Proceedings of the 16th Conference on Computer Science and Intelligence Systems pp. 301–311

DOI: 10.15439/2021F89
ISSN 2300-5963 ACSIS, Vol. 25

A random forest-based approach for survival curves comparison: principles, computational aspects, and asymptotic time complexity analysis

Lubomír Štěpánek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
lubomir.stepanek@vse.cz
&
Institute of Biophysics and Informatics
First Faculty of Medicine
Charles University
Salmovská 1, Prague, Czech Republic
lubomir.stepanek@lf1.cuni.cz

Filip Habarta
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
filip.habarta@vse.cz

Ivana Malá
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
malai@vse.cz

Luboš Marek
Department of Statistics and Probability
Faculty of Informatics and Statistics
University of Economics
nám. W. Churchilla 4, 130 67 Prague, Czech Republic
marek@vse.cz

***Abstract*—The log-rank test and Cox's proportional hazard model can be used to compare survival curves but are limited by strict statistical assumptions. In this study, we introduce a novel, assumption-free method based on a random forest algorithm able to compare two or more survival curves. A proportion of the random forest's trees with sufficient complexity is close to the test's p-value estimate. The pruning of trees in the model modifies trees' complexity and, thus, both the method's robustness and statistical power. The discussed results are confirmed using a simulation study, varying the survival curves and the tree pruning level.**

I. Introduction

COMPARING two or more survival curves is relatively common in many applied areas such as biomedicine, econometrics, management, and others. When the curves are statistically significantly different, it may help treat the groups that are the curves built by in appropriate (separate) ways.

As typical for survival analysis, the variable of our interest usually describes a (time) development of proportions of individuals who have not experienced the event of interest yet (until each considered time point) in each consecutive time point of the time period of our interest.

Such a time development is commonly plotted using orthogonal polygonal lines, also known as *survival curves* in a two-dimensional (survival) plot, sometimes called Kaplan-Meier plot [1]. Since groups of individuals that are about to be compared have their own time developments of non-experiencing the event of interest, one Kaplan-Meier survival plot may include more than only one curve, as shown in Fig. 1.

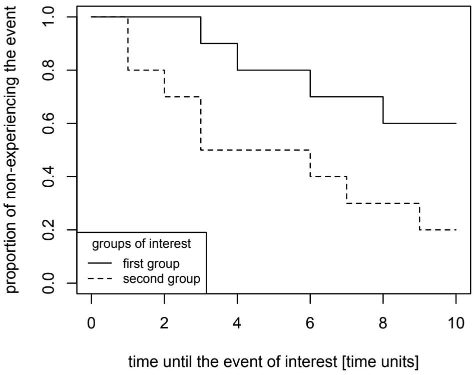

Fig. 1. Two time-to-event survival curves for two groups of interest in Kaplan-Meier (survival) plot.

Typically, following the logic behind the time development of non-experiencing the event of our interest, there are time points placed on the horizontal axis of a survival plot and proportions of subjects with no experience of the event of interest on a vertical axis.

Regardless of the total length of the time period of interest, it has in principle be finite and, consequently, we cannot get any piece of information whether the individuals not experiencing the event of interest in the time period would register the event after an end of the period, or not. That is why the time-to-event (survival) data are also called *right censored* data.

Since the experiencing of the event of interest is usually irreversible in time within the scope of classical survival analysis (the event is, e. g., a death, diagnosis, bankruptcy, failure, etc.), a subject registering the event whenever in the time period of the interest, continues to stay in this state till the end of the referenced time (the time of the right censoring). Thus, the survival curves are monotonous and nonincreasing. Intuitively, when the development of the event of interest differs between two groups, we may expect their survival curves are hardly similar and relatively far from each other within each considered time point.

If there are two survival curves to be statistically compared, the log-rank test as a tool of choice is usually performed [2] and commonly implemented, e. g. in R language and statistical environment [3] and its library `survival` [4]. However, a usage of the log-rank test is limited by statistical assumptions, that (i) censoring should not affect anyhow the observed events and (ii) the censoring should occur equally or at least near-equally in both compared groups, generating the survival curves. Also, (iii) the group's sizes are expected to be large enough, enabling the log-rank test's χ^2 statistics to fulfill its asymptotic properties.

To overcome the limitations done by the statistical assumptions of the original log-rank test or to increase its robustness or statistical power, several modifications of the traditional log-rank test were published. The first approach is to modify the hazard functions slightly, i. e., functions of rates of events based on fixed proportions of the events in the past, and relax their assumptions to increase their robustness as suggested by [5]. Then, another option is to introduce new covariates (variables) to enrich the model comparing two survival curves and increase the robustness, published in [6]. Also, employing various weighting schemes for individual observations, usually growing significance for earlier ones, may increase the statistical power of the test as investigated originally in [7] and then improved in [8], [9] and [10]. Finally, robust combinatorial and exact calculations of all possible combinations of the event experiencing and non-experiencing subjects for given total numbers of subjects in the compared groups are researched in [11] and [12] and using survival curves' finite combinatorial geometry in [13].

An advantage of the latter approach, based on robust combinatorial computations when comparing two survival curves, is that makes possible an estimation of asymptotic time complexity, as is in details commented by [14], [15], [16] and partly by [13], too.

When one wants to statistically compare three or more survival curves, there is an option to use Cox's proportional hazard model or a score-rank test based on Cox's model. Unfortunately, Cox's proportional model is also limited – it assumes that hazard proportions for each group are constant across all considered time points, which is often not met in practice. Some more robust versions of Cox's model were derived to minimize the violation of the constant hazards' ratio by real-world data, e. g. based on an idea of stratification of each group into subgroups according to their hazard similarity; however, those advanced models are usually limited by other, more complex assumptions [17], though.

The decision (or regression) trees and random forests are classical algorithms used for classification or regression problems. An idea to apply decision trees and random forest on survival tasks and right-censored, time-to-event data originate from [18], but initial thoughts rather aimed to a robust estimation of hazard functions' parameters, e. g. Nelson-Aalen estimator etc. The decision trees and random forests are naturally assumption-free robust, and fully non-parametric, especially in comparison to the log-rank test or the Cox's proportional hazard model, which is a property also utilized in this study and by the proposed alternative method for survival curves comparison.

The proceeding proceeds as follows. Firstly, in the section *Traditional methods for survival curves comparing and random forests revisited*, we shortly remind the fundamental principles of the log-rank test, Cox's proportional hazard model, decision trees, and random forests. We also discuss assumptions and limitations of the named methods that create room for new approaches that are less dependent on statistical assumptions.

