Journal of Data Science 0 (0), 1–21
??? 2022

DOI: 10.6339/22-JDS1060
Computing in Data Science

Subpopulation Treatment Effect Pattern Plot (STEPP) Methods with R and Stata

Sergio Venturini[1,]*, Marco Bonetti[2,3], Ann A. Lazar[4,5], Bernard F. Cole[6], Xin Victoria Wang[7,8], Richard D. Gelber[7,8], and Wai-Ki Yip[9]

[1] *Department of Economic and Social Sciences, Università Cattolica del Sacro Cuore, Via Bissolati 74, 26100 Cremona, Italy*

[2] *Carlo F. Dondena Research Centre, Università Commerciale L. Bocconi, Via Röntgen 1, 20136 Milano, Italy*

[3] *Bocconi Institute for Data Science and Analytics, Università Commerciale L. Bocconi, Via Röntgen 1, 20136 Milano, Italy*

[4] *Division of Oral Epidemiology, Department of Preventive and Restorative Dental Sciences, University of California, San Francisco, 3333 California Street, San Francisco, CA 94118, USA*

[5] *Division of Biostatistics, Department of Epidemiology and Biostatistics, University of California, San Francisco, 3333 California Street, San Francisco, CA 94118, USA*

[6] *Department of Mathematics and Statistics, University of Vermont, 33 Colchester Avenue, Burlington, VT 05405, USA*

[7] *Department of Data Science, Dana-Farber Cancer Institute, 450 Brookline Avenue, Mailstop CLS 11007, Boston, MA 02215, USA*

[8] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02215, USA*

[9] *Agenus, Inc., 3 Forbes Road, Lexington, MA 02421, USA*

Abstract

We introduce the *stepp* packages for R and Stata that implement the subpopulation treatment effect pattern plot (STEPP) method. STEPP is a nonparametric graphical tool aimed at examining possible heterogeneous treatment effects in subpopulations defined on a continuous covariate or composite score. More pecifically, STEPP considers overlapping subpopulations defined with respect to a continuous covariate (or risk index) and it estimates a treatment effect for each subpopulation. It also produces confidence regions and tests for treatment effect heterogeneity among the subpopulations. The original method has been extended in different directions such as different survival contexts, outcome types, or more efficient procedures for identifying the overlapping subpopulations. In this paper, we also introduce a novel method to determine the number of subjects within the subpopulations by minimizing the variability of the sizes of the subpopulations generated by a specific parameter combination. We illustrate the packages using both synthetic data and publicly available data sets. The most intensive computations in R are implemented in Fortran, while the Stata version exploits the powerful Mata language.

Keywords *clinical trial; interaction; subgroup analysis; subpopulation; treatment-covariate interaction*

*Corresponding author. Email: sergio.venturini@unicatt.it.

Received May 5, 2022; Accepted July 3, 2022

1 Introduction

Results from randomized clinical trials (RCTs) provide the basis for evidence-based medicine by comparing the benefits of alternative therapies. Usually, the assessment of effectiveness for the different therapeutic strategies is based on the entire cohort of patients enrolled in the study. However, the extent of the treatment effect can be heterogeneous among subsets of patients defined by prognostic factors. Instead of the traditional "one-size-fits-all" treatment recommendation, understanding the interplay between treatment effect and covariates of interest may provide the necessary information to customize treatment for individuals to maximize the benefit. A similar situation also occurs in other fields, such as epidemiology or the social sciences. Depending on the field, the terms that are commonly used to refer to these assessments are *interaction* or *moderation analysis* (see for example VanderWeele, 2015, Hayes, 2017, Imai, 2017).

A simple but effective approach to perform this kind of evaluation consists of the estimation of the treatment effect within (disjoint) subsets of the patient population, a practice that is often referred to as *subgroup analysis*. Performing subgroup analysis is in general a challenging task (Lagakos, 2006; Wang et al., 2007; Pocock, 2008). The two major statistical concerns are the inflation of false positive rates due to repeated testing and the lack of power to detect different treatment effects across subgroups. Traditionally, patients are divided into subgroups according to the median, quartiles or other convenient cut-points of one or more covariates of interest, and treatment comparisons are then performed within each subgroup. Unfortunately, the convenient cut-points do not necessarily create clinically relevant subgroups. Furthermore, there may be few patients in some subgroups, thus reducing the precision of the groupwise estimates of the treatment effect. For survival data, treatment-covariate interactions can be analyzed using regression methods such as the Cox proportional hazards model (Cox, 1972) or cumulative incidence models (Gray, 1988; Fine and Gray, 1999). In these cases the focus of the analysis lies on the significance test of the interaction's coefficient. In this work we focus on another method for examining possible treatment effect heterogeneity, the subpopulation treatment effect pattern plot (STEPP) method first introduced by Bonetti and Gelber (2000). STEPP is an exploratory graphical tool designed to help researchers investigate the potential heterogeneity of treatment effects and facilitate the interpretation of estimates of treatment effect derived from different but potentially overlapping subsets of patients. The method is aimed at determining whether the magnitude of the treatment effect changes as a function of the values of the covariate used to define the subpopulations. STEPP addresses some of the concerns associated with traditional subgroup analysis. Indeed, the main advantages of STEPP are the fact that it requires few assumptions and that it provides a graphical display to show potentially complex interactions, thus assisting researchers in the interpretation of the results. While traditional statistical methods for subgroup analysis divide the population into disjoint subgroups, STEPP takes a different approach by constructing *overlapping* subpopulations along the continuum of a continuous covariate of interest (e.g., a biomarker), thereby improving the precision of the estimated treatment effect within the subgroups. For each subpopulation, an estimate of treatment effect is computed. Such treatment effect estimates are clearly correlated, as neighboring subpopulations share patients. The estimates are represented graphically in different diagrams to help the researcher interpret the results, together with relevant inference.

