row (= percent of variable category) on the right panel. P-values can else be non-adjusted or adjusted with the FDR, Holm or Bonferonni methods



/!\ Selecting "all" variables can take a long time to run if you have a big dataset. Be particularly careful if you have not cleaned your data yet and if you have variables with multiple categories. Be also aware that dates (01/01/2020 format for example) will be encoded as categorical variables and should be discarded from the descriptive table.

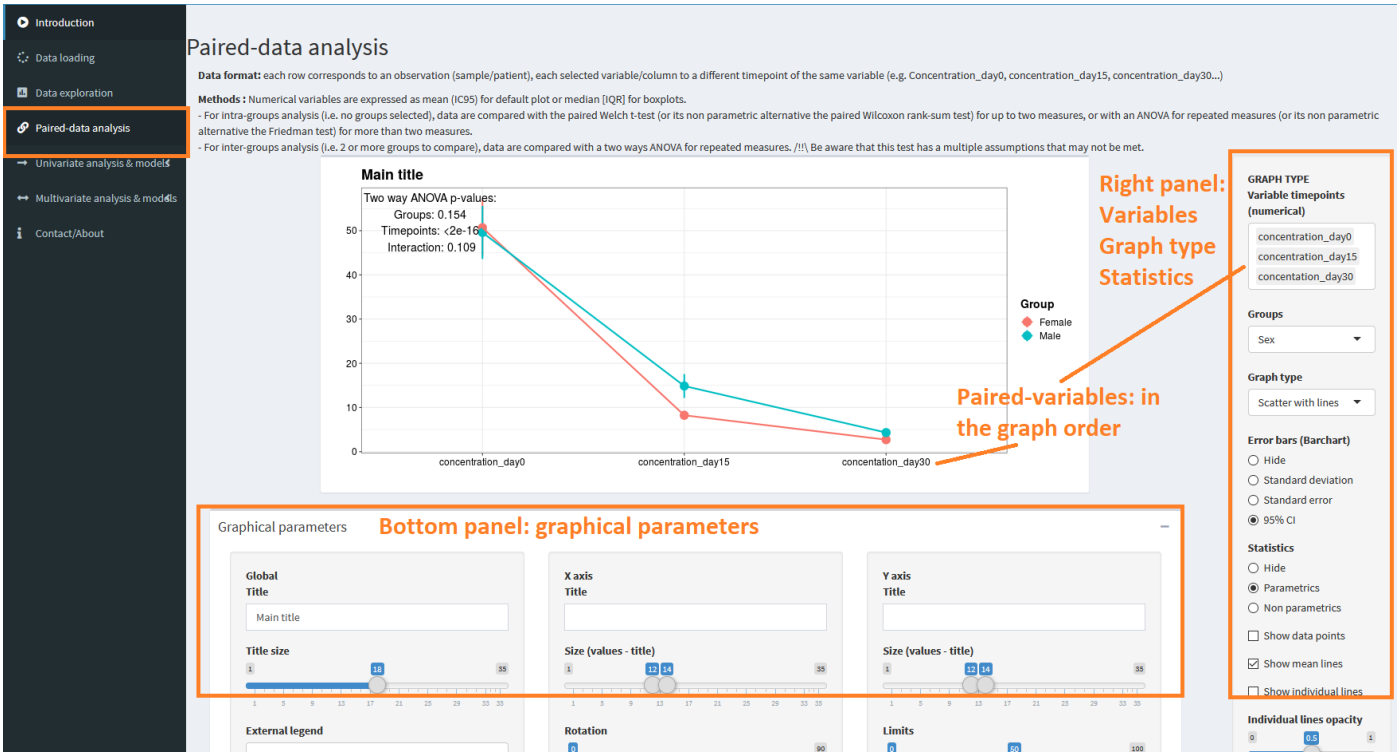
d) Custom graph

This module will allow you to plot customized graph fitting your needs for simple exploration and comparisons. Graph type and statistics parameters can be set in the right control panel. Graphical parameters such as font size, colors and legends can be modified in the bottom panel. The graph can be save by right-clicking on it.