Then, in the section *The proposed method for survival curves comparing*, we introduce a novel alternative for two or more survival curves comparing, based on random forest-based generating of multiple decision trees, using variables derived from original time-to-event data of compared groups of individuals in their nodes. The level of the trees' pruning is adjustable as a hyperparameter; it enables to control a complexity of the trees, i. e., an average number of nodes and leaves per tree in the forest. If a given tree in the random forest is able to classify whatever new observation in each of the groups (described by its survival curve), i. e., there exists at least one leaf node for each group assigning the observation to such a group, that tends to be contradictory the null hypothesis, claiming there are no statistical differences between the groups (and their survival curves). A proportion of the trees with sufficient complexity to all trees in the forest serves by definition as an estimate of p-value as would be analogously[1] returned by the traditional log-rank test, i. e., a conditional probability of collecting data as extreme or even more given there is no difference between the survival curves. Since the p-value is partially determined by the proportion of sufficiently complex trees to all trees of the random forest, the level of pruning may affect the robustness or statistical

[1] A numerical value of the p-value returned by the log-rank test and by the proposed method are not supposed to be equal, as discussed in the following sections.

power of the proposed random forest-based inference test, as is discussed more in details later.

The asymptotic time complexity of the p-value estimation, assuming the random forest model building, is then derived, and, finally, in the section *Simulation study*, a preliminary simulation study is performed to confirm the theoretically derived properties of the method. Besides others, the introduced approach offers a way how to compare more than two survival curves without any assumptions needed to be met.

II. Traditional methods for survival curves comparing and random forests revisited

Firstly, we remind principles and assumptions of the log-rank test and Cox's proportional hazard model that facilitates a better understanding of their limitations, which, consequently, opens room for improvements in survival curves comparing. We also recapitulate the logic of the decision trees and random forest heavily equipped in the proposed method for survival curves comparison.

A. Principles, assumptions, and limitations of the log-rank test

Principles of the log-rank test. Let's assume k distinct time points where the event of interest could take a place; the j-th time point is marked as t_j, where $j \in \{1, 2, 3, \dots, k\}$, and all the time points are ordered in a tuple $(t_1, t_2, \dots, t_k)^T$. Also, let's suppose there are two groups of subjects, marked by subscripts 1 and 2, respectively. For each of the time points, let's say for the j-th one (t_j) there are $r_{1,j}$ and $r_{2,j}$ individuals at risk (they have not experienced the event of interest yet or have been censored) in the group 1 and group 2, respectively, and $d_{1,j}$ and $d_{2,j}$ individuals who experienced the event in the group 1 and group 2, respectively. Thus, following the previous logic, we can construct a (contingency) table I.

TABLE I
NUMBERS OF INDIVIDUALS EXPERIENCING THE EVENTS OF INTEREST IN BOTH GROUPS (1 AND 2) AT TIME t_j.

	event of interest at the event time t_j		
group	yes	no	total
1	$d_{1,j}$	$r_{1,j} - d_{1,j}$	$r_{1,j}$
2	$d_{2,j}$	$r_{2,j} - d_{2,j}$	$r_{2,j}$
total	d_j	$r_j - d_j$	r_j

The log-rank test checks the null hypothesis H_0 that both groups experienced identical rates of the events of interest in time (also called *hazard functions*) [2], conditional on fixed rates in the past are the same. Under the null hypothesis H_0, the observed numbers of individuals experiencing the events could be considered as random variables $D_{1,j}$ and $D_{2,j}$ following a hypergeometric distribution with parameters $(r_j, r_{i,j}, d_j)$ for both $i \in \{1, 2\}$. Thus, the expected value of the variable $D_{i,j}$ is $\mathbb{E}(D_{i,j}) = r_{i,j}\frac{d_j}{r_j}$ and variance is $\mathrm{var}(D_{i,j}) = \frac{r_{1,j}r_{2,j}d_j}{r_j^2}\left(\frac{r_j-d_j}{r_j-1}\right)$ for both $i \in \{1, 2\}$. Finally, under the null hypothesis H_0, we can compare the observed numbers of events of interest, $d_(i, j)$, for all $j \in \{1, 2, 3, \dots, k\}$, to their expected values $\mathbb{E}(D_{i,j}) = r_{i,j}\frac{d_j}{r_j}$. So, the test statistic for both $i \in \{1, 2\}$ is then

$$\chi^2_{\text{log-rank}} = \frac{\left(\sum_{j=1}^{k} d_{i,j} - \mathbb{E}(D_{i,j})\right)^2}{\sum_{j=1}^{k} \mathrm{var}(D_{i,j})} = = \frac{\left(\sum_{j=1}^{k} d_{i,j} - r_{i,j}\frac{d_j}{r_j}\right)^2}{\sum_{j=1}^{k} \frac{r_{1,j}r_{2,j}d_j}{r_j^2}\left(\frac{r_j-d_j}{r_j-1}\right)}, \tag{1}$$

which follows under H_0 a χ^2 distribution with 1 degree of freedom, $\chi^2_{\text{log-rank}} \sim \chi^2(1)$. For feasible large r_j, i. e. at least $r_j \geq 30$, a square root of $\chi^2_{\text{log-rank}}$ follows a standard normal distribution, $\sqrt{\chi^2_{\text{log-rank}}} \sim \mathcal{N}(0, 1^2)$. Since $\chi^2_{\text{log-rank}} \sim \chi^2(1)$, the statistics $\chi^2_{\text{log-rank}}$ can be uniquely transformed into p-value, which stands for a conditional probability of obtaining the test statistics $\chi^2_{\text{log-rank}}$ at least as extreme as the statistics actually observed, under the assumption that the null hypothesis H_0 reflects the reality.

Assumptions and limitations of the log-rank test. The right censoring of the data should not affect the occurrences of the event of interest in both groups anyhow. Also, the proportions of censored observations are supposed to be of (nearly) equal size in both groups. Otherwise, the test statistic $\chi^2_{\text{log-rank}}$ calculated using (1) could be biased for $i = 1$, or for $i = 2$.

Then, putting together the equation (1), so the test statistic $\chi^2_{\text{log-rank}}$ follows a χ^2 distribution, and the table I, both the initial total number of individuals r_0 at risk and initial number $r_0 - d_0$ not experiencing the event, should be large enough. Otherwise, so-called Cochrane criteria for minimal sample size for χ^2 tests are not met and the $\chi^2_{\text{log-rank}}$ statistics could not fulfill the χ^2 asymptotic properties; or, analogously, both the numerator and denominator of the statistics (1) are relatively small and an estimate of the $\chi^2_{\text{log-rank}}$ statistics is numerically unstable.

All the named issues may decrease the robustness or statistical power of the log-rank test.

Furthermore, by investigating the denominator of the equation (1), we can easily realize the test statistic $\chi^2_{\text{log-rank}}$ is the highest when the denominator $\sum_{j=1}^{k} \mathrm{var}(D_{i,j})$ is as low as possible given the values $d_{i,j}$ and $r_{i,j}$ for all $i \in \{1, 2\}$ and $j \in \{1, 2, 3, \dots, k\}$. This holds just when the proportions $\frac{r_{1,j}}{r_j} = \frac{r_{1,j}}{r_{1,j}+r_{2,j}}$ and $\frac{r_{2,j}}{r_j} = \frac{r_{2,j}}{r_{1,j}+r_{2,j}}$ are both constant (and mutually different enough) across all the time points $(t_1, t_2, \dots, t_k)^T$, and then the log-rank test is the most statistically powerful, i. e. its ability to reject the null hypothesis H_0 claiming the survival curves are equivalent, when they are in fact different, is maximal possible. That is common issue decreasing the test power – the mentioned proportions are typically not constant when a "trend" of the survival curves change a lot, when the curves change their mutual distance or when they even cross themselves one or more times.