In this article, we introduce the *stepp* packages that implement all the methods currently available for conducting a STEPP analysis. We developed the same package for the R and Stata software. The R version includes some Fortran code to perform the most intensive computations, while the Stata version of the package exploits the power of the Mata language. The R current version of the *stepp* package is available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/stepp and it can be installed directly using, for example, `install.packages("stepp")`, while the development version of the package can be retrieved from https://github.com/steppdev/stepp and installed from within R with the command `devtools::install_github("steppdev/stepp")`. The Stata version is available at https://github.com/sergioventurini/stepp, where the reader will also find some installation guidelines. The work by Royston and Sauerbrei (2009) also describes an implementation of the STEPP approach through the `stepp_window` and `stepp_tail` commands, but these are limited to the original STEPP method (Bonetti and Gelber, 2000). Moreover, these commands are only available for Stata. The *stepp* packages we present here provide a more comprehensive and current implementation of STEPP.

Other frequentist and Bayesian approaches have been developed over the years to address the same problem (see for example Simon et al., 1995; Simon, 2002; Royston and Sauerbrei, 2007; Foster et al., 2011; Zhao et al., 2013; Li et al., 2015). A common approach is based on the mutivariable fractional polynomials (MFPs) methodology (Royston and Sauerbrei, 2004, 2008, 2009). Fractional polynomials have been introduced by Royston and Altman (1994) as an extension of polynomial models for determining the functional form of a continuous predictor. These models are well suited for nonlinear data and present advantages over the standard polynomials both in terms of range of curve shapes and avoidance of undesirable artifacts such as edge effects and waves. MFPs represent the generalization of fractional polynomials to more than one variable. Substantially, MFPs are used to test for a treatment-covariate interaction by fitting separate models within the treatment groups and then assess the differences by means of likelihood ratio tests. Most of these methods have been implemented in specific software tools. For example, the approach based on MFPs is implemented in the `mfpi` Stata command and it is described in Royston and Sauerbrei (2009). To install it run the command `net sj 9-2 st0164` followed by `net install st0164.pkg` within a Stata session. The `mfpi` command takes advantage of the `mfp` command included in Stata (Royston and Ambler, 1998).

This article first reviews the STEPP methods and then moves to present the package features, focusing on the R version. More specifically, the work is organized as follows: Section 2 reviews the key methodological aspects of the different STEPP approaches. Section 3 provides a description of the features available in the *stepp* packages and shows how they can be applied to simulated data. In Section 4 we apply the package functionalities to a data set that is available within the packages. Finally, Section 5 summarizes and provides some closing remarks.

2 The STEPP Methodology

We consider the general situation of a random sample of n units on which we collected the values of an outcome measure Y, a treatment indicator X as well as a covariate Z. We consider here only the case of a unidimensional covariate Z. The typical context where the STEPP method is applied is a clinical trial in which subjects are randomized to one of two treatments and a baseline covariate is observed for all subjects. We assume that the covariate Z is continuous or ordered categorical, and that it does not change over time. Moreover, we let Z take values in the range $[Z_{\text{min}}, Z_{\text{max}}] \in \mathbb{R}$. Then, a STEPP analysis is typically conducted through the following steps:

1. generate the subpopulations based on the covariate of interest Z;
2. estimate the treatment effect within each subpopulation using one of the available measures

depending on the outcome type or the context (i.e., survival data with or without competing risks, numeric, binary or count);

3. perform inference by constructing pointwise or simultaneous confidence regions around the collection of treatment effects and by computing a permutation test for the equality of the treatment effects across the different subpopulations;
4. report the results in some summary plots, which represent the main STEPP output (together with additional text output).

In the next sections we provide more details for each step.

2.1 Generation of the Subpopulations

Given the observed covariate values z_i for each sample unit, the first step consists of the generation of the subpopulations by selecting a finite collection of points $\{(l_j, u_j), j = 1, \ldots, K\}$ for the covariate Z such that $l_j + \eta \leqslant u_j$ for all $j = 1, \ldots, K$, where η is a positive constant (for the technical justification of why we need to add η see Bonetti and Gelber (2004), in particular Appendix 1). Therefore, each pair (l_j, u_j) defines a particular subset of observations, those for which $l_j \leqslant z_i \leqslant u_j$. We will denote the corresponding set of indexes (i.e., the subpopulation) as $\mathcal{P}_j$. It must clearly hold that $Z_{\text{min}} \leqslant l_j < u_j \leqslant Z_{\text{max}}$ for all $j = 1, \ldots, K$. Different approaches have been advanced to specify the (l_j, u_j) pairs, in particular:

- the *unit-based sliding window* approach, originally referred to simply as sliding window (Bonetti and Gelber, 2000, 2004);
- the *event-based sliding window* approach, which has been introduced more recently to improve the stability of the STEPP results in some contexts (Lazar et al., 2016);
- the *tail-oriented* approach (Bonetti and Gelber, 2000).