Introduction
Data loading
Data exploration
Paired-data analysis
Univariate analysis & models
Multivariate analysis & models
Contact/About
Numerical variables
Categorical variables
Descriptive table
Custom graph
Main title
ANOVA p-value: 0.269
Age
FAB
M0
M1
M2
M3
M4
M5
M6
M7
Right panel:
Variable selection
Graph and statistics type
GRAPH TYPE
Groups
FAB
Exclude NA (groups)
Group filter
Variable
Age
Graph type
Boxplot
Error bars (Barchart)
Hide
Standard deviation
Standard error
95% CI
Statistics
Hide
Parametrics
Non parametrics
Show data points (scatter)
Run
Bottom panel: graphical parameters
Graphical parameters
Global
Title
Main title
Title size
External legend
None
Colors palette
X axis
Title
FAB
Size (values - title)
Rotation
Y axis
Title
Age
Size (values - title)
Y Limits

3. PAIRED-DATA ANALYSIS

In this module, you can easily plot and explore paired-data. First select the groups you want to compare (or none if you want to follow a variable in one group) and then select the different timepoints (i.e. paired data) on the right panel. Variables should be selected in the order you want them to appear on the graph, from left to right. Here again, multiple parameters can be tweaked, concerning the graph type and statistics on the right panel, and the graphical parameters on the bottom panel. A brief paragraph explains the methods used.