B. Principles, assumptions, and limitations of Cox proportional hazard model

Principles of Cox proportional hazard model. The Cox proportional hazard model is frequently used to model relationships between the hazard function of the event of interest, defined as a probability that a subject experiences the event of interest in a small time interval, given that the individual survived up to the beginning of the interval, and explanatory variables. If one of the explanatory variables is categorical, thus dividing an entire sample into two or more groups, then the Cox proportional hazard model could serve as a method for statistical comparing of more than two groups and their survival curves. The hazard function $h(t)$ depending on explanatory variables as suggested by Cox [19], follows for individual i form

$$\log h(t) = \log h_0(t) + \boldsymbol{\beta}^T \boldsymbol{x_i}, \qquad (2)$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \ldots)^T$ is a vector of estimated linear coefficients to explanatory variables and $\boldsymbol{x_i} = (1, x_{i,1}, x_{i,2}, \ldots)^T$ is a vector of values of the explanatory variables for group i. The formula (2) could be after exponentiation rewritten also as

$$h(t) = h_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{x_i}},$$

by which we can see for two groups 1 and 2 that

$$\frac{h(t \mid x_1)}{h(t \mid x_2)} = \frac{h_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{x_1}}}{h_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{x_2}}} = \frac{e^{\boldsymbol{\beta}^T \boldsymbol{x_1}}}{e^{\boldsymbol{\beta}^T \boldsymbol{x_2}}},$$

thus, the hazard ratio for any two groups 1 and 2 is forced to be constant, considering the model (2) and a fact that once estimated coefficients $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots)^T$ and input data $\boldsymbol{x_i} = (1, x_{i,1}, x_{i,2}, \ldots)^T$ are given, therefore constant. The parameters in the Cox model (2) can be estimated by a partial likelihood [20].

When exists $j \in \{1, 2, 3, \ldots\}$ so that β_j is a linear coefficient of a categorical variable classifying observations into two or more groups (with their survival curves), then one can consider the Cox approach as an alternative for the log-rank test with the exception there are more than two survival curves to be compared. Wald t-tests indicate significant statistical differences between the categorical variable levels, thus also in groups' survival curves.

Assumptions and limitations of Cox proportional hazard model. However, while Cox's regression is widely used for event prediction in survival analysis or for comparing more than two survival curves, it has rigid statistical limitations [21]. Particularly, Cox's model assumes that ratios of hazards for any two subjects (individuals or groups) are constant across all time points; that is why the model is called "proportional hazard". However, real survival data often violate this assumption. For instance, supposing two survival curves for two groups as in Fig. 2, such that the curves cross each other, their hazards could not be proportionally constant. Even more, when one of the curves drops to zero while the other levels off similarly to Fig. 3, also, the ratio of the hazard functions could not be constant.

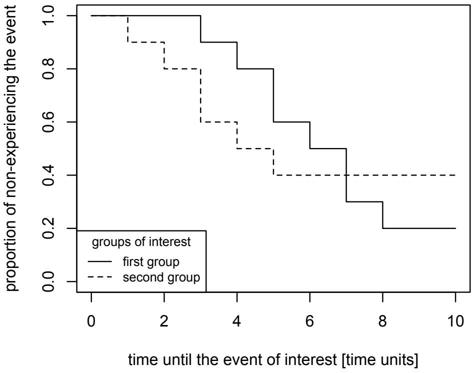

Fig. 2. An example of a pair of survival curves crossing each other.

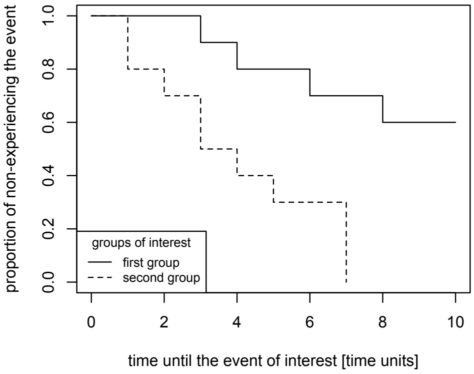

Fig. 3. An example of a pair of survival curves so that one drops to zero while the other levels off.

C. Principles of the random forests

Before we conclude up basic principles of the random forests' algorithm, we remind fundamental pieces of knowledge about decision trees that build up a random forest model.

Principles of the decision trees. The decision trees (also called classification trees) that belong to the CART family of trees (classification and regression trees) are sets of rules that partition the hyperspace of all explanatory variables into disjunctive hyper-rectangles and fit simple (constant) models there, each time minimizing a given criterion [22].

More specifically, the decision trees classify an observation depicted by a vector of values $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \ldots, x_{i,k})^T$ for k explanatory variables into one of m target classes, i. e. classes of a response categorical variable, where $[k, m] \in \mathbb{N}^2$.

The logic behind a tree induction is described by the flowchart in Fig. 4. Initially, one root node is set, and the tree induction algorithm searches for a node decision rule, i. e. such an explanatory variable and a logical formula containing the explanatory (splitting) variable and its relationship to some constant or subset that minimizes a given criterion. When the optimal node rule is found, the node rule enables to split (binary partition) the dataset into two parts following the logic of the slitting variable and splitting point (the first part contains values larger than or equal to the splitting point, the other contains the rest of dataset). Two new child nodes for the corresponding two parts of the dataset are added to the growing tree. The procedure of searching a node rule, i. e. a splitting variable and splitting point, is repeated for each fresh added (child) node until the part of the dataset that is logically constrained by a set of decision rules coming from the root node till the last (leaf) one, includes observations of only one target class. This strategy of the tree growing is called a top-down induction of a decision tree (TDIDT).

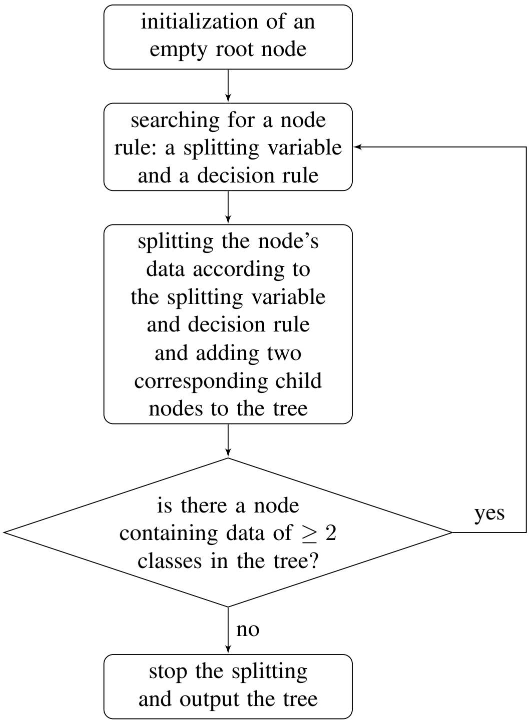
initialization of an empty root node
searching for a node rule: a splitting variable and a decision rule
splitting the node's data according to the splitting variable and decision rule and adding two corresponding child nodes to the tree
is there a node containing data of ≥ 2 classes in the tree?
yes
no
stop the splitting and output the tree

Fig. 4. A top-down induction of a decision tree (TDIDT).