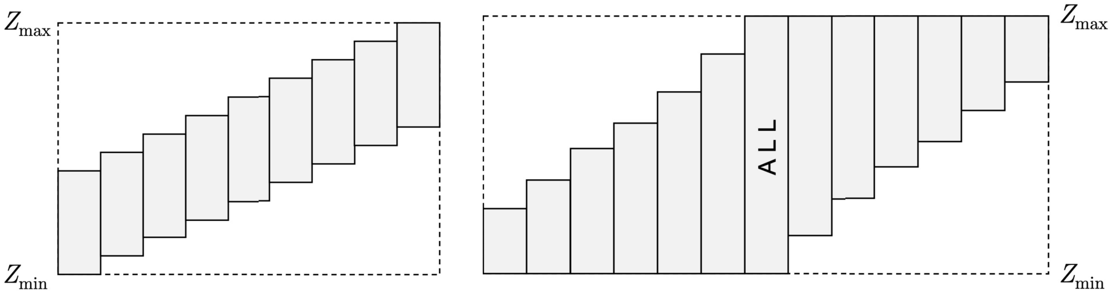
The distinctive feature of STEPP as compared to other methods such as subgroup analysis is that, no matter which windowing system is chosen, the subpopulations that are produced overlap. We now briefly illustrate the different approaches developed so far.

2.1.1 Unit-based Sliding Window Subpopulations

The unit-based sliding window approach is implemented by specifying two parameters usually denoted as r_1 and r_2, with $r_1 < r_2 < n$. The former number, r_1, provides the (approximate) extent of the overlap between the subpopulations in terms of number of units in common by each pair of consecutive subpopulations. The latter parameter, r_2, indicates the (approximate) size of each subpopulation. More specifically, for a fixed choice of r_1 and r_2 (we discuss the choice of the r_1 and r_2 values later, in particular see Section 2.1.4):

1. the first subpopulation $\mathcal{P}_1$ is defined as the set of units with covariate value z_i in between $l_1 = Z_{\text{min}}$ and u_1 corresponding to the sample $(r_2/n \times 100)$th percentile of the covariate. In case the exact percentile does not exist, u_1 is chosen to allow $\mathcal{P}_1$ include at least r_2 units;
2. the second subpopulation $\mathcal{P}_2$ is defined by choosing l_2 so that at most r_1 units fall between u_1 and l_2, while u_2 is identified to allow $\mathcal{P}_2$ include at least r_2 units;
3. the same process is then iterated till the last subpopulation.

Note that setting $r_1 = \lfloor np_1 \rfloor$ and $r_2 = \lfloor np_2 \rfloor$, with p_1 and p_2 so that $0 < p_1 < p_2 < 1$, the implementation above guarantees that the proportion of units in each subpopulation (except for the last one, which is defined residually) converges to p_2 as $n \to \infty$ (assuming that the covariate Z is continuous). In addition, it can be shown (Bonetti and Gelber, 2004) that the number of subpopulations converges to the smallest integer greater than or equal to $[1 + (1 - p_2)/(p_2 - p_1)]$.



(a) Subpopulations using the (unit-based) sliding window approach.

(b) Subpopulations using the tail-oriented approach.

Figure 1: Comparison of different approaches for the generation of subpopulations in a STEPP analysis.

A generic unit-based sliding window pattern is represented in Figure 1a, where the horizontal axis reports the K subpopulations while the vertical axis provides the covariate values. Currently, unit-based sliding window is the recommended approach and thus we have chosen it as the default in the packages.

2.1.2 Event-based Sliding Window Subpopulations

In specific applications such as in clinical trials with a time-to-event (i.e., survival) outcome, the generation of subpopulations through the unit-based sliding window approach may encounter problems because not enough events may be available in each subpopulation to guarantee a statistically sound estimation of the treatment effects. As a consequence, in presence of sparse and imbalanced events across treatments the STEPP analysis produces unstable results and an inflated type I error rate of the interaction test. A solution that has been proposed is to generate the subpopulations by pre-specifying the number of *events* instead of the number of units within each subpopulation (Lazar et al., 2016). This allows to produce treatment effect estimates that have similar variances across the subpopulations.

In analogy with the unit-based approach, one can implement the event-based window pattern by specifying two numbers, e_1 and e_2, where e_1 represents the largest number of events in common among consecutive (overlapping) subpopulations of each treatment group, while e_2 indicates the minimum number of events in each treatment group of each subpopulation. It must clearly hold that $e_2 > e_1$. The overlapping subpopulations are then constructed as follows:

1. units are ordered from the lowest to highest value of the covariate Z;
2. the first subpopulation $\mathcal{P}_1$ is formed by those units with at least e_2 events within each treatment group with the lowest covariate values;
3. the second subpopulation $\mathcal{P}_2$ is defined by removing units with $(e_2 - e_1)$ events with the lowest covariate values from $\mathcal{P}_1$ and replacing them with the next set of units with $(e_2 - e_1)$ events in the ordered list;
4. the process continues iteratively till all units have been included in at least one subpopulation and each subpopulation includes at least e_2 events per each treatment.

In cases when the last subpopulation does not include at least e_2 events, it will be combined with the previous subpopulation. In the context of competing risks, the e_1 and e_2 values denote the number of events related to the cause of interest. Note that currently the event-based sliding window approach has been developed only for competing risks analyses. The extension to other survival endpoints is one of the future directions in the STEPP agenda.