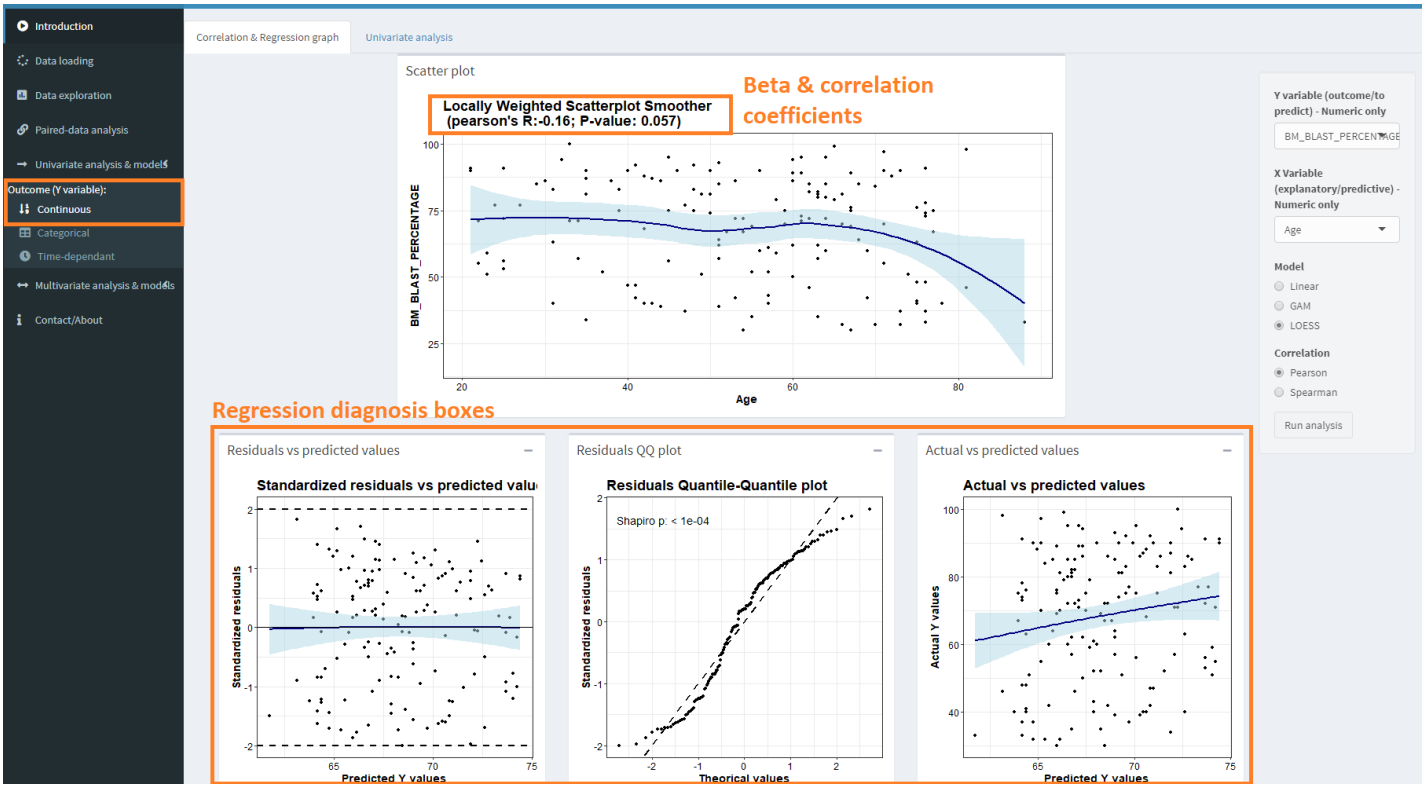


4. UNIVARIATE ANALYSIS

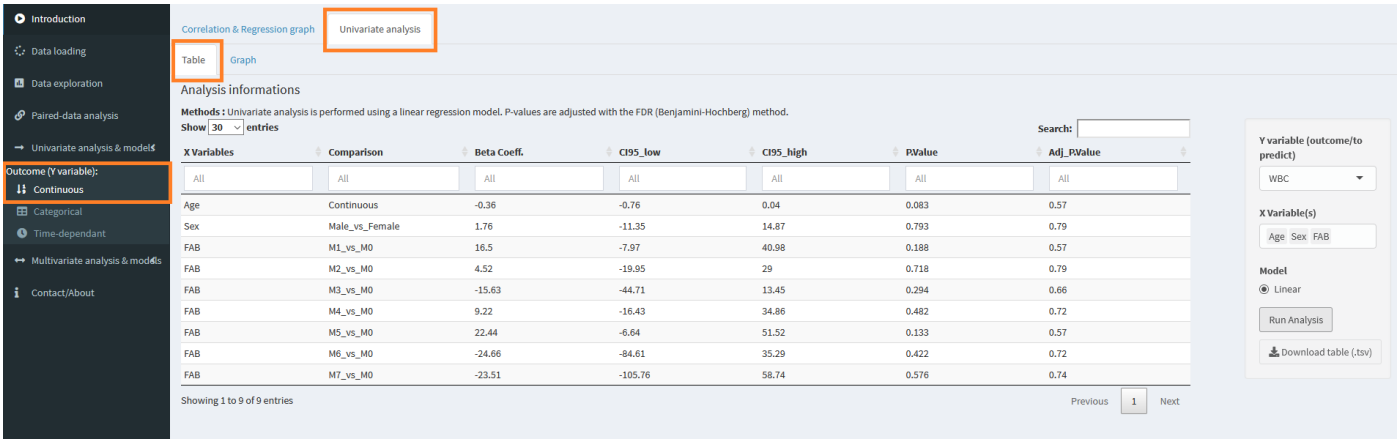
Once you have explored your dataset and checked your data quality (*e.g.* missing values, encoding errors, variables to excludes, variables to keep, numerical variable distribution allowing to perform parametric tests…) you can start your univariate analysis.

a) Continuous outcome

The first panel of the continuous outcome section allows you to plot correlation and regression line between two numerical variables. Regression model (linear, generalized additive or LOESS models) together with correlation type (Pearson or Spearman) can be selected in the right control panel. The main plot shows coefficients and p-values in the title. The three bottom boxes can be used for regression diagnosis and linear assumption tests.



The second panel (Univariate analysis) lead to the table panel. As in the descriptive table in the data exploration module, you can manually select X variables or select “All” to automatically study all variables. The run analysis button will output for each variable the linear regression’s beta-coefficient, its 95% confidence interval, p-value and FDR adjusted p-value. The “comparison” column precise the type of X-variable studied: in case of categorical variable, it will show which comparison have been performed. The above example shows that the Age variable has been studied as a numeric (continuous) variable, while the FAB variable is a categorical one for which each level is compared to the baseline (M0). Clicking on the “graph” panel after having run the analysis will lead to a graphical view of the same table.



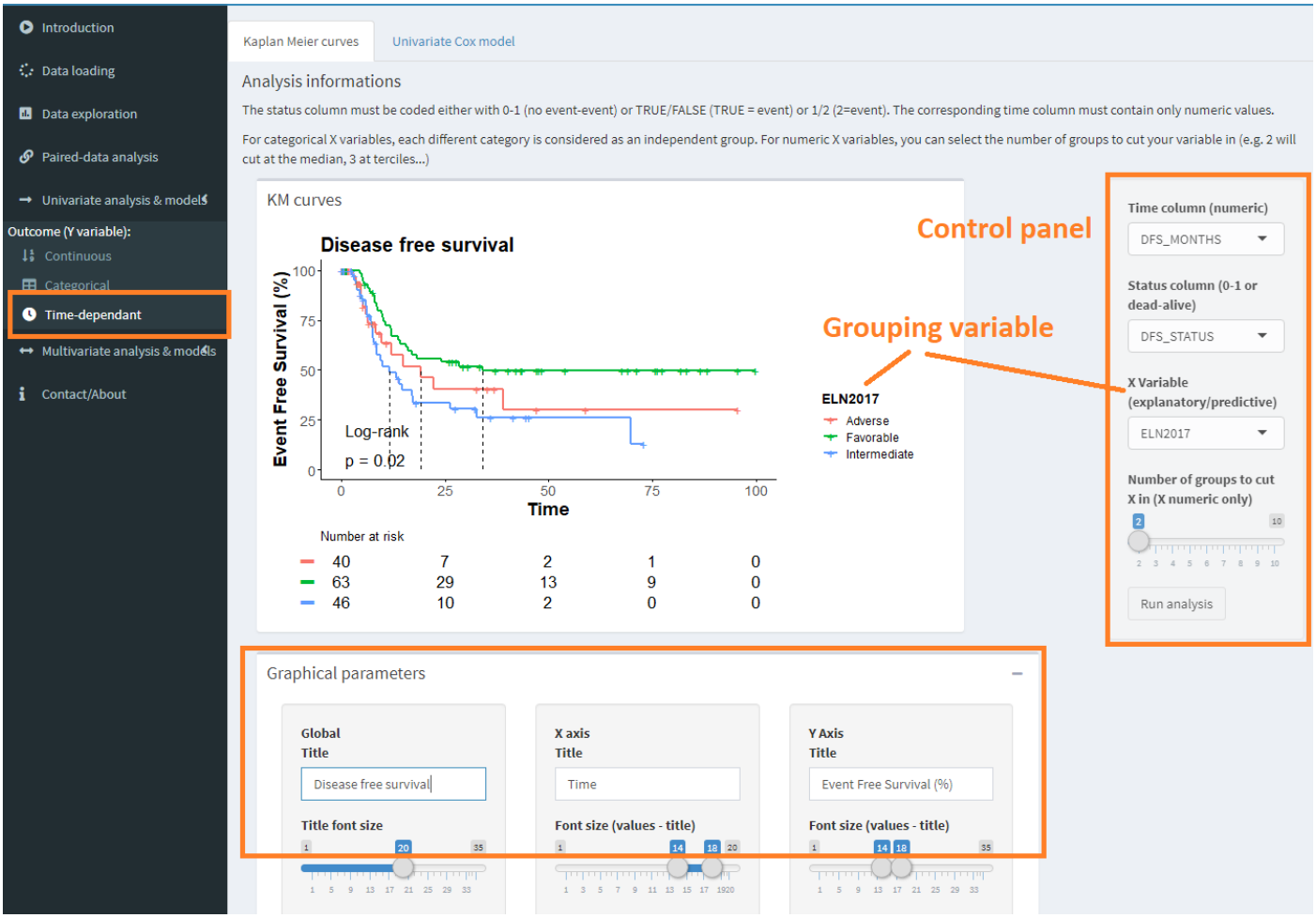
/!\ Note that for categorical variables, each level of the variables is compared to the baseline level (which is the first level by alphabetical order).

b) Categorical outcome

Univariate analysis module for categorical outcome works exactly like the one for continuous outcome except that you do not have the first graph panel. Only binomial model is implemented for the moment (i.e. you can only compare two groups in your Y variable). The beta-coefficient is replaced by the Odds ratio.

c) Time-dependent outcome

In this module, you can explore censored data such as survival data provided in the example dataset. The first panel allows you to draw Kaplan-Meier survival curves. You have to first select your time column (which should be a numeric variable, DFS_MONTHS in our example corresponding to the disease free survival in months) and the your status column (which should be encoded either with 0-1 (no event-event) or TRUE/FALSE (TRUE = event) or 1/2 (2=event), DFS_STATUS in our example encoded with 0 = no relapse, 1 = relapse). You can select a grouping variable (or no grouping variable with "Whole_cohort") in the right control panel. If your grouping variable is a numerical variable, you can select the number of groups you want to cut-it in. For example, selecting 3 groups with cut your numerical variables in tercile (4 in quartile and so on). Here again, graphical parameters can be tweaked in the bottom panel.



The second panel works exactly like for the other panels, except that hazard ratios are represented instead of beta coefficient or odds ratio.

5. MULTIVARIATE ANALYSIS

All the modules of the multivariate analysis are designed exactly like the one from the univariate analysis. The main difference is that the analysis will here be performed in a multivariate way. The coefficients shown (beta, odds ratio or HR) are thus calculated in a multivariate model taking into account all the selected X-variables, and not independalty for each variable as in the univariate module.

Keep in mind that as a rule of thumb, each additional X variable in your multivariate model require ~10 samples/patients to have enough power. If you want to include 6 variables in your model, you should aim to have at least 60 samples/patients!