Let $\sigma(\bullet)_j$ be a proportion of a target class j in all observations constrained by rules coming from the root till the node n_t. If the node n_t is a leaf one, it classifies into the class j^* so that $j^* = \operatorname{argmax}_{j\in\{1,2,\dots,m\}} \{\sigma(\bullet)_j\}$.

The given criterion minimized in searching for node n_t rule is an *impurity measure*, $Q_{n_t}(T)$, such as misclassification error

$$Q_{n_t}(T) = 1 - \sigma(\bullet)_j,$$

or Gini index

$$Q_{n_t}(T) = \sum_{j=1}^{m} \sigma(\bullet)_j (1 - \sigma(\bullet)_j),$$

or deviance (cross-entropy)

$$Q_{n_t}(T) = -\sum_{j=1}^{m} \sigma(\bullet)_j \cdot \log \sigma(\bullet)_j.$$

We can easily see that the higher the $\sigma(\bullet)_j$ as a proportion of a target class j in the node n_t is, the lower whatever kind of the named impurity measures is, as expected.

Following the logic of the top-down induction of a decision tree depicted in Fig. 4, a final tree cannot have lower than maximal possible complexity; even a leaf node including only two observations of two different target classes is once more split into two child leaf nodes. To overcome this issue, *overfitting*, besides some naive approaches like a fixed maximal number of nodes per a tree, etc., a procedure called *pruning* is frequently applied. The pruning is based on numerical estimating of the statistics *cost–complexity function* following the form

$$C_\kappa(T) = \sum_{n_t\in\{\boldsymbol{n_t}\}} |\{\boldsymbol{x}_{n_t}\}| \cdot Q_{n_t}(T) + \kappa \cdot |\{\boldsymbol{n_t}\}|, \qquad (3)$$

where $\{\boldsymbol{n_t}\}$ is a set of leaf nodes of the tree and $\{\boldsymbol{x}_{n_t}\}$ is a set of all observations constrained by rules coming from the root till the node n_t. The idea is to find a subtree T_κ so that $T_\kappa \subset T$ for a given κ that minimizes the statistics $C_\kappa(T)$, i. e. $T_\kappa = \operatorname{argmin}_T \left\{ \sum_{n_t\in\{\boldsymbol{n_t}\}} |\{\boldsymbol{x}_{n_t}\}| \cdot Q_{n_t}(T) + \kappa \cdot |\{\boldsymbol{n_t}\}| \right\}$. The $\kappa \geq 0$ is a hyperparameter (a tuning parameter) and governs the trade-off between a tree complexity or size (low values of κ) and goodness of fit to the data (large values of κ).

Principles of the random forests. Once we can generate classification trees as described above, construction of a random forest is relatively easy. Random forests are finite sets of (distinct) decision trees so that each tree classifies an observation depicted by a vector of values $\boldsymbol{x_i} = (x_{i,1}, x_{i,2}, \dots, x_{i,k})^T$ for k explanatory variables into one of $m < \infty$ target classes [18]. The eventual classification into the final class is done using a voting scheme – the final class $j^* \in \{1, 2, \dots, m\}$ is the one that a subset of the random forest's trees classifying just into the class j^* is the largest one among all subsets of the random forest's trees. More technically, $j^* = \operatorname{argmax}_{j\in\{1,2,\dots,m\}} \{\#\text{ of trees classifying into the class } j\}$. In case of a tie, i. e. there are two or more target classes the forest's trees would classify with maximum frequency into, one of them is picked randomly.

A bit different in the random forest's tree induction is the fact that only $k^* < k$ variables are considered as possible splitting variables in each searching for the node rule. Instead, the subset of k^* variables of the original set of all k explanatory variables is selected randomly using bootstrapping to ensure the pre-selected k^* variables are as much uncorrelated as possible. Other details of the trees inductions are the same as described above. A flowchart of the random forest model building is in Fig. 5.

Assumptions and limitations of the trees and forests. There are neither other technical assumptions nor limitations of the random forests usage worth to be discussed.

III. The proposed method for survival curves comparing

We introduce the novel method for statistical comparison of two or more time-to-event developments of individuals' groups, depicted by their survival curves.

Firstly, data that are on the input of the method have to be transformed. Each individual is originally described

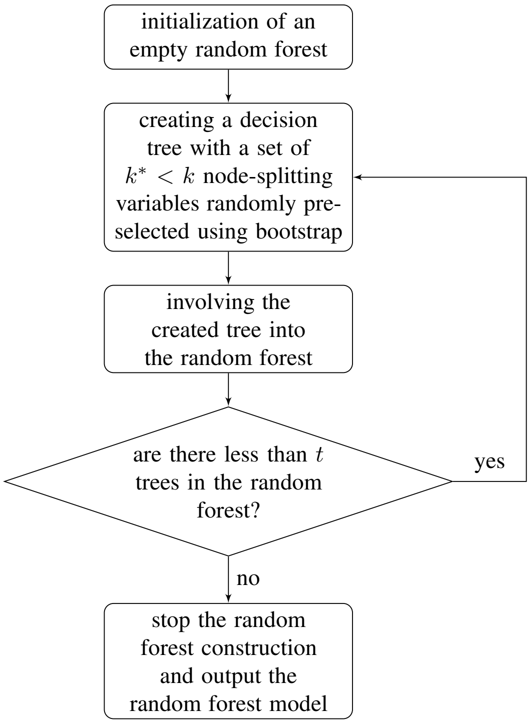
initialization of an empty random forest
creating a decision tree with a set of $k^* < k$ node-splitting variables randomly pre-selected using bootstrap
involving the created tree into the random forest
are there less than t trees in the random forest?
yes
no
stop the random forest construction and output the random forest model

Fig. 5. A construction of the random forest model involving t decision trees.

using their group affiliation, a time to event of interest (or to censoring), and whether they experienced the event of interest (or have been censored). Then, using the original data, for each individual, a sequence of weighted point estimates of probabilities that they did not experience the event of interest in a given time point and the group affiliation is created. That enables introducing new variables (their number is equal to the number or all considered time points) that are used as splitting variables in tree inductions when a random forest model is built.

Once the data are transformed, a random forest model is constructed. Each tree of the random forest either can classify into two or more classes that are represented as the group affiliations, or cannot to classify into any classes at all (then it is necessarily a root node tree), based on its complexity (size). See also Fig. 6.