2.1.3 Tail-oriented Subpopulations

In the tail-oriented framework the (l_j, u_j) pairs are set so that $\{l_j = Z_{\min}, u_j \in [Z_{\min}, Z_{\max}]\}$, or $\{l_j \in [Z_{\min}, Z_{\max}], u_j = Z_{\max}\}$. In the former case, the lower bounds l_j are all set to $Z_{\min}$ while the u_j upper bounds are defined as $u_1 < u_2 < \cdots < u_K = Z_{\max}$. Therefore, the jth subpopulation $\mathcal{P}_j$ contains all units for which $z_i \leqslant u_j$. In other words, the first approach defines the subpopulations as (strictly) increasing sets of units, with the last one including all the n sample units. This situation is represented in the left part of Figure 1b, where the horizontal axis reports the K subpopulations and the vertical axis represents the covariate values. The label “ALL” refers to the entire sample. Alternatively, the second case listed above corresponds to the opposite situation where the u_j are all fixed at $Z_{\max}$ and the l_j are chosen so that $Z_{\min} = l_1 < l_2 < \cdots < l_K$. In this case the subpopulations form a sequence of decreasing index sets for which $z_i \geqslant l_j$. This situation is represented in the right part of Figure 1b starting from the central rectangle labeled as “ALL” up to the rightmost one.

In the tail-oriented approach the cutoffs l_j or u_j may be defined according to different criteria such as: (1) a constant number of units added (removed) from each subpopulation compared to the next one, (2) values that are particularly relevant in the context of the application under scrutiny, for example disease-specific cutoffs that are of clinical relevance, (3) the observed values of a discrete covariate.

We note that, due to its definition, the aim of the tail-oriented framework is primarily to focus the analysis on the impact of increasing values of the covariate to the right, and of decreasing values of the covariate to the left on the magnitude of the treatment effect. Furthermore, it is usually recommended that any testing procedure be applied separately to the left and right parts of the plot shown in Figure 1b. The tail-oriented approach was introduced in the original STEPP paper (Bonetti and Gelber, 2000), because the overall “ALL” result is featured in the center of the plot as an anchor against which increasing or decreasing covariate values are depicted. It is, therefore, included in the packages although most recent applications of STEPP favor the sliding window approach.

2.1.4 Selection of the Number of Subpopulations

The number of STEPP subpopulations generated by the sliding window approach is defined by the r_1 and r_2 parameters together with the distribution of the covariate values. Ideally, the selection of r_1 and r_2 results in a sufficient number of subpopulations for a meaningful analysis, with each containing a number of subjects as near to the specified r_2 value as possible. However, in some cases, the sliding window approach will result in uneven numbers of subjects (usually with one subpopulation being much smaller than the others). To address this issue, we have implemented a simple algorithm that provides the (r_1, r_2) choice that minimizes the variability of the subpopulations sizes. More specifically, let $R_1 = \left[r_1^{\min}, r_1^{\max}\right]$ and $R_2 = \left[r_2^{\min}, r_2^{\max}\right]$ be the ranges of the two parameters one wants to consider. Then, for each pair $\{r_1, r_2\}$, with $r_1 \in R_1$, $r_2 \in R_2$ and $r_1 < r_2$, the unit-based sliding window algorithm is run and the variance of the corresponding subpopulation sizes is computed. The values r_1^* and r_2^* are then identified as those corresponding to the smallest variance in the numbers of subjects within the subpopulations. Note that the identification of the r_1^* and r_2^* values is performed without knowledge of the p-value associated to each parameter combination, thus ensuring that no “fishing for significance” is performed. The same approach can also be extended to selecting the subpopulations in a balanced manner but with respect to the number of events.

2.2 Treatment Effect Estimation

After the overlapping subpopulations have been generated, the treatment effect θ_j for subpopulation j, with $j = 1, \dots, K$, is estimated using the appropriate method depending on the type of the outcome variable and context. Currently, STEPP has been developed to deal with the following cases:

- time-to-event (survival) outcomes when the treatment effect θ_j is measured as the difference $S_A(t^*) - S_B(t^*)$ in survival at a fixed time point t^* between treatment arms A and B; in this context the survival functions are estimated through the Kaplan-Meier estimator (Bonetti and Gelber, 2004);
- time-to-event (survival) outcomes in presence of competing risks when the treatment effect θ_j is represented by the log hazard ratio, which is estimated by a first-order approximation of the partial likelihood (Lazar et al., 2016). In particular, in that paper the authors show how to extend the "O minus E methodology" (Peto et al., 1977), where O denotes the observed number of events among the treated subjects and E is the corresponding expected (log-rank) number of events, to the competing risks setting (Fine and Gray, 1999) for the cases of no censoring, complete censoring and randomly censored data;
- a generic (i.e., non time-to-event) outcome Y with a distribution belonging to the exponential family; in this context treatment effects θ_j are estimated via generalized linear models (GLMs) and can be expressed in either absolute or relative terms (Yip et al., 2016). More specifically, denoting the treatment group as G, the absolute treatment effect corresponds to the difference between the treatment-specific outcome expected values, that is $[E(Y|G = A) - E(Y|G = B)]$, for all outcome types, while the relative effect is measured by the ratio of the treatment-specific outcome expected values, $E(Y|G = A)/E(Y|G = B)$, for Gaussian and Poisson distributed outcomes, or as the odds ratio, $\frac{E(Y|G=A)/[1-E(Y|G=A)]}{E(Y|G=B)/[1-E(Y|G=B)]}$, for binary outcomes.

