Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

OmicSelector: Docker-based application and R package for biomarker signature selection from high-throughput omic experiments and deep learning model development.

Konrad Stawiski
Medical University of Lodz

Marcin Kaszkowiak
Medical University of Lodz

Damian Mikulski
Medical Univeristy of Lodz

Dipanjan Chowdhury
Dana-Farber Cancer Institute

Wojciech Fendler
Medical University of Lodz

Abstract

The crucial phase of modern biomarker discovery studies is a selection of most promising features from the results of high-throughput screening assays. Here, we present the OmicSelector - Docker-based web application and R package that facilitates the analysis of such experiments. OmicSelector provides a consistent and overfitting-resilient pipeline that integrates 94 feature selection approaches based on 25 distinct variable selection methods. It identifies and ranks the best feature sets, basing on 12 modeling algorithms (including GPU-based deep learning) with hyperparameter optimization in hold-out or cross-validation. OmicSelector provides classification performance metrics for proposed feature sets, which allow researchers to choose the overfitting-resistant biomarker set with the most significant diagnostic potential. Lastly, it allows for development, validation and implementation of deep learning feedforward neural networks (up to 3 hidden layers) on selected signature. Application performs extensive grid search of hyperparameters including balancing and preprocessing with additional autoencoders. The pipeline is applicable for selecting candidate circulating or tissue miRNAs, RNAs, methylation data, metabolites, or proteins. The tool is open-source and available at https://biostat.umed.pl/OmicSelector/.

Keywords: feature selection, biomarker, data-mining, next-generation sequencing, omics, deep learning, artificial neural network, R, Docker.

1. Introduction

Broad-scale treatment personalization is one of the most significant modern medicine challenges, requiring accurate and cost-effective diagnostic tests. Such methods rely heavily on biomarkers, which are usually discovered using omic techniques. Although high-throughput experiments enable us to gather the biological measurements of an extensive amount of biomarker candidates, translating the results to the clinical bedside remains troublesome.