The more trees of sufficient complexity able to classify into the classes (equal to the groups of individuals, described by their survival curves melted into the transformed variables as mentioned above) are in the forest, the more likely we can reject the null hypothesis that there is no difference between the survival curves (or the groups of individuals' time-to-event development). Thus, a proportion of trees that classify into all the classes, to all the trees of the random forest is very close to a point estimate of the p-value, i. e. the probability we incorrectly reject the null hypothesis of no difference between the survival curves, assuming the null hypothesis is true. Thus, the p-value is a probability of a wrong decision and should be as low as possible whenever we consider rejecting the null hypothesis. The first type error rate, i. e. the incorrect rejection of the null hypothesis when it is true, can be controlled by setting the parameter κ of the random forest's tree complexity (or the tree pruning).

So, the proposed method fulfills all feasible demands on inference testing. We also discuss some of the method's properties, particularly its asymptotic time complexity. The first type error rate is simulated in the simulation study with varying κ tuning parameters. The introduced method is able to compare more than two survival curves, and since it utilizes a random forest tree-based algorithm, it is practically assumption-free. This is where it surpasses both the log-rank test and Cox's regression.

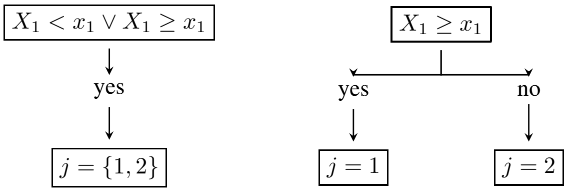

Fig. 6. An example of a root node tree (on the left) not capable to classify into any class unambiguously, and an example of a tree with "sufficient" complexity (on the right) able to classify into two classes ($j = 1$ and $j = 2$).

A. Data transformation and preparation for random forest model building

Initial time-to-event survival data includes n observations; for each of them, we have a piece of information about the time to the event of interest (or to the censoring) and whether the event of interest or the censoring occurred. By adopting the mathematical notation from the section about the log-rank test, for each considered time point t_j, where $j \in \{1, 2, \dots, k\}$ and $k \in \mathbb{N}$, we can calculate for the group i, where $i \in \{1, 2, \dots, m\}$, a proportion $r_{i,j}$ of individuals that are at risk (of the event of interest or the censoring) in the j-th time point. Similarly, one can estimate a for the group i, where $i \in \{1, 2, \dots, m\}$, a proportion $d_{i,j}$ of individuals who experienced the event of interest (or the censoring) in the j-th time point. Putting those estimates together, for the group i, where $i \in \{1, 2, \dots, m\}$, we can make a point estimate of a probability $\hat{p}_{i,j}$ that an individual from the group would not experience the event of interest (or the censoring) in the j-th time point, so

$$\hat{p}_{i,j} = 1 - \frac{d_{i,j}}{r_{i,j}}. \tag{4}$$

Such an estimate is made k-times for all time points $\{t_1, t_2, \dots, t_k\}$, by getting $(\hat{p}_{i,1}, \hat{p}_{i,2}, \dots, \hat{p}_{i,k})$, but such a vector of values is common for all individuals of the group i. However, it could be personalized using an operator $\delta_{\nu,j}$ for ν-th individual, where $\nu \in \{1, 2, \dots, n\}$, following the form

$$\delta_{\nu,j} = \begin{cases} 1, & \nu\text{-th individual did not experience the event} \\ & \text{of interest in } j\text{-th time point} \\ 0, & \nu\text{-th individual experienced the event} \\ & \text{of interest in } j\text{-th time point,} \end{cases}$$

assuming that ν-th individual belongs to the group i. So by modifying the formula (4) using the operator $\delta_{\nu,j}$ we get

$$\delta_{\nu,j}\hat{p}_{i,j} = \delta_{\nu,j}\left(1 - \frac{d_{i,j}}{r_{i,j}}\right). \tag{5}$$

The logic of the formula (5) enables to get mutually distinct vectors of values $(\delta_{\nu,1}\hat{p}_{i,1}, \delta_{\nu,2}\hat{p}_{i,2}, \ldots, \delta_{\nu,k}\hat{p}_{i,k})^T$ for each individual in the group i, which increases natural variability of the data.

Finally, still assuming that ν-th individual belongs to the group i, where $\nu \in \{1, 2, \ldots, n\}$, there are n new vectors $(\delta_{\nu,1}\hat{p}_{i,1}, \delta_{\nu,2}\hat{p}_{i,2}, \ldots, \delta_{\nu,k}\hat{p}_{i,k})^T$ that could be arranged in a matrix of n rows and k columns, which creates a new dataset suitable as an input for the decision tree induction; the k variables could serve as possible splitting variables in the trees' nodes. The j-th variable of the dataset could be interpreted as a personalized point estimate of probability of non-experiencing the event of interest. The $(k+1)$-th variable in the dataset is a target one – categorical variable describing a group affiliation $i \in \{1, 2, \ldots, m\}$ of each observation[2].

B. Construction of the random forest model behind the novel method

The random forest model is built following the algorithms sketched in Fig. 4 and Fig. 5. Variables used as node splitting variables come from the newly created dataset, containing k "explanatory" variables and a target one, as described more in the previous subsection.

Number $t \in \mathbb{N}$ as a count of the trees in the random forest as well as the level of the trees' pruning determined by parameter $\kappa \geq 0$ may vary, as is more explained later.

C. Statistical inference behind the novel method

As already mentioned, the main purpose of the introduced method is to statistically compare two or more survival curves depicting a time-to-event development of distinct groups of individuals. Intuitively, when a large number of the (adequately pruned) trees involved in the random forest model is able to classify into two or more classes, i. e. groups determined by their survival curves, then it is hard to suppose the groups and their survival curves are (statistically) without any difference.

Similarly to the log-rank test or the Cox's regression, let the null hypothesis H_0 claim that there is no statistical difference between the $m > 1$ survival curves[3], and let the alternative hypothesis H_1 claim the contradiction, so

H_0 : *No statistical difference between the m survival curves.*

H_1 : *Statistical difference between the m survival curves.*

Whenever the log-rank test or the Cox's model based on Wald t-test rejects – based on test statistics – the appropriate null hypothesis H_0 in favor of the alternative hypothesis H_1, is equivalent to a situation the test's p-value is lower than or equal to an apriori set level of significance α, usually equal to 0.05. Since the introduced method is in practice assumption-free and non-parametric, the only way to evaluate the statistical inference about the null hypothesis is to estimate the p-value and compare it to the previously set significance level α.

By definition, the p-value is a probability of gaining data evidence at least as extreme as the data evidence actually observed, under the assumption the null hypothesis is true. Let t_c be a number of trees in the random forest that are in contradiction to the null hypothesis under the null hypothesis. The random forest contains exactly t trees. We can easily realize that, given the value for the κ parameter, the value of t_c is equal to the number of all the trees classifying into more than only one class (which is naturally in contradiction to the null hypothesis). Let the $n_c(\tau)$ be a number of classes the tree τ classifies into. Then we can derive

$$t_c = |\{\forall \text{ tree} \in \text{random forest} : n_c(\text{tree}) \geq 2.\}|$$

Then, assuming all trees are inducted randomly regardless of their complexity, the p-value is estimated by $\hat{p}$ so that

$$\begin{aligned}\hat{p} &= P(\text{getting data at least as extreme as the observed} \mid H_0) = \\ &= P(|\{\forall \text{ tree} \in \text{random forest} : n_c(\text{tree}) \geq 2\}| \geq t_c \mid H_0) = \\ &= P(|\{t_c, t_c+1, \ldots, t\}| \mid H_0) = \\ &= \frac{|\{t_c, t_c+1, \ldots, t\}|}{t} = \\ &= \frac{t - t_c + 1}{t} = \\ &= 1 - \frac{t_c - 1}{t}.\end{aligned} \tag{6}$$

Thus, from the formula (6) results that the p-value's estimate is equal to the fraction of $1 - \frac{t_c-1}{t}$. That result is also intuitive. If the initial number t_c of trees in the random forest that are complex enough and classify into two or more classes (and more groups with their survival curves) is in general low, then such a random forest as an entire model is not "so much" in contradiction to the null hypothesis, claiming there are no differences between the classes (and survival curves). Finishing the idea, since the t_c is relatively low, then the fraction $p\text{-value} = 1 - \frac{t_c-1}{t}$ is relatively large, close to 1 and unlikely to be lower than $\alpha(= 0.05)$ which is required for the null hypothesis rejection. On the other hand, when the initial value of t_c is large, i. e. there are many trees in the forest with sufficient complexity classifying into two or more classes (and thus, standing against the null hypothesis), then – because of the large value of t_c – the fraction $p\text{-value} = 1 - \frac{t_c-1}{t}$ is relatively low and likely below the level α. That likely results in the null hypothesis rejection.

The number of trees t in the random forest determines maximum decimal precision of the p-value estimate. When the precision of d decimal digits is required for the p-value estimate, then t has to be $t > 10^d$ or better $t > 10^{d+1}$ to ensure the next-to-last digit (as the d-th decimal digit) is feasibly estimated.

[2] Make a note that across the entire paper, the mathematical notation is consistent – there are m groups depicted by their survival curves, but there are also m target classes of decision trees. Furthermore, there are k time points, and for each of them, a new variable is created within the transformation to the new dataset, thus containing k variables. Finally, the bootstrap behind the random forest model construction also pre-selects $k^* < k$ node variables.

[3] Two or more survival curves; in general $m \in \{2, 3, 4, \ldots\}$ curves.

The κ parameter determines how complex the trees in the random forest would be, i. e. how significant the pruning of the trees should be. Inspecting the formula (3), we can simply realize that if $\kappa = 0$, then there is no cost for large tree complexity and the trees in the random forest are generally very complex (of large size). Then, whenever there are at least two observations in the transformed dataset so the they are assigned to different two groups, all the trees (because of the unlimited complexity) in the forest would classify those observations into their groups (classes), i. e. that for each tree τ is $n_c(\tau) \geq 2$, which results into the equity $t_c = t$ and, thus, p-value estimate of p-value $= 1 - \frac{t_c - 1}{t} = 1 - \frac{t-1}{t} = \frac{1}{t} \approx 0$. If p-value ≈ 0, then also p-value $\approx 0 < \alpha$ which, consequently, tends to rejection of the null hypothesis, very likely a *false* rejection that increases the first error type rate. However, high chance of the null hypothesis rejection means also the high statistical power, i. e. the rejection of the null hypothesis when this is not true.

If $\kappa > 0$, then in general the trees' complexity (size) decreases and also not all of the trees are complex enough to classify into more than one class (the are only root node trees); this means that there are trees τ in the random forest so that $n_c(\tau) \leq 1$, and, finally, $t_c < t$. So, p-value estimate is p-value $= 1 - \frac{t_c - 1}{t} > 0$ and it could be, but also could not be below α.

To conclude this, low values of κ tend to decrease values of p-value and increase the statistical power and the first type error rate, and vice versa. However, the more exact relationship between κ and α could be only roughly estimated using simulations due to the stochastic character of the random forests.

D. A brief asymptotic time complexity analysis and fundamental approaches on the p-value estimation

An atomic unit of the random forest model is a decision tree, inducted following the flowchart 4 and algorithm 1. As long as there is a node containing data of ≥ 2 classes, constrained by all node rules coming from root to the node, the data splitting and growing of the tree continues. If the classes in the data are well balanced as well as the growing tree, the splitting partitions the subdatasets roughly in halves, and the average depth of the tree would be $\log n$ and the time complexity would be $\Theta(\log n)$, assuming one split of a node takes 1 time unit. However, on the other hand, when the classes are not well balanced across the dataset, the splitting cuts the subdatasets into 1 and $n - 1$ observations, which takes n steps in total and the depth of the tree is n. Consequently, the asymptotic time complexity is $\Theta(n)$, assuming one split of a node takes 1 time unit.

Within each node splitting, both for a splitting variable among k variables and through the sample size n is searched, the time complexity $\Theta(\bullet)$ of a decision tree building is somewhere in between being in $\Theta(k \cdot n \cdot \log n)$ (the best-case scenario) and $\Theta(k \cdot n \cdot n)$ (the worst-case scenario), so that

$$\Theta(kn \log n) \leq \Theta(\bullet) \leq \Theta(kn^2).$$

Algorithm 1: The top-down induction of decision trees (TDIDT) following the logic of the flowchart 4

```
Data: a $n \times (k+1)$ dataset with transformed variables
Result: a decision tree
$T = (\{\boldsymbol{n}\})$         // a tree $T$ with a set ;
                  // of nodes $\boldsymbol{n}$;
$\{\boldsymbol{n}\} = \{\mathrm{root}\}$    // initially, the tree $T$
;
                  // is a root;
$\sigma(\bullet)_j$            // a node criterion;
while $\exists$ a node $\in \{\boldsymbol{n}\}$ so that data constrained by all
  node rules coming from root to this node belong to
  $\geq 2$ classes do
    find for the node a splitting variable and splitting
      point minimizing the $\sigma(\bullet)_j$;
    add to the node two child nodes $n_1$ a $n_2$;
    $\{\boldsymbol{n}\} := \{\boldsymbol{n} \cup \{n_1, n_2\}\}$ ;
    $T := (\{\boldsymbol{n}\})$ // update the tree using
      the new node set $\boldsymbol{n}$ ;
end
a completely inducted tree $T$;
```