2.3 Inference

To perform inference on the vector of treatment effects $(\theta_1, \dots, \theta_K)$ for the K subpopulations $\mathcal{P}_1, \dots, \mathcal{P}_K$, one must derive the joint distribution of the corresponding vector of estimates $(\widehat{\theta}_1, \dots, \widehat{\theta}_K)$. Bonetti and Gelber (2004) have shown that under mild conditions the asymptotic distribution of the vector of estimates $(\widehat{\theta}_1, \dots, \widehat{\theta}_K)$ is multivariate normal with mean vector $(\theta_1, \dots, \theta_K)$ and covariance matrix Σ, that is

$$\sqrt{n}\begin{pmatrix} \widehat{\theta}_1 - \theta_1 \\ \vdots \\ \widehat{\theta}_K - \theta_K \end{pmatrix} \xrightarrow{d} N_K\left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \Sigma\right), \tag{1}$$

where Σ can be consistently estimated from the data. Note that the specific form of the matrix Σ depends on the outcome type considered (see Section 2.2).

Thanks to the result above, one can then compute simultaneous confidence regions around the collection of estimators. More specifically, a $(1 - \alpha) \times 100\%$ rectangular confidence region is defined as $\{\theta_j \in \widehat{\theta}_j \pm \gamma \cdot z_{\alpha/2} \cdot \widehat{\sigma}_j,\ j = 1, \dots, K\}$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$-th percentile for a standard normal random variable, $\widehat{\sigma}_j = \left[\widehat{\mathrm{var}}(\widehat{\theta}_j)\right]^{1/2}$ is a consistent estimator of $\left[\mathrm{var}(\widehat{\theta}_j)\right]^{1/2}$, while γ is obtained by solving numerically the following equation

$$P\left(\bigcap_{j=1}^{K}\left\{\theta_j \in \widehat{\theta}_j \pm \gamma \cdot z_{\frac{\alpha}{2}} \cdot \widehat{\sigma}_j\right\}\right) = 1 - \alpha$$

for a sample of random variables generated from the estimated asymptotic distribution of the estimates. The quantity γ represents the widening of the marginal confidence intervals that is necessary to produce the desired simultaneous coverage of the confidence region.

Together with a confidence region for the vector of treatment effects, one may also desire to test the null hypothesis of equality of treatment effects, that is $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$, which corresponds to the absence of interaction between the covariate of interest (i.e., across subpopulations) and the treatment effect. One possibility is to develop an “omnibus” test using a quadratic form based on the asymptotic normality of the estimators. However, it has been shown (Bonetti and Gelber, 2000) that such a test is too sensitive to the particular choice of the subpopulations and thus it is not recommended. An alternative approach is to base the test on the following supremum test statistic

$$T = \max\left\{\frac{\left|\widehat{\theta}_j - \widehat{\theta}_{ALL}\right|}{\left[\widehat{\mathrm{var}}\left(\widehat{\theta}_j - \widehat{\theta}_{ALL}\right)\right]^{1/2}},\ j = 1, \ldots, K\right\}, \tag{2}$$

where $\widehat{\theta}_{ALL}$ represents the treatment effect estimate computed on all units in the study and $[\widehat{\mathrm{var}}(\widehat{\theta}_j - \widehat{\theta}_{ALL})]^{1/2}$ is a consistent estimator of $[\mathrm{var}(\widehat{\theta}_j - \widehat{\theta}_{ALL})]^{1/2}$. The asymptotic null distribution of T can be estimated by sampling repeatedly from the asymptotic distribution of the (scaled) vector $(\widehat{\theta}_1 - \theta_1, \ldots, \widehat{\theta}_K - \theta_K, \widehat{\theta}_{ALL} - \theta_{ALL})$ under the null hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K = \theta_{ALL}$, and a Monte Carlo p-value can thus be produced.

The finite sample properties of the procedures above for the case of survival outcomes are discussed in Bonetti et al. (2009). The results show an inflation of the type I error probability even for large sample sizes, which clearly may lead to false rejections of the null hypothesis. The same results also show that the coverage of the confidence region can be considered satisfactory only for sample sizes equal to or above 500. To overcome the limitations of the inferential procedures described above, in the same paper the authors propose to adopt a permutation distribution approach to inference (see for example Pesarin, 2001) in which one permutes the covariate values across the units within each treatment group and then re-computes the test statistic (2) on the permuted samples. The variances are also estimated from the permuted samples. This procedure returns a sample from the permutation distribution of the test statistic T that can be used for testing. More specifically, the permutation p-value is computed as the proportion of times the permutation-based statistic is more extreme than the observed test statistic value.

Different simulation studies have shown that the permutation-based test improves the performance of the asymptotic approach in all cases, that is for survival outcomes (Bonetti et al., 2009), for survival outcomes with competing risks (Lazar et al., 2016), as well as for outcomes from the exponential family (Yip et al., 2016).

2.4 The Subpopulation Treatment Effect Pattern Plot

The main output of a STEPP analysis is a plot of the estimated treatment effects $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ against the median covariate value in each subpopulation. In addition, the plot also reports the corresponding confidence regions and the subpopulation sizes. The treatment effect estimates may be reported either in absolute or relative terms. We remark that the ability to detect heterogeneity both graphically and via statistical testing may depend on the scale used for measuring the effects. As such, the treatment effects may be statistically significant on one scale but not on the other. Finally, note that the overall treatment effect (i.e., θ_{ALL}) is, in general, not a linear combination of the subpopulations' treatment effects. The output of a STEPP analysis also includes the description of the subpopulations and the corresponding treatment effect estimates.