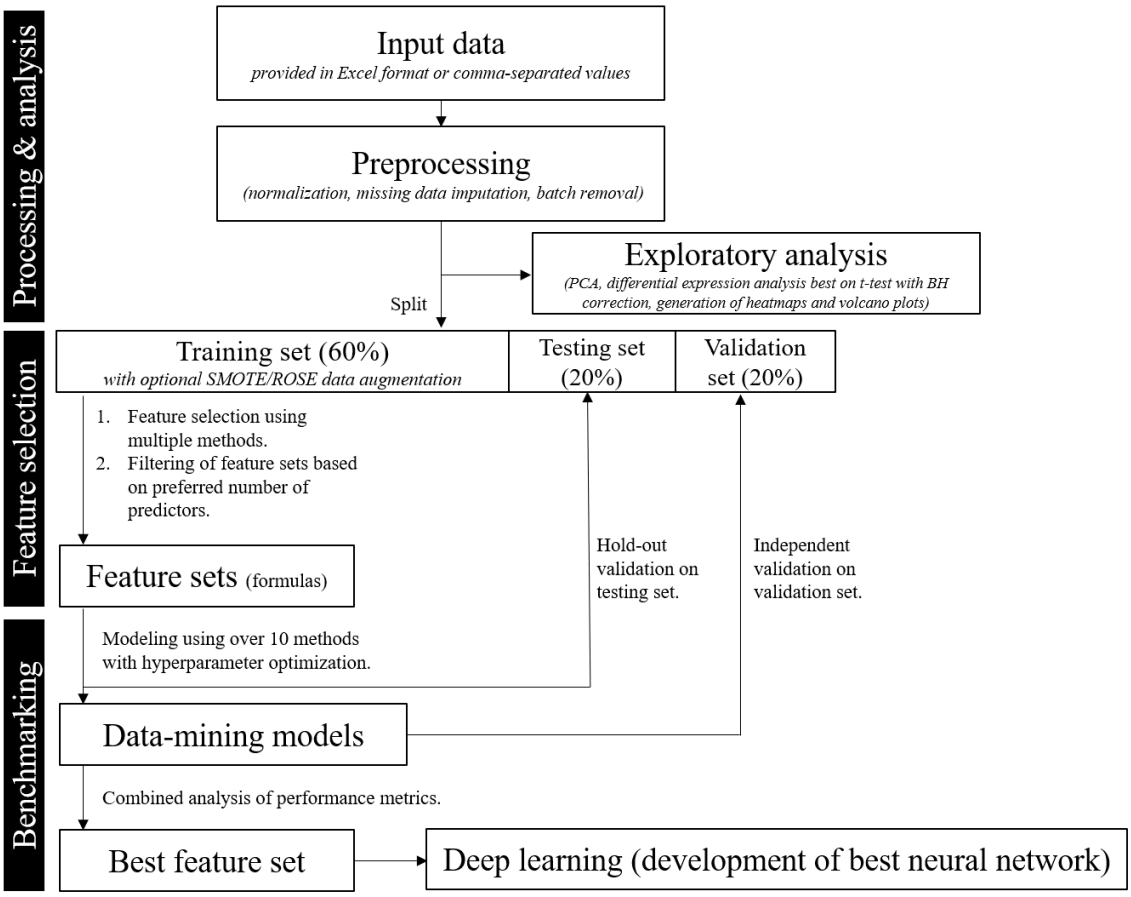
The typical biomarker study comprises of discovery and validation phases. (Goossens, Nakagawa, Sun, and Hoshida (2015)) In the former, high-throughput screening is usually performed to measure the values of multiple features. Those are further assessed to determine their diagnostic potential. In the validation phase, only selected variables are measured, typically in a new set of samples, with a cheaper and/or more accessible method. Our team has been working on microRNA (miRNA) biomarkers for radiation (Dinh, Fendler, Chałubińska-Fendler, Acharya, O'Leary, Deraska, D'Andrea, Chowdhury, and Kozono (2016)) and cancer (Elias, Fendler, Stawiski, Fiascone, Vitonis, Berkowitz, Frendl, Konstantinopoulos, Crum, Kedzierska, Cramer, and Chowdhury (2017)), but trouble with the reproducibility of selected biomarker performance (Acharya, Fendler, Watson, Hamilton, Pan, Gaudiano, Moskwa, Bhanja, Saha, Guha, Parmar, and Chowdhury (2015); Fendler, Malachowska, Meghani, Konstantinopoulos, Guha, Singh, and Chowdhury (2017); Małachowska, Tomasik, Stawiski, Kulkarni, Guha, Chowdhury, and Fendler (2020)) or reference identification (Pagacz, Kucharski, Smyczynska, Grabia, Chowdhury, and Fendler (2020)). Similar challenges, caused by bias and overfitting, hindered the attempts of other groups to develop validated, efficient omic-driven biomarkers. (Dobbin, Cesano, Alvarez, Hawtin, Janetzki, Kirsch, Masucci, Robbins, Selvan, Streicher, Zhang, Butterfield, and Thurin (2016))

Cohorts used in the discovery phase are usually small due to the high cost of high-throughput assays, which makes the experiments vulnerable to overfitting and results in false-positive biomarker candidates that fail in external validation. (Smialowski, Frishman, and Kramer (2009)) For example, a recent review of serum miRNA biomarkers for pancreatic cancer (Xue, Jia, Ren, Lindsay, and Yu (2019)) highlights how various miRNA sets are chosen in different studies, with each study reporting unrealistically optimistic results. Thus, correct and overfitting-resistant feature selection is critical in biomarker studies.

In this paper, we try to tackle this problem by designing software for systematic, overfitting-resistant, and informative feature selection. The analytical steps of our package entail (Figure 1): splitting of the dataset into training, testing and validation sets, differential expression analysis and performing up to 94 different feature selection procedures on the training set. Feature sets (formulas) are further validated by training 12 models of various architectures with hyperparameter optimization based on hold-out- or cross-validation. Our toolset enables the users to make an informed decision about the most appropriate feature selection method and informs them about their predictive abilities using different modeling approaches. Finally, as the most flexible method, users are able to train and implement final deep feed-forward neural network (up to 3 hidden layers, with or without autoencoders; grid search of hyperparameters) for classification (diagnostic) problem.