When a random forest model containing t trees is constructed, the tree induction as introduced above is repeated t times. That being said, the asymptotic time complexity $\Theta(\bullet\bullet)$ of a random forest model building is in between

$$\Theta(tkn \log n) \leq \Theta(\bullet\bullet) \leq \Theta(tkn^2). \tag{7}$$

One model of the random forest provides one (point) estimate of the p-value using the formula (6). In comparison, the estimation of the χ^2 statistics using the formula (1) takes only $\Theta(2k+1)$ time units since is based on a ratio of two summations of k elements. Fortunately, the time complexity (7) is still polynomial. Furthermore, the building of the random forest with the complexity of (7) could be parallelized; then, asymptotic memory complexity rather than the time complexity could become an issue. In theory, if the random forest building would be parallelized into $\ell \leq t$ independent slave processes each inducting a bunch of $\frac{t}{\ell}$ trees, the time complexity (7) would be reduced to $\Theta\left(\frac{t}{\ell} kn \log n\right) \leq \Theta(\bullet\bullet) \leq \Theta\left(\frac{t}{\ell} kn^2\right)$. Eventually, for $\ell = t$, the random forest building could take the same computing time as only one single tree induction,

$$\begin{gathered} \Theta\left(\frac{t}{\ell} kn \log n\right) \leq \Theta(\bullet\bullet) \leq \Theta\left(\frac{t}{\ell} kn^2\right) \\ \Theta\left(\frac{t}{t} kn \log n\right) \leq \Theta(\bullet\bullet) \leq \Theta\left(\frac{t}{t} kn^2\right) \\ \Theta\left(kn \log n\right) \leq \Theta(\bullet\bullet) \leq \Theta\left(kn^2\right). \end{gathered}$$

When we want to estimate the p-value rather using a confidence interval than only using a point, we need to repeat the random forest building many times, let us say $f \gg 0$ times. As a result, we get a set of random forests that might also be called *a primeval superforest of random forests*. The primeval superforest of random forests construction is of a time complexity $\Theta(\bullet\bullet\bullet)$, so that

$$\Theta(ftkn\log n) \leq \Theta(\bullet\bullet\bullet) \leq \Theta(ftkn^2). \tag{8}$$

However, for a given dataset, a point estimate of the p-value is usually supposed to suffice for purposes of routine statistical inference. The primaveral superforest of random forests of the complexity (8) may be applied rather for experimental reasons when e. g., a posterior distribution of the p-values is about to be investigated.

IV. Simulation study

We compared the log-rank test and the proposed method using several simulations of many pairs of survival curves to get preliminary simulated results, although the method – in contrast to the log-rank test – can compare more than only two survival curves. The curves in pairs were assumed they were not significantly different. We calculated the first type error rates, i. e., rates of false test results that two statistically non-different survival curves are (falsely) detected as different. Also, the lower value of the first type error is, the more robust such a method is. The simulation was repeated for different κ parameter values to illustrate how the value of κ determines the first type error rates.

For generating of the pairs of survival curves, we applied the negatively exponential survival function as follows,

$$s(t) = \rho\left(e^{-\frac{5+\varepsilon}{200}t}\right)$$

where ε is a random white noise term following a standard normal distribution, $\varepsilon \sim \mathcal{N}(0, 1^2)$, and $\rho(\bullet)$ is a function rounding its argument to the nearest multiplier of 0.01 using a half rule, e. g. $\sigma(0.012) = 0.01$, $\sigma(0.350) = 0.35$ or $\sigma(0.048) = 0.05$. A group of negatively exponential survival functions following the formula $s(t) = \rho\left(e^{-\frac{5+\varepsilon}{200}t}\right)$ is in Fig. 7.

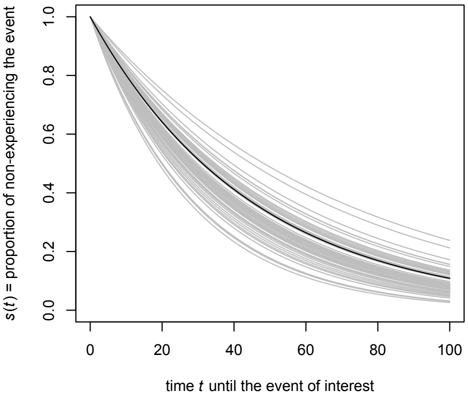

Fig. 7. An example of a group of negatively exponential survival functions following the formula $\rho^{-1}(s(t)) = \rho^{-1}\left(e^{-\frac{5+\varepsilon}{200}t}\right) = e^{-\frac{5+\varepsilon}{200}t}$ for different random values of $\varepsilon \sim \mathcal{N}(0, 1^2)$.

There were $\eta = 1000$ pairs of significantly non-different survival curves generated in total, and for each $\kappa \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, the curves were compared using the log-rank test and the above-proposed method. The number of trees in each random forest was always $t = 1000$. Numbers of cases where p-value was lower than or equal to $\alpha = 0.05$ regardless of the method were summed up, by which we got the point estimates of the first type error rates, as illustrated in table II. The simulation study was performed using R programming language and environment [3]. More on numerical applications of R language to various areas is in [23]–[27].

TABLE II
Point estimates of the first type error rates both for the log-rank test and the proposed method for different values of tuning parameter κ, based on the simulations described above.

	method		
	log-rank test	proposed method	κ
# of simulated cases in total	1000	1000	0.1
# of cases p-value ≤ 0.05	53	65	
first type error rate estimate	0.053	0.065	
# of simulated cases in total	1000	1000	0.3
# of cases p-value ≤ 0.05	48	52	
first type error rate estimate	0.048	0.052	
# of simulated cases in total	1000	1000	0.5
# of cases p-value ≤ 0.05	52	31	
first type error rate estimate	0.052	0.031	
# of simulated cases in total	1000	1000	0.7
# of cases p-value ≤ 0.05	46	14	
first type error rate estimate	0.046	0.014	
# of simulated cases in total	1000	1000	0.9
# of cases p-value ≤ 0.05	55	4	
first type error rate estimate	0.055	0.004	

While the log-rank test returned a point estimate of the first type error rate about 0.050 (regardless of κ since the χ^2 statistics following the formula (1) is not a function of the κ), point estimates of the first type error rates output by the introduced method progressively decreased with increasing value of κ, see table II. What is more, the proposed method seems to be more robust than the log-rank test for large values of κ, based on the simulations above.

V. Conclusion remarks

Survival curves could be compared by the log-rank test when they are only two or by the Cox proportional hazard model if there are more than two curves. However, both methods are limited by statistical assumptions.

We introduced a novel, assumption-free method for survival curves comparison based on a random forest algorithm. Firstly, it requires deriving new variables using the point estimates of modified (personalized) probabilities of non-experiencing the event of interest across all time points. Using those variables as node splitting ones, the random forest model can be built. A subtraction between 1 and a proportion of trees with sufficient complexity, capable of classifying into two or more classes, i. e. groups determined by their survival curves, to all trees of the forest, is a point estimate of p-value of the proposed method. Parameter κ determines the random forest trees' complexity, and, thus, by increasing the parameter, the first type error rate decreases, and robustness of the method increases, as was also illustrated within the simulation study.

The asymptotic time complexity of the random forest-based method is higher than the one for the log-rank test but still polynomial and could be parallelized, too.

The random forest-based method seems to overcome the risk of violations of statistical assumptions of the traditional techniques comparing survival curves and, furthermore, could compare more than two survival curves. Eventually, the method and its computational optimization could also inspire a new R package development.

VI. Acknowledgement

This paper is supported by the grant OP VVV IGA/A, CZ.02.2.69/0.0/0.0/19_073/0016936 with no. 18/2021, which has been provided by the Internal Grant Agency of the Prague University of Economics and Business.

References

[1] E. L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (June 1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452. URL: https://doi.org/10.1080/01621459.1958.10501452.

[2] Huimin Li, Dong Han, Yawen Hou, et al. "Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods". In: *PLOS ONE* 10.1 (Jan. 2015). Ed. by Zhongxue Chen, e0116774. DOI: 10.1371/journal.pone.0116774. URL: https://doi.org/10.1371/journal.pone.0116774.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[4] Therneau T. *survival: A Package for Survival Analysis in R*. Vienna, Austria, R package version 3.1-12. URL: https://CRAN.R-project.org/package=survival/.

[5] F. Kong. "Robust covariate-adjusted logrank tests". In: *Biometrika* 84.4 (Dec. 1997), pp. 847–862. DOI: 10.1093/biomet/84.4.847. URL: https://doi.org/10.1093/biomet/84.4.847.

[6] Rui Song, Michael R. Kosorok, and Jianwen Cai. "Robust Covariate-Adjusted Log-Rank Statistics and Corresponding Sample Size Formula for Recurrent Events Data". In: *Biometrics* 64.3 (Dec. 2007), pp. 741–750. DOI: 10.1111/j.1541-0420.2007.00948.x. URL: https://doi.org/10.1111/j.1541-0420.2007.00948.x.

[7] Richard Peto and Julian Peto. "Asymptotically Efficient Rank Invariant Test Procedures". In: *Journal of the Royal Statistical Society. Series A (General)* 135.2 (1972), p. 185. DOI: 10.2307/2344317. URL: https://doi.org/10.2307/2344317.

[8] Georg Heinze, Michael Gnant, and Michael Schemper. "Exact Log-Rank Tests for Unequal Follow-Up". In: *Biometrics* 59.4 (Dec. 2003), pp. 1151–1157. DOI: 10.1111/j.0006-341x.2003.00132.x. URL: https://doi.org/10.1111/j.0006-341x.2003.00132.x.

[9] Song Yang and Ross Prentice. "Improved Logrank-Type Tests for Survival Data Using Adaptive Weights". In: *Biometrics* 66.1 (Apr. 2009), pp. 30–38. DOI: 10.1111/j.1541-0420.2009.01243.x. URL: https://doi.org/10.1111/j.1541-0420.2009.01243.x.

[10] Chenxi Li. "Doubly robust weighted log-rank tests and Renyi-type tests under non-random treatment assignment and dependent censoring". In: *Statistical Methods in Medical Research* 28.9 (July 2018), pp. 2649–2664. DOI: 10.1177/0962280218785926. URL: https://doi.org/10.1177/0962280218785926.

[11] Donald G. Thomas. "Exact and asymptotic methods for the combination of 2 × 2 tables". In: *Computers and Biomedical Research* 8.5 (Oct. 1975), pp. 423–446. DOI: 10.1016/0010-4809(75)90048-8. URL: https://doi.org/10.1016/0010-4809(75)90048-8.

[12] Cyrus R. Mehta, Nitin R. Patel, and Robert Gray. "Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2 × 2 Contingency Tables". In: *Journal of the American Statistical Association* 80.392 (Dec. 1985), p. 969. DOI: 10.2307/2288562. URL: https://doi.org/10.2307/2288562.

[13] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. "Analysis of asymptotic time complexity of an assumption-free alternative to the log-rank test". In: *Proceedings of the 2020 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2020. DOI: 10.15439/2020f198. URL: https://doi.org/10.15439/2020f198.

[14] Karl Mosler. *Multivariate dispersion, central regions, and depth : the lift zonoid approach*. New York: Springer, 2002. ISBN: 0387954120.

[15] Tomasz Smolinski. *Computational intelligence in biomedicine and bioinformatics : current trends and applications*. Berlin: Springer, 2008. ISBN: 978-3-540-70776-9.

[16] Alexander Kulikov. *Combinatorial pattern matching : 25th annual symposium, CPM 2014 Moscow, Russia, June 16-18, 2014, proceedings*. Cham: Springer, 2014. ISBN: 978-3-319-07565-5.

[17] Nihal Ata Tutkun and Muhammet Tekin. "Cox Regression Models with Nonproportional Hazards Applied to Lung Cancer Survival Data". In: *Hacettepe Journal of Mathematics and Statistics Volume* 36 (Jan. 2007), pp. 157–167.

[18] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324. URL: https://doi.org/10.1023/a:1010933404324.

[19] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246. URL: http://www.jstor.org/stable/2985181.

[20] Lubomír Štěpánek, Filip Habarta, Ivana Malá, et al. "A Machine-learning Approach to Survival Time-event Predicting: Initial Analyses using Stomach Cancer Data". In: *2020 International Conference*

on e-Health and Bioengineering (EHB). IEEE, Oct. 2020. DOI: 10.1109/ehb50910.2020.9280301. URL: https://doi.org/10.1109/ehb50910.2020.9280301.

[21] Xiaonan Xue, Xianhong Xie, Marc Gunter, et al. "Testing the proportional hazards assumption in case-cohort analysis". In: *BMC Medical Research Methodology* 13.1 (July 2013). DOI: 10.1186/1471-2288-13-88. URL: https://doi.org/10.1186/1471-2288-13-88.

[22] Leo Breiman. *Classification and regression trees*. New York: Chapman & Hall, 1993. ISBN: 9780412048418.

[23] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Evaluation of facial attractiveness for purposes of plastic surgery using machine-learning methods and image analysis". In: *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, Sept. 2018. DOI: 10.1109/healthcom.2018.8531195. URL: https://doi.org/10.1109/healthcom.2018.8531195.

[24] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language". In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[25] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-Learning and R in Plastic Surgery – Evaluation of Facial Attractiveness and Classification of Facial Emotions". In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Sept. 2019, pp. 243–252. DOI: 10.1007/978-3-030-30604-5_22. URL: https://doi.org/10.1007/978-3-030-30604-5_22.

[26] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Machine-learning at the service of plastic surgery: a case study evaluating facial attractiveness and emotions using R language". In: *Proceedings of the 2019 Federated Conference on Computer Science and Information Systems*. IEEE, Sept. 2019. DOI: 10.15439/2019f264. URL: https://doi.org/10.15439/2019f264.

[27] Lubomír Štěpánek, Pavel Kasal, and Jan Měšťák. "Evaluation of Facial Attractiveness after Undergoing Rhinoplasty Using Tree-based and Regression Methods". In: *2019 E-Health and Bioengineering Conference (EHB)*. IEEE, Nov. 2019. DOI: 10.1109/ehb47216.2019.8969932. URL: https://doi.org/10.1109/ehb47216.2019.8969932.