3 The *stepp* Package

The *stepp* packages allow one to perform the STEPP analyses described in Section 2 and produce the corresponding plots for all outcomes and treatment effect measures. In this section we provide a more practical description of the package characteristics focusing in particular on the R implementation.

The *stepp* package uses the S4 framework and it is built around the following six classes (more details can be found in the package documentation):

- `stwin`, which defines objects to represent the window system type to use for generating the overlapping subpopulations in a STEPP analysis,
- `stsubpop`, which defines objects representing the subpopulations, the main component of a STEPP analysis,
- `steppes`, which provides the object class containing the results regarding a specific STEPP analysis,
- `stmodelKM`, which defines objects are for a STEPP analysis involving survival outcomes with no competing risks,
- `stmodelCI`, which defines objects for the STEPP analyses of survival outcomes in presence of competing risks,
- `stmodelGLM`, which defines objects for the STEPP analyses of outcomes with distribution belonging to exponential family.

For all these classes, the package provides specific methods for initializing, printing, summarizing and plotting, some of which will be illustrated in more detail in the rest of this section.

The steps required to perform a STEPP analysis with the *stepp* package are:

1. choose the windowing system to use and the corresponding parameters by defining a new `stwin` object
2. define a new `stsubpop` object that will contain the overlapping subpopulations to use in the STEPP analysis
3. populate the `stsubpop` object by generating the subpopulations with a call to the `generate()` generic function
4. create a new `steppes` object which will enclose all the results of the analysis
5. depending on the problem context, create a new model object of class either `stmodelKM`, `stmodelCI` or `stmodelGLM`
6. perform the analysis by calling the `estimate()` and `test()` generic functions; the former computes the treatment effect estimates, while the latter performs the permutation-based test
7. finally, the results contained in the `steppes` object can be represented graphically by calling the corresponding `plot` method.

As a demonstration of this workflow, we present an example using some simulated data contained in the simdataKM object available in the package (the code to generate these data can be retrieved from the simdataKM documentation). These data can be used to perform a STEPP analysis for a survival outcome without competing risks. First, we create a new stwin object using the unit-based sliding window approach with $r_1 = 200$ and $r_2 = 300$:

```
library(‘‘stepp’’)
data(‘‘simdataKM’’, package = ‘‘stepp’’)
swin <- new(‘‘stwin’’, type = ‘‘sliding’’, r1 = 200, r2 = 300)
```

Then, we create a new stsubpop object and generate the overlapping subpopulations using the covariate values in the covar column of the simdataKM data frame:

```
subp <- new(‘‘stsubpop’’)
subp <- generate(subp, win = swin, covariate = simdataKM$covar)
summary(subp)
```

```
Window type: sliding
Number of patients per subpopulation (patspop r2): 300
Largest number of patients in common among consecutive subpopulations (minpatspop r1): 200
Number of subpopulations created: 8

Subpopulation summary information (including all treatments)
                                     Covariate Summary                  Sample
     Subpopulation         Median       Minimum       Maximum             size
           1                47.79       33.1156       51.3941              300
           2                50.42       45.8536       53.2784              300
           3                52.46       49.3066       55.0570              300
           4                54.30       51.4095       56.7860              300
           5                56.09       53.2844       58.9007              300
           6                57.62       55.0621       61.0561              300
           7                59.86       56.8181       64.1091              300
           8                62.38       58.9031       77.2494              300
```

The output shows that 8 subppopulations have been generated and it reports the corresponding minimum, maximum and median covariate values as well as their sizes, which in this example correspond exactly to the r_2 parameter value.

The next step requires the creation of a new steppes object, which will then be populated with the estimates produced by the estimate() function as well as with the results of the permutation test returned by the test() function. In this example we are interested in assessing the treatment effect measured as the difference in the survival functions at time 4 (output partly omitted):

```
res <- new(‘‘steppes’’)
modelKM <- new(‘‘stmodelKM’’, coltrt = simdataKM$trt,
  survTime = simdataKM$time, censor = simdataKM$censor,
  trts = c(1, 2), timePoint = 4)
res <- estimate(res, subp, modelKM)

set.seed(101)
```

```
nperm <- 500
res <- test(res, nperm)
print(res, estimate = TRUE, cov = FALSE, test = TRUE)
```

```
Sample size in treatment 1: 479
Sample size in treatment 2: 521
Total sample size (excluding missing data): 1000

[...]

Survival differences at time point and hazard ratio estimates

trt 1 vs. trt 2

Survival differences at time point 4
Comparing trt 1 vs. trt 2

                         Survival
     Subpopulation      Difference      Std. Err.
           1             -0.3147          0.0567
           2             -0.2337          0.0594
           3             -0.2299          0.0594
           4             -0.2030          0.0592
           5             -0.2384          0.0577
           6             -0.1464          0.0585
           7             -0.1280          0.0595
           8             -0.0263          0.0621
        Overall          -0.1910          0.0324

Hazard ratio estimates
     Subpopulation        Log HR       Std. Err.       Hazard Ratio
           1            1.328788       0.206001            3.78
           2            0.803914       0.188752            2.23
           3            0.703132       0.177723            2.02
           4            0.500376       0.175574            1.65
           5            0.638177       0.172646            1.89
           6            0.399095       0.172902            1.49
           7            0.474916       0.169297            1.61
           8            0.173798       0.172673            1.19
        Overall         0.632647       0.098281            1.88

Supremum test results
trt 1 vs. trt 2
Interaction p-value based on Kaplan-Meier estimates: 0.006
Interaction p-value based on hazard ratio estimates: 0

Chi-square test results
Interaction p-value based on Kaplan-Meier estimates: 0.028
```

The results show the presence of a significant treatment-covariate interaction because the supremum test p-value is smaller than the usual 0.05 significance level. Moreover, the tables reported in the output permit to describe how the covariate interacts with the treatment by providing the results in the different subpopulations. The rows labeled `Overall` correspond to the estimation results for the whole sample. The numbers in the tables allow to conclude that treatment 2 appears to be more effective than treatment 1 because it provides a higher probability to survive. More specifically, the differences in survival decrease as the covariate values increase, with the survival difference in the last subpopulation being very close to zero. Finally, we note that in the output above we skipped printing the estimate of the asymptotic covariance matrix Σ defined in (1) because we specified `cov = FALSE` in the print method.

To aid the user in getting the results, the *stepp* package also provides the `stepp.test()` wrapper function that allows to automate some of the steps described above. The package also includes constructor functions for the different object classes described above. More specifically, the constructor functions are `stepp.win()`, `stepp.subpop()`, `stepp.KM()`, `stepp.CI()` and `stepp.GLM()`. The aim of these functions is substantially to hide to the end user the call to the `new()` function when creating a new instance for the corresponding object classes. In particular, `stepp.test()` directly implements steps 4, 6 and 7, that is the creation of a new `steppes` object, estimation of the model and execution of the permutation test. So, the code provided above for the STEPP analysis on the `simdataKM` data can be recast more compactly as follows:

```
set.seed(101)
res <- stepp.test(subp, modelKM, nperm)
```

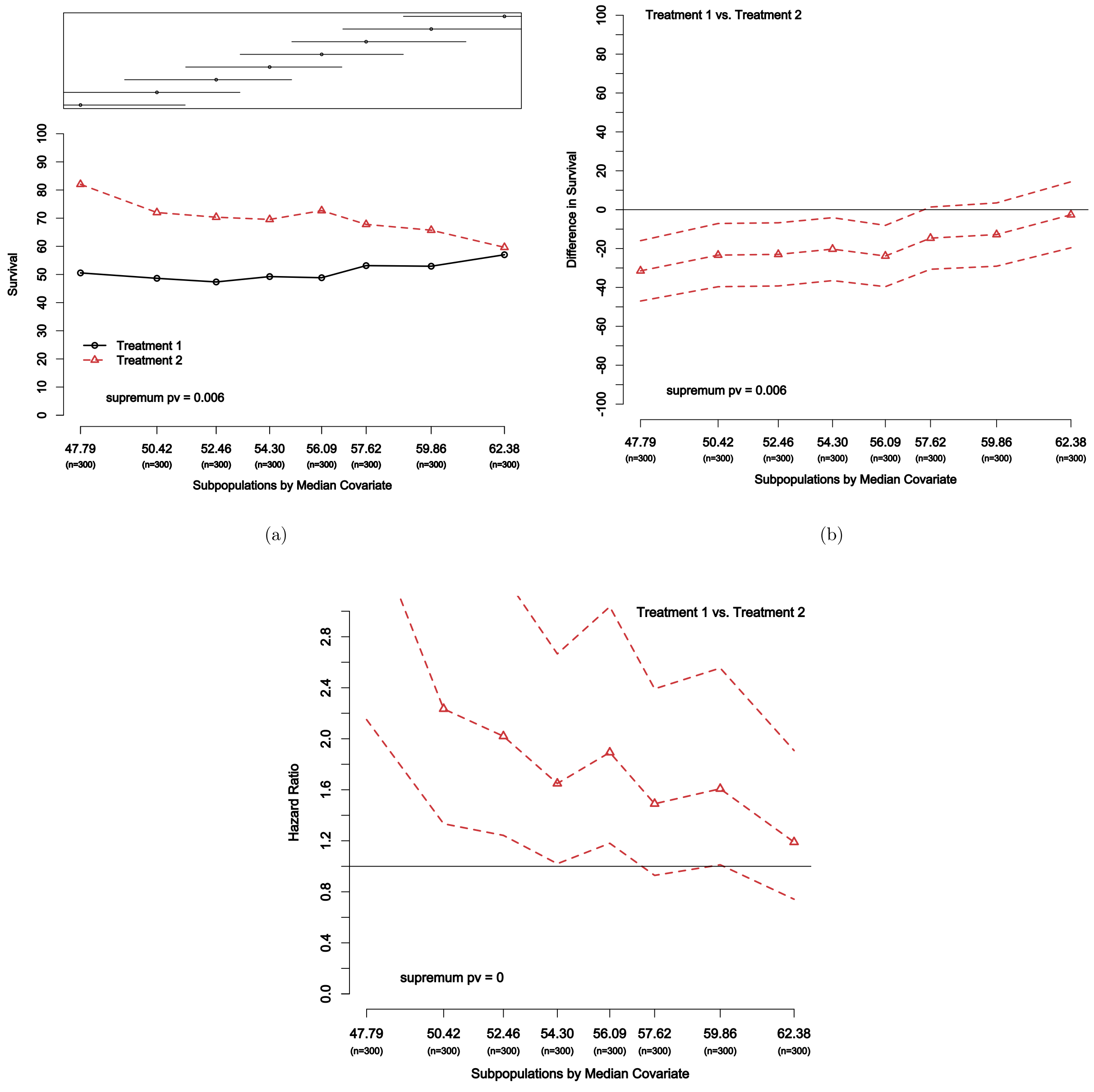
After the STEPP model has been specified and fitted, the results can be represented graphically using the `plot` method for `steppes` objects available in the package. The graphs produced are those for the outcome measure, absolute and relative treatment effect estimates. The `plot` method allows to set a number of graphical parameters whose details can be found in the package documentation. Here, we demonstrate the capabilities of the commands by concluding the example on the `simdataKM` simulated data and producing the three graphs described above (see Figure 2):

```
plot(res, subplot = TRUE, ylabel = ‘‘Survival’’,
  tlegend = c(‘‘Treatment 1’’, ‘‘Treatment 2’’),
  legend_diff = c(1, 2))
```

The picture confirms the same conclusions we reached by inspecting the numerical output, that is the data support the presence of a significant treatment-covariate interaction since most of the treatment effect estimates (on both scales) appear to be significantly different from zero across the subpopulations. In particular, the effect of treatment 2 relative to treatment 1 diminishes as the covariate value increases.

4 Application

In this section we present an application of the STEPP approach using the R version of the *stepp* package. The Stata code to perform the same analyses is available in the do file included in the supplementary material. The example involves the estimation of the cumulative incidence of breast cancer recurrence for postmenopausal women data collected as part of a randomized clinical trial with two treatment groups (Lazar et al., 2010, 2016). A second application, which is provided in the supplementary material available online, focuses on a clinical trial that aims at evaluating the effect of oral aspirin as a chemoprevention agent against colorectal adenomas (Yip et al., 2016). In this case the endpoint is binary (presence or not of adenomas), and thus a GLM analysis will be appropriate. In both examples we explore the presence of a treatment-covariate interaction using the STEPP tools described above.



(a)

(b)

(c)

Figure 2: `simdataKM` data. Plot of: (a) survival function (Kaplan-Meier) estimates, (b) absolute treatment effect measured as the difference in survival estimates, (c) relative treatment effect measured as the hazard ratio. Panels (b) and (c) also report the corresponding 95% confidence regions.

Here, we illustrate how to fit a STEPP model using the data from the BIG (Breast International Group) 1-98 study, an international, double-blind, phase III randomized clinical trial of 8010 postmenopausal women with hormone receptor–positive early invasive breast cancer. Patients were randomly assigned to receive one of four adjuvant endocrine therapy groups: letrozole, tamoxifen, or sequences of letrozole to tamoxifen or tamoxifen to letrozole. A first BIG

1-98 report (Breast International Group (BIG) 1-98 Collaborative Group et al., 2005; Coates et al., 2007) presented results indicating that letrozole significantly reduced the cumulative incidence of breast cancer recurrence as compared with tamoxifen in presence of two competing risks, second non-breast primary event and death prior to breast cancer recurrence. A known important prognostic factor of breast cancer is the Ki-67 biomarker, a nuclear protein present in cycling cells that represents an indicator of tumor proliferation and is associated with the extent of chemotherapy efficacy (Gerdes et al., 1983; Clahsen et al., 1999). In particular, high values of Ki-67 are predictive of a strong response to preoperative chemotherapy. The primary endpoint in the BIG 1-98 study was disease-free survival (DFS), which is defined as the time from randomization to the first of the following events: (1) invasive recurrence of breast cancer in local, regional, or distant sites, (2) a new invasive breast cancer in the contralateral breast, (3) any second nonbreast malignancy, or (4) death as a result of any cause.

Of the whole sample of patients in the trial, we consider here only women that were randomized to receive 5 years of monotherapy with either letrozole or tamoxifen who had tumors with centrally confirmed estrogen receptor expression and tumor material available for Ki-67 determination in a central laboratory. This selection resulted in a subset of 2685 women (Viale et al., 2008). In the following we focus on a STEPP analysis of 4-year DFS, where the 4-year time point was chosen to match the time point used in the original BIG 1-98 report. Furthermore, we use the recurrence of breast cancer as the primary outcome of interest, while non-breast second malignancies and deaths without recurrence were considered competing risks. Our aim here is that of investigating the presence of a treatment-covariate interaction by focusing on potential patterns of treatment effect for varying levels of the Ki-67 biomarker. The data for the BIG 1-98 study are included in the *stepp* package as an object called `bigCI`:

```
data(‘‘bigCI’’, package = ‘‘stepp’’)
```

The data set contains the following four columns:

- `trt`, the treatment indicator, where value `1` indicates the letrozole arm while value `2` refers to the tamoxifen arm;
- `time`, the survival times, that is the length of time in years from randomization to the occurrence of the event of interest, with mean and median values equal to 3.953 and 3.580 years respectively;
- `event`, the event indicator, where value `0` means no event occurred, value `1` indicates the recurrence of breast cancer and value `2` refers to competing risks (i.e., either non-breast second malignancies or death without recurrence);
- `ki67`, the Ki-67 biomarker measurements, from 0 to 90%, which show a strongly right-skewed distribution.

Table 1 shows the number of patients that experienced each one of the competing events. In the following, the presence of a treatment-covariate interaction is investigated with reference to the heterogeneity of treatment effects both in absolute and relative terms. Absolute treatment effects are measured as the difference in 4-year cumulative incidence of breast cancer recurrence, while relative effects are measured by the subdistribution hazard ratio (Fine and Gray, