Evaluate Cutpoints user manual

July 2017

Contents

Introduction....3
System architecture....4
System requirements and installation....5
Launching the application....6
Data format and upload....8
    Data format....8
    Data upload....9
Example data....9
Methods for cutpoint determination....10
    Two - group methods....10
        1. Methods based on significance of correlation with binary outcome:....10
        2. Methods based on significance of correlation with survival time:....10
        3. Adaptive (manual) selection of the cutpoint value:....10
    Three - group methods....11
Plots....11
    ROC curve....12
    Histogram....13
    Kaplan - Meier plot....14
    Standardized log-rank statistics plot....15
    Heatmap....16
    Saving the plots....16
References....17

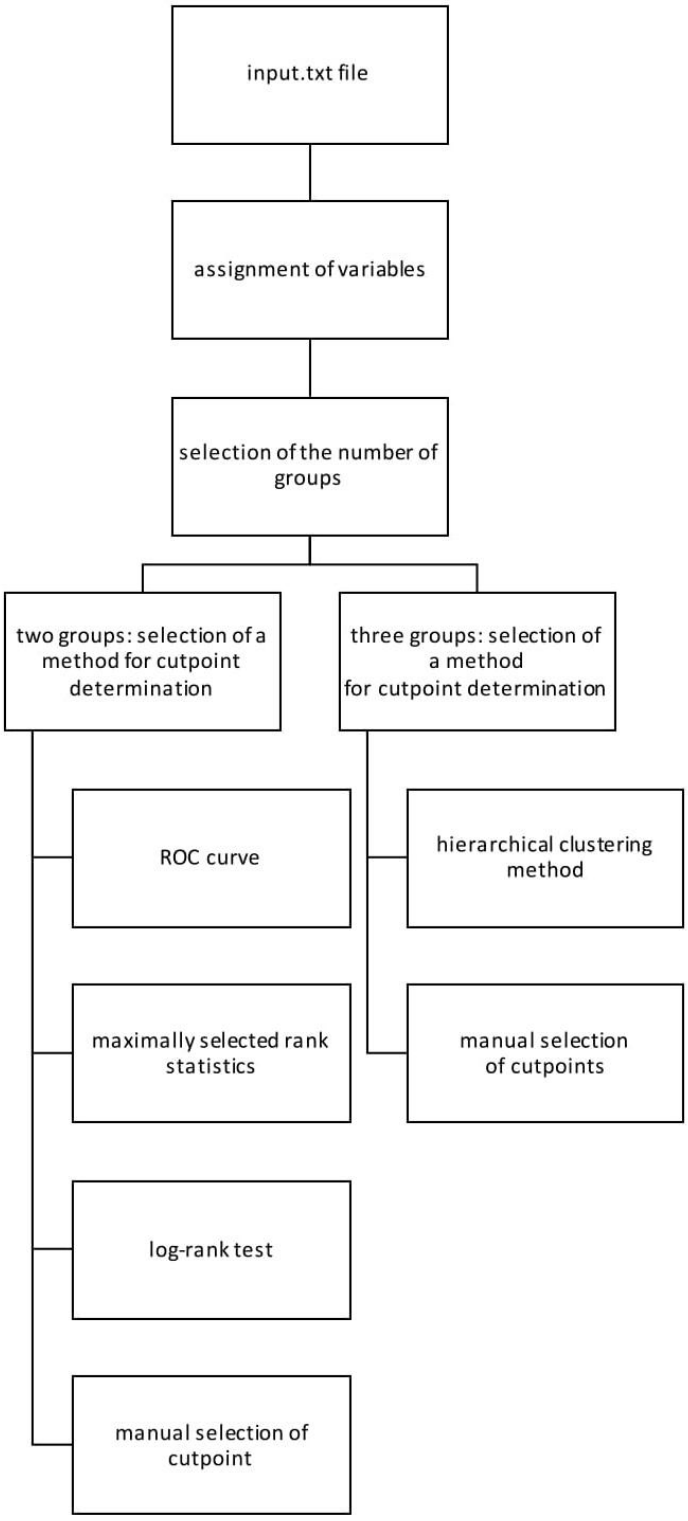

Figure 1. General workflow of Evaluate Cutpoints application.

Introduction

Evaluate Cutpoints is R Shiny - based user - friendly application for the determination and optimization of cutpoints in biological data, which employs few different algorithms. The application is freely available at http://wnbikp.umed.lodz.pl/Evaluate-Cutpoints/ in form of R scripts to be downloaded and launched locally on your computer. The graphical output of Evaluate Cutpoints can be presented in scientific publications with no limitations, but we kindly ask you to cite the publication.

Usually, the data, which result from molecular experiments, are derived in form of continuous or categorical binary variables. To evaluate their clinical significance and understand their meaning the key step is the determination optimal cutpoint with respect to which cohort patients will be stratified into two or three groups differing in prognosis. Evaluate Cutpoints offers algorithms for stratification into two or three groups employing distinct cutpoints optimization methods and resulting graphical output. Each method is described in more details in following sections of the manual.

The general workflow of application and available algorithms is shown in Figure 1. Briefly:

1. the input dataset can be any tab - separated file. Observations should be placed in rows, columns should represent variables.
2. The next step is to select the number of groups that patients will be stratified into (two or three).
3. Then, the user defines variables required to perform the analysis: biomarker, survival time and outcome. If the user wants to split the cohort into two groups, one also determines whether the cutpoint should be estimated manually or based on the significance of correlation with the binary outcome or survival. Additionally, the user indicates the statistical method that will be used to calculate the cutpoint. If the user selects stratification of the population into three groups, one also indicates if the cutpoint should be estimated using the hierarchical clustering method, or whether cutpoints will be chosen manually.

4. The query is then processed by the server. The result includes estimated cutpoints and graphical visualization.

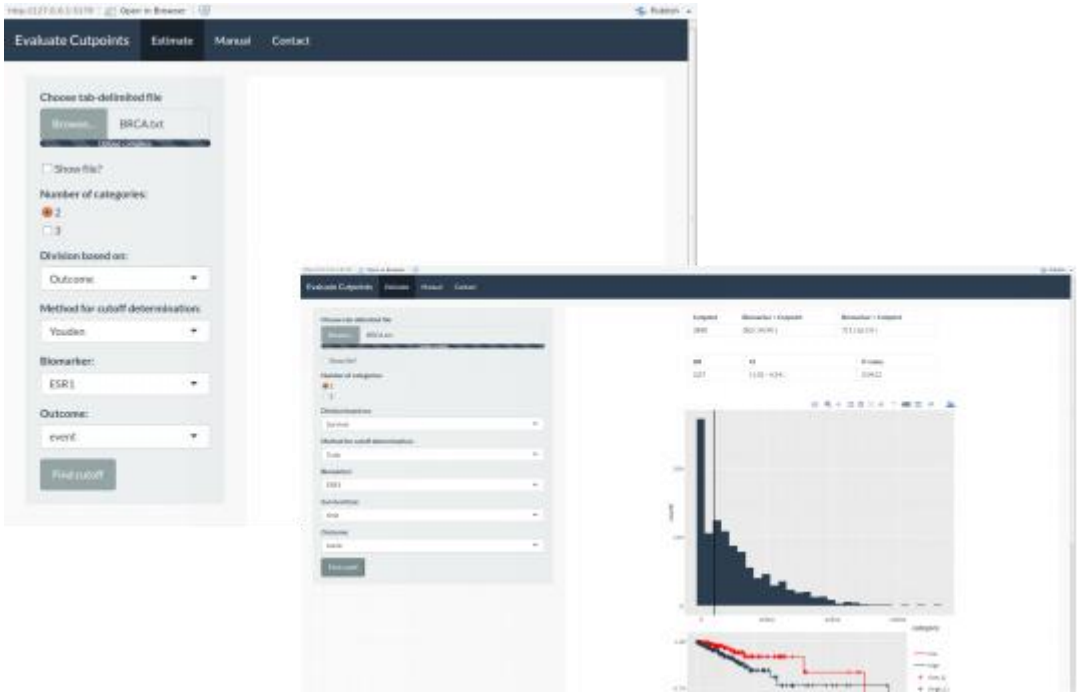

Figure 2. Screenshot of Evaluate Cutpoints application.

System architecture

Evaluate Cutpoints is the application developed using the R language, Shiny framework and R packages: `survival`, `survMisc`, `OptimalCutpoints` [1], `maxstat` [2], `rolr`, `ggplot2`, `GGally` and `plotly`. It consists of two main layers – the first one dynamically generates HTML, the second separates the data analysis in real - time. Software is available through the web interface locally launched by the user.

System requirements and installation

Evaluate Cutpoints is the application developed and based on R language. Therefore the user is obliged to install the newest version of R environment, RStudio application and adjacent R packages to ensure proper functioning of the application.

R environment is freely available for every operation system and may be downloaded and installed from R project webpage (https://cran.r-project.org/) according to the instructions. RStudio Desktop is freely available for every operation system graphical interface for R, which also enables Shiny applications to run. It may be downloaded from RStudio home webapge (https://www.rstudio.com/) and installed according to instructions. The R package that is essential for initialization of the Evaluate Cutpoints is shiny. The installation process of shiny R package and initialization of Evaluate Cutpoints application is described hereafter.

Launching the application

Right after installation of required software (R, RStudio) start with installation of shiny package.

1. Launch RStudio.
2. In the Console window type:

   `install.packages("shiny")`

   and confirm with Enter.
3. Successful installation of the package will be terminated with following messages:

```
> install.packages("shiny")
Installing package into ‘/home/akb/R/x86_64-pc-linux-gnu-library/3.4’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/src/contrib/shiny_1.0.3.tar.gz'
Content type 'application/x-gzip' length 2273603 bytes (2.2 MB)
==================================================
downloaded 2.2 MB

* installing
* source
* package ‘shiny’ ...
** package ‘shiny’ successfully unpacked and MD5 sums checked
** R
** inst
** preparing package for lazy loading
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (shiny)

The downloaded source packages are in
        ‘/tmp/Rtmp0Ko2h5/downloaded_packages’
```

4. All four downloaded scripts (server.R, global.R, ui.R, styles.css) place in the same localization on your computer.

5. Open with RStudio server.R file. To open the file click on it with the right mouse button and select Open with -> RStudio or go to RStudio -> Ctrl+O -> server.R. Properly opened script should generate in RStudio new window above Console (Fig. 3).

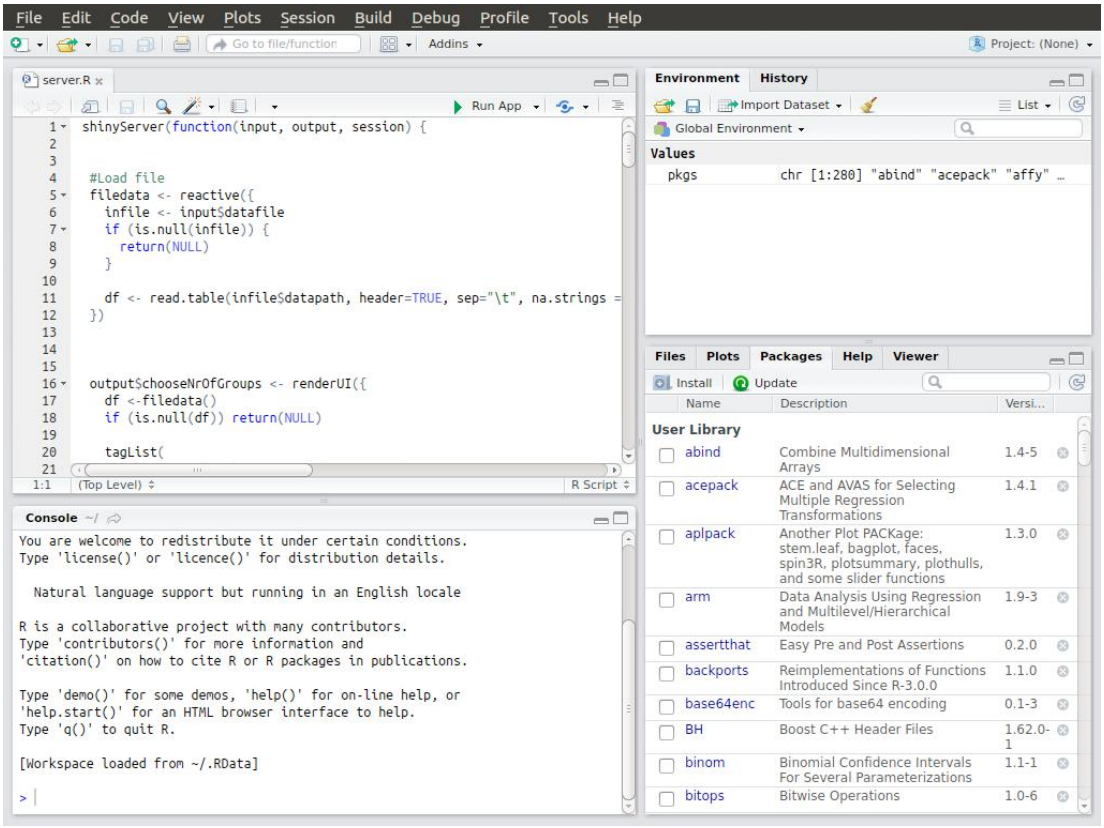

Figure 3. RStudio with opened server.R.

6. In the right corner of opened server.R click Run App and wait.

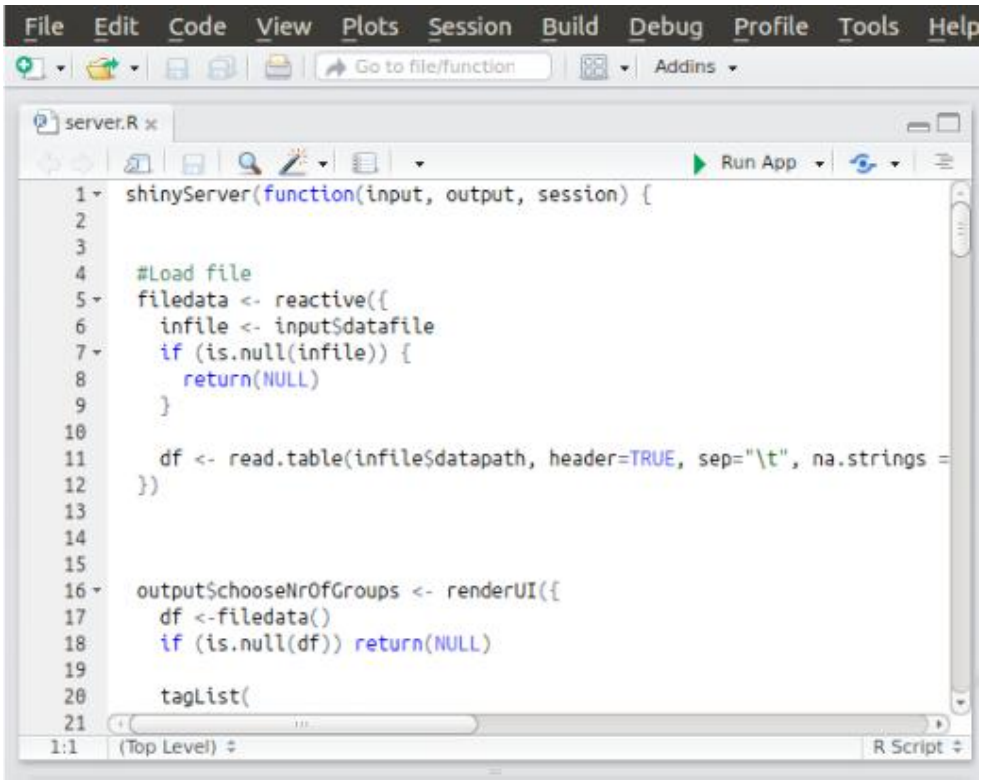

Figure 4. Run Evaluation Cutpoints.

7. A new browser window should open with launched ready-to-use Evaluate Cutpoints application.

Data format and upload

Data format

1. Tab - delimited .txt file.
2. Observations should be placed in rows.
3. Variables should be placed in columns.
4. Each column must have a header describing its content.
5. Binary categorical variables should be represented as 0 and 1.
6. Empty cells should be labeled as `NA` (not available).
7. The decimal separator should be . (dot) instead of , (comma).

8. All table content should be numeric excepting column headers.
9. The application accepts negative values.

Data upload

1. In Evaluate Cutpoints application window choose your tab - delimited file from your computer and click Find cutoff, which will be followed by expansion of bars with possible algorithms to be selected for analysis.
2. Select number of categories, into which observations will be stratified.
3. Select method of the division:
    a) outcome,
    b) survival,
    c) rolr (available only for three groups),
    d) adapt.
4. Select method of cutpoint determination:
    a) Youden index or ROC for outcome,
    b) cutp or maxstat for survival.
5. Select your variables for biomarker, time and outcome.
6. Click Find cutoff.

Example data

We hereby released two example data sheets for analysis using Evaluate Cutpoints application. This can be done by downloading the chosen .txt file and uploading it into the application. The example data include part of The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) breast cancer cohort such as expression profiling (RNAseqV2, RSEM normalized; data status of 28th Jan 2016) and clinical information for over 1000 patients, which are publicly and freely available. Example analyses may be performed by using expression measurements of estrogen (*ESR1*), progesterone (*PGR*) or human epidermal growth factor 2 (*ERBB2*) receptors as biomarker,

immunohistochemical determination of receptor presence or disease recurrence as outcome variable and time as survival time.

Methods for cutpoint determination

Two - group methods

Stratification of the population into two groups can be performed by applying four different algorithms based on significance of correlation with binary outcome or survival time. The user can also determine the cutpoint value manually.

1. **Methods based on significance of correlation with binary outcome:**

Youden index and minimization of the distance between PROC plot and point (0,1). Application uses the R package OptimalCutpoints to generate ROC plots and estimate: cutpoints, positive predictive value (PPV), negative predictive value (NPV), sensitivity (Se), specificity (Sp), positive diagnostic likelihood ratio (LR+), negative diagnostic likelihood ratio (LR-), false positives (FP), false negatives (FN). User may find more details in `OptimalCutpoints` user's manual at CRAN repositories.

2. **Methods based on significance of correlation with survival time:**

maximally selected rank statistics and Cox proportional hazard model. Calculation of maximally selected rank statistics and estimation of cutpoint (the point with the most significant split based on the standardized log - rank test) is performed with the use of `maxstat` R package. As for the second method, application uses `coxph` function from the `survival` R package to fit Cox proportional hazard model to the binary (outcome) and continuous (survival time and biomarker value) covariates. Cutpoint is then computed with the `cutp` function (`survMisc` R package). User may find more details in `maxstat`, `survival` and `survMisc` user's manual at CRAN repositories.

3. **Adaptive (manual) selection of the cutpoint value:**

user can select a cutpoint value based on a scalable, interactive heatmap (generated with the use of `plotly` R package) that illustrates all cutpoints (biomarker values arranged from lowest to highest). The intensity of the color (from blue to yellow) of

each field represents the probability value, calculated using the `coxph` function from the survival R package. Selection of a field on the heatmap results in generation of Kaplan-Meier plot and a table with estimated hazard ratio, 95% confidence intervals and p-value. Cutpoint adaptability increases the value of the algorithm, since the statistically significant cutpoint value may not be optimal regarding biology or medicine. In such cases, despite lower statistical significance, we may receive better information from a clinical point of view. User may find more details in `plotly` and `survival` user's manual at CRAN repositories.

Three - group methods

Stratification of the population into three groups based on survival time and binary outcome can be executed in the application with the hierarchical clustering method [3] (`rhier` function, `rolr` R package). Firstly, the algorithm splits the cohort into two groups by estimation of the optimal cutpoint with the highest log-rank statistics. The procedure is then repeated in the resulting groups to obtain two supplementary cutoff values. Second optimal cutpoint is the one with larger test statistics. Application omits all rows (observations) with `NA` values. User can also manually select two cutpoints from the sliders to observe the changes in the Kaplan-Meier plot and pairwise comparison of hazard ratio, 95% confidence intervals and p-values between the resulting groups. User may find more details in `rolr` user's manual at CRAN repositories.

Plots

Evaluate Cutpoints application generates graphical output representing cutpoints determined by one of available methods (all available methods has been described in previous section of this manual). In the following sections we present all types of plots that may be generated by Evaluate Cutpoints application using example data. The user can generate and save all graphical output (described below).

ROC curve

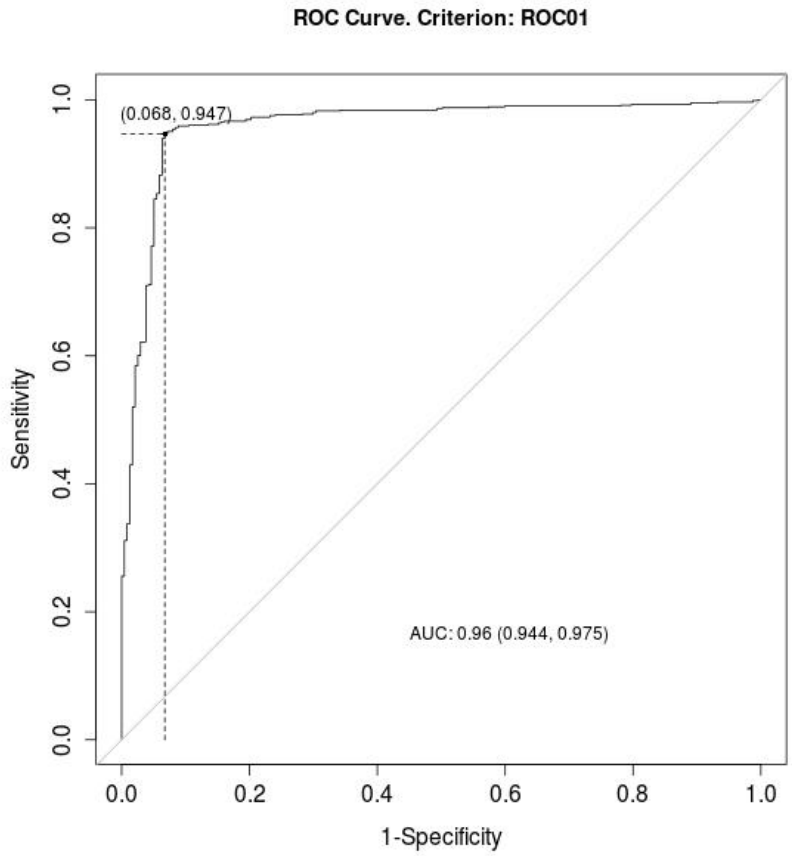

Figure 5. ROC curve as a result of performed Youden index method with optimal cutpoint for *ESR1* and the quality of the prediction assessed by the area under the curve (AUC).

ROC curve presented in Figure 5 has been generated by applying Youden index outcome method using example data for the prediction of immunohistochemically determined status of estrogen receptor (*ESR1*) by its expression. The quality of evaluation is expressed by area under curve (AUC) and specificity and sensitivity measures at determined cutpoint. ROC curve is also extended with statistic table that contains values for cutpoints, positive predictive value (PPV), negative predictive value (NPV), sensitivity (Se), specificity (Sp), positive diagnostic likelihood ratio (LR+), negative diagnostic likelihood ratio (LR-), false positives (FP), false negatives (FN).

Histogram

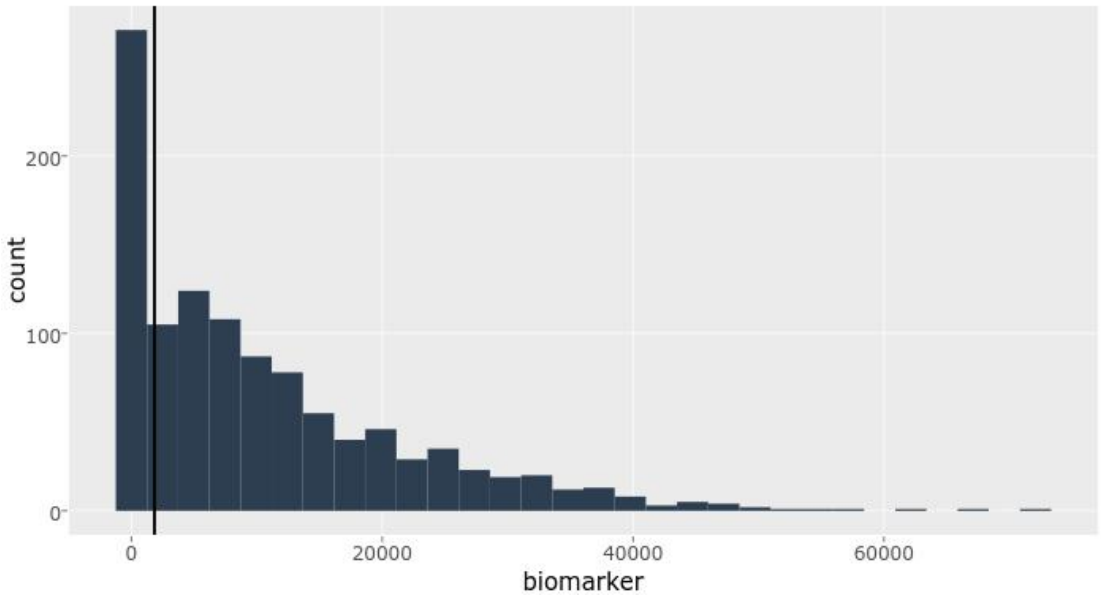

Figure 6. Population of example dataset stratified into two groups based on computed cutpoint of *ESR1* expression.

Histogram (Fig. 6) shows biomarker (*ESR1*) expression with determined cutpoint (vertical line) followed by number of individuals (samples) stratified and assigned to each group.

Kaplan - Meier plot

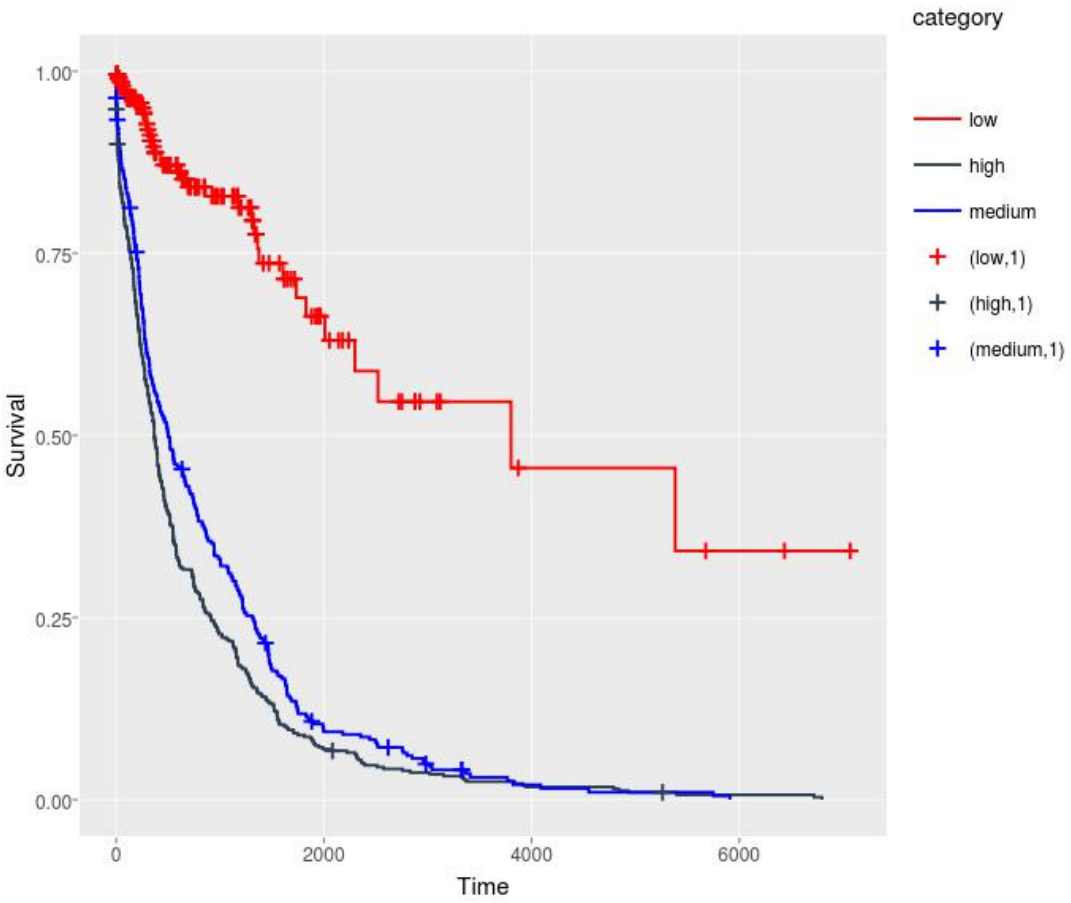

Figure 7. Survival analysis for three groups performed at optimized cutpoint.

Kaplan-Meier curve for two or three groups (Fig. 7) has been plotted for optimized cutpoint based on chosen method. Additionally, all survival analysis statistics like hazard ratio (HR) with 95% confidence intervals (CI) and p-value are enclosed in adjacent table.

Standardized log-rank statistics plot

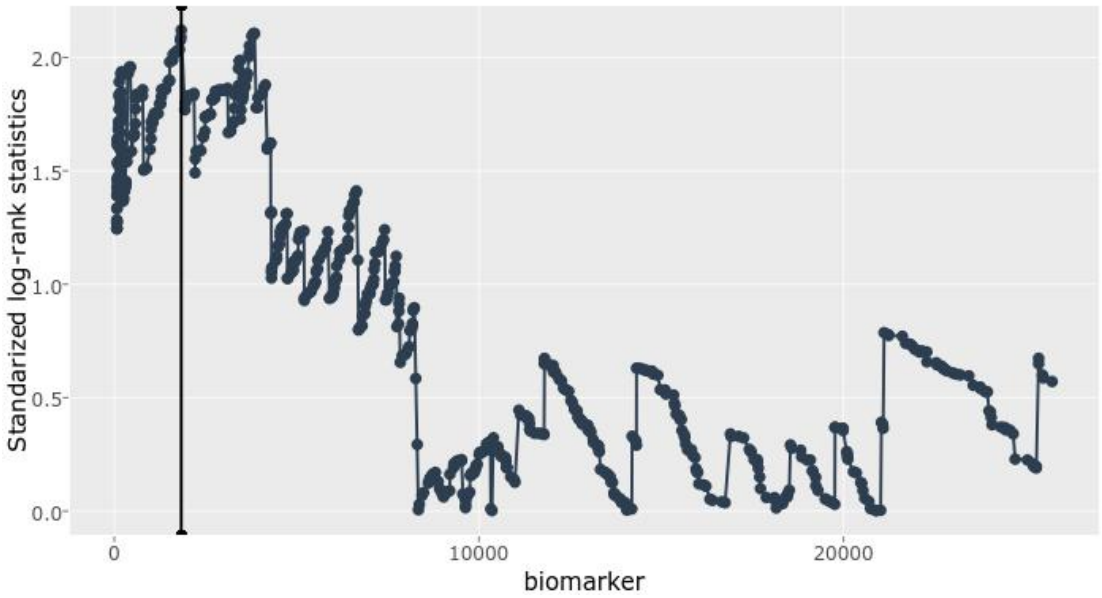

Figure 8. Standarized log-rank statistics plot.

Figure 8 shows standardized log-rank statistics along biomarker expression at determined optimal cutpoint.

Heatmap

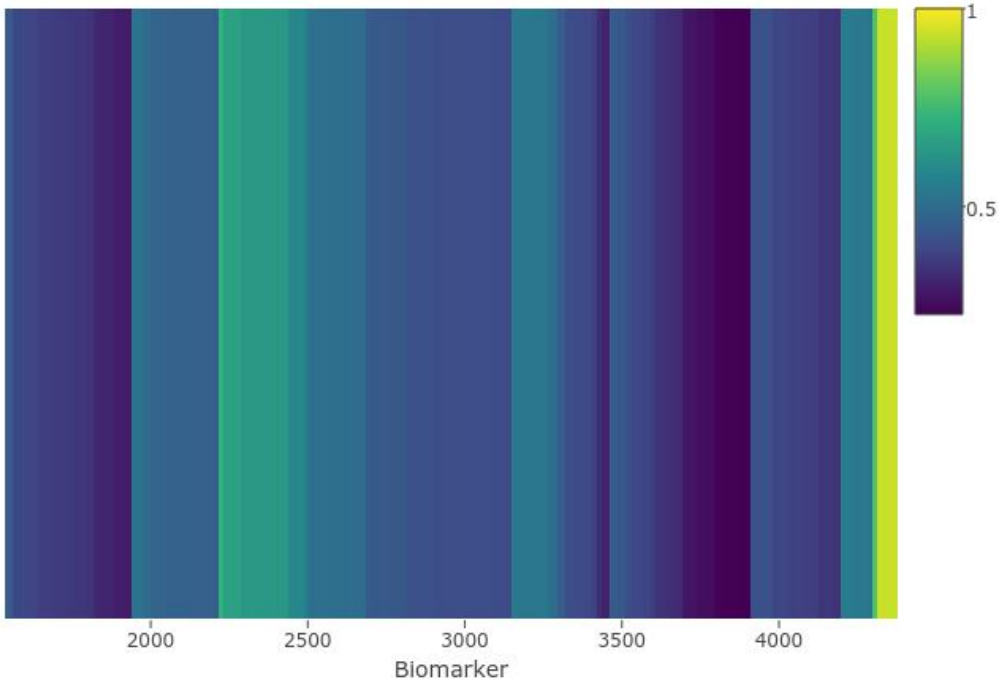

Figure 9. Heatmap for manual cutpoint selection.

Figure 9 shows heatmap for manual cutpoint selection and the intensity of the colors corresponds with statistical significance (p-value) of each biomarker expression value. The darker color the more significant is the chosen cutpoint.

Saving the plots

Each plot can be exported and saved locally on your computer through the icon located in the bar above the figure. The bar contains all available options for figures like exporting to png, zooming, auto-scaling etc.

References

1. López-Ratón M, Rodríguez-Álvarez M, Cadarso-Suárez C, Gude-Sampedro F. OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. J Stat Softw Artic. 2014;61(8):1–36.

2. Hothorn T, Lausen B. On the exact distribution of maximally selected rank statistics. Comput Stat Data Anal. 2003;43(2):121–37.

3. Leblanc M, Crowley J. Survival Trees by Goodness of Split. J Am Stat Assoc. 1993;88(422):457–67.