2. Implementation

Figure 1: ***The pipeline of OmicSelector analysis.*** *Abbreviations: PCA - principal component analysis, BH - Benjamini-Hochberg procedure, SMOTE/ROSE - data balancing methods explained in the main text.*




```
R> library(OmicSelector)
R> sessionInfo()

R version 4.0.3 (2020-10-10)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.5 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.2.20.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
```

```
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
[1] OmicSelector_1.0.0 MASS_7.3-53

loaded via a namespace (and not attached):
[1] compiler_4.0.3    snow_0.4-3        parallel_4.0.3
[4] tools_4.0.3       codetools_0.2-18  doParallel_1.0.16
[7] iterators_1.0.13  foreach_1.5.1
```

N	ID	Description
1	all [1]	Get all features (all features staring with 'hsa' in the name). We assume that the most frequent application of the pipeline will be for human-related expression measurements.
2	sig [2]sigtop [2]sigtopBonf[2]sigtopHolm [2]topFC [2]sig_SMOTE [2]sigtop_SMOTE [2]sigtopBonf_SMOTE [2]sigtopHolm_SMOTE [2]topFC_SMOTE [2]	Selects features significantly differently expressed between classes by performing unpaired t-test with and without correction for multiple testing.We get: **sig** - all significant (adjusted p-value less or equal to 0.05) miRNAs with comparison using unpaired t-test and after the Benjamini-Hochberg procedure; **sigtop** - sig limited only to the number of features preffered by an user (selecting top after sorting by p-value), **sigtopBonf** - uses Bonferroni instead of BH correction, **sigtopHolm** - uses Holm–Bonferroni instead of BH correction, **topFC** - selects prefered number of features based on decreasing absolute value of fold change in differential analysis.
3	fcsig [3]fcsig_SMOTE [3]	Features significantly differently expressed with absolute log2FC greater than 1. (Thus, features significantly up- or down-regulated in the higher magnitudes)
4	cfs [4]cfs_SMOTE [4]cfs_sig [4]cfs_SMOTE_sig [4]	Correlation-based feature selection (CFS) - a heuristic algorithm selecting features that are highly correlated with class (binary) and lowly correlated with one another. It explores a search space in best-first manner, until stopping criteria are met.
5	classloop [5] classloop_SMOTE [6] classloop_sig [7] classloop_SMOTE_sig [8]	Classifier loop - performs multiple classification procedures using various algorithms (with embedded feature ranking) and various performance metrices. Final feature selection is done by combining the results. Modeling methods used: support vector machines, linear discriminant a nalysis, random forest and nearest shrunken centroid. Features are selected based on the AUC ROC and assessed in k-fold cross-validation according to the documentation.
6	fcfs [9] fcfs_SMOTE [10] fcfs_sig [11] fcfs_SMOTE_sig [12]	An algorithm similar to CFS, though exploring search space in greedy forward search manner (adding one, most attractive, feature at the time, until such addition does not improve set's overall quality). Based on Wang et al. 2005 and documented here.
7	fwrap [13] fwrap_SMOTE [14] fwrap_sig [15] fwrap_SMOTE_sig [16]	A decision tree algorithm and forward search strategy documented here.
8	AUC_MDL [17] AUC_MDL_SMOTE [20] AUC_MDL_sig [23] AUC_MDL_SMOTE_sig [26]	Feature ranking based on ROC AUC and minimal description length (MDL) discretization algorithm documented here.
9	SU_MDL [18] SU_MDL_SMOTE [21] SU_MDL_sig [24] SU_MDL_SMOTE_sig [27]	Feature ranking based on symmetrical uncertainty and minimal description length (MDL) discretization algorithm documented here.
10	CorrSF_MDL [19] CorrSF_MDL_SMOTE [22] CorrSF_MDL_sig [25]	Feature ranking based on CFS algorithm with forward search and minimal description length (MDL) discretization algorithm documented here.

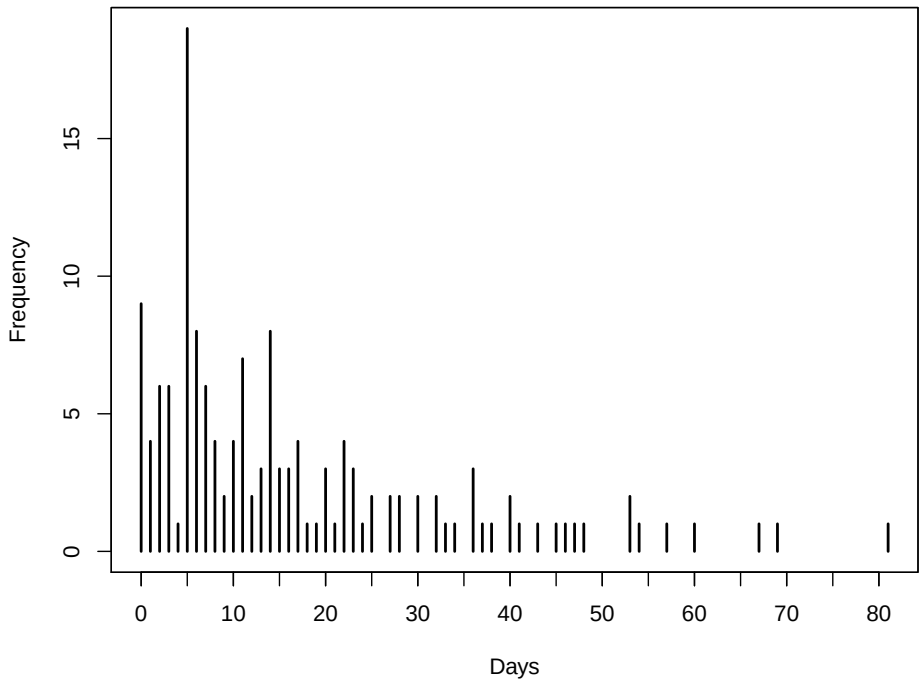


Figure 2: Frequency distribution for number of days absent from school.

3. Illustrations

For a simple illustration of basic Poisson and NB count regression the `quine` data from the **MASS** package is used. This provides the number of `Days` that children were absent from school in Australia in a particular year, along with several covariates that can be employed as regressors. The data can be loaded by

```
R> data("quine", package = "MASS")
```

and a basic frequency distribution of the response variable is displayed in Figure 2.

> For code input and output, the style files provide dedicated environments. Either the "agnostic" `{CodeInput}` and `{CodeOutput}` can be used or, equivalently, the environments `{Sinput}` and `{Soutput}` as produced by `Sweave()` or **knitr** when using the `render_sweave()` hook. Please make sure that all code is properly spaced, e.g., using `y = a + b * x` and *not* `y=a+b*x`. Moreover, code input should use "the usual" command prompt in the respective software system. For R code, the prompt `"R> "` should be used with `"+ "` as the continuation prompt. Generally, comments within the code chunks should be avoided – and made in the regular LaTeX text instead. Finally, empty lines before and after code input/output should be avoided (see above).

As a first model for the `quine` data, we fit the basic Poisson regression model. (Note that JSS prefers when the second line of code is indented by two spaces.)

```
R> m_pois <- glm(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
+    family = poisson)
```

To account for potential overdispersion we also consider a negative binomial GLM.

```
R> library("MASS")
R> m_nbin <- glm.nb(Days ~ (Eth + Sex + Age + Lrn)^2, data = quine)
```

In a comparison with the BIC the latter model is clearly preferred.

```
R> BIC(m_pois, m_nbin)

       df      BIC
m_pois 18 2046.851
m_nbin 19 1157.235
```

Hence, the full summary of that model is shown below.

```
R> summary(m_nbin)

Call:
glm.nb(formula = Days ~ (Eth + Sex + Age + Lrn)^2, data = quine,
    init.theta = 1.60364105, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0857  -0.8306  -0.2620   0.4282   2.0898

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00155    0.33709   8.904  < 2e-16 ***
EthN        -0.24591    0.39135  -0.628  0.52977
SexM        -0.77181    0.38021  -2.030  0.04236 *
AgeF1       -0.02546    0.41615  -0.061  0.95121
AgeF2       -0.54884    0.54393  -1.009  0.31296
AgeF3       -0.25735    0.40558  -0.635  0.52574
LrnSL        0.38919    0.48421   0.804  0.42153
EthN:SexM    0.36240    0.29430   1.231  0.21818
EthN:AgeF1  -0.70000    0.43646  -1.604  0.10876
EthN:AgeF2  -1.23283    0.42962  -2.870  0.00411 **
EthN:AgeF3   0.04721    0.44883   0.105  0.91622
EthN:LrnSL   0.06847    0.34040   0.201  0.84059
SexM:AgeF1   0.02257    0.47360   0.048  0.96198
SexM:AgeF2   1.55330    0.51325   3.026  0.00247 **
SexM:AgeF3   1.25227    0.45539   2.750  0.00596 **
SexM:LrnSL   0.07187    0.40805   0.176  0.86019
AgeF1:LrnSL -0.43101    0.47948  -0.899  0.36870
AgeF2:LrnSL  0.52074    0.48567   1.072  0.28363
AgeF3:LrnSL       NA         NA      NA       NA
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.6036) family taken to be 1)

    Null deviance: 235.23  on 145  degrees of freedom
Residual deviance: 167.53  on 128  degrees of freedom
AIC: 1100.5

Number of Fisher Scoring iterations: 1


              Theta:  1.604
          Std. Err.:  0.214

 2 x log-likelihood:  -1062.546
```

4. Summary and discussion

- As usual ...

Computational details

If necessary or useful, information about certain computational details such as version numbers, operating systems, or compilers could be included in an unnumbered section. Also, auxiliary packages (say, for visualizations, maps, tables, ...) that are not cited in the main text can be credited here.

The results in this paper were obtained using R 4.0.3 with the **MASS** 7.3.53 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

All acknowledgments (note the AE spelling) should be collected in this unnumbered section before the references. It may contain the usual information about funding and feedback from colleagues/reviewers/etc. Furthermore, information such as relative contributions of the authors may be added here (if any).

References

Acharya SS, Fendler W, Watson J, Hamilton A, Pan Y, Gaudiano E, Moskwa P, Bhanja P, Saha S, Guha C, Parmar K, Chowdhury D (2015). "Serum microRNAs are early indicators of survival after radiation-induced hematopoietic injury." *Science Translational Medicine*, **7**(287). ISSN 19466242. doi:10.1126/scitranslmed.aaa6593. URL https://pubmed.ncbi.nlm.nih.gov/25972001/.

Dinh TKT, Fendler W, Chałubińska-Fendler J, Acharya SS, O'Leary C, Deraska PV, D'Andrea AD, Chowdhury D, Kozono D (2016). "Circulating miR-29a and miR-150 correlate with delivered dose during thoracic radiation therapy for non-small cell lung cancer." *Radiation Oncology*, **11**(1). ISSN 1748717X. doi:10.1186/s13014-016-0636-4.

Dobbin KK, Cesano A, Alvarez J, Hawtin R, Janetzki S, Kirsch I, Masucci GV, Robbins PB, Selvan SR, Streicher HZ, Zhang J, Butterfield LH, Thurin M (2016). "Validation of biomarkers to predict response to immunotherapy in cancer: Volume II - clinical validation and regulatory considerations." *Journal for ImmunoTherapy of Cancer*, **4**(1), 77. ISSN 20511426. doi:10.1186/s40425-016-0179-0. URL https://jitc.bmj.com/lookup/doi/10.1186/s40425-016-0179-0.

Elias KM, Fendler W, Stawiski K, Fiascone SJ, Vitonis AF, Berkowitz RS, Frendl G, Konstantinopoulos P, Crum CP, Kedzierska M, Cramer DW, Chowdhury D (2017). "Diagnostic potential for a serum miRNA neural network for detection of ovarian cancer." *eLife*, **6**. ISSN 2050084X. doi:10.7554/eLife.28932.

Fendler W, Malachowska B, Meghani K, Konstantinopoulos PA, Guha C, Singh VK, Chowdhury D (2017). "Evolutionarily conserved serum microRNAs predict radiation-induced fatality in nonhuman primates." *Science Translational Medicine*, **9**(379). ISSN 19466242. doi:10.1126/scitranslmed.aal2408. URL https://pubmed.ncbi.nlm.nih.gov/28251902/.

Goossens N, Nakagawa S, Sun X, Hoshida Y (2015). "Cancer biomarker discovery and validation." doi:10.3978/j.issn.2218-676X.2015.06.04. URL /pmc/articles/PMC4511498/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4511498/.

Małachowska B, Tomasik B, Stawiski K, Kulkarni S, Guha C, Chowdhury D, Fendler W (2020). "Circulating microRNAs as Biomarkers of Radiation Exposure: A Systematic Review and Meta-Analysis." *International Journal of Radiation Oncology Biology Physics*, **106**(2), 390–402. ISSN 1879355X. doi:10.1016/j.ijrobp.2019.10.028.

Pagacz K, Kucharski P, Smyczynska U, Grabia S, Chowdhury D, Fendler W (2020). "A systemic approach to screening high-throughput RT-qPCR data for a suitable set of reference circulating miRNAs." *BMC Genomics*, **21**(1). ISSN 14712164. doi:10.1186/s12864-020-6530-3. URL https://pubmed.ncbi.nlm.nih.gov/32005151/.

Smialowski P, Frishman D, Kramer S (2009). "Pitfalls of supervised feature selection." doi:10.1093/bioinformatics/btp621. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815655/.

Xue J, Jia E, Ren N, Lindsay A, Yu H (2019). "Circulating microRNAs as promising diagnostic biomarkers for pancreatic cancer: A systematic review." *OncoTargets and Therapy*, **12**, 6665–6684. ISSN 11786930. doi:10.2147/OTT.S207963.

URL /pmc/articles/PMC6707936/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6707936/.

A. More technical details

Appendices can be included after the bibliography (with a page break). Each section within the appendix should have a proper section title (rather than just *Appendix*).

For more technical style details, please check out JSS's style FAQ at https://www.jstatsoft.org/pages/view/style#frequently-asked-questions which includes the following topics:

- Title vs. sentence case.
- Graphics formatting.
- Naming conventions.
- Turning JSS manuscripts into R package vignettes.
- Trouble shooting.
- Many other potentially helpful details...

B. Using BibTeX

References need to be provided in a BibTeX file (`.bib`). All references should be made with `\cite`, `\citet`, `\citep`, `\citealp` etc. (and never hard-coded). This commands yield different formats of author-year citations and allow to include additional details (e.g., pages, chapters, ...) in brackets. In case you are not familiar with these commands see the JSS style FAQ for details.

Cleaning up BibTeX files is a somewhat tedious task – especially when acquiring the entries automatically from mixed online sources. However, it is important that informations are complete and presented in a consistent style to avoid confusions. JSS requires the following format.

- JSS-specific markup (`\proglang`, `\pkg`, `\code`) should be used in the references.
- Titles should be in title case.
- Journal titles should not be abbreviated and in title case.
- DOIs should be included where available.
- Software should be properly cited as well. For R packages `citation("pkgname")` typically provides a good starting point.

Affiliation:

Konrad Stawiski, M.D.
Department of Biostatistics and Translational Medicine
Medical University of Lodz
Mazowiecka 15
92-215 Lodz, Poland
E-mail: konrad.stawiski@umed.lodz.pl, konrad@konsta.com.pl
URL: https://biostat.umed.pl/person/?surname=stawiski, https://konsta.com.pl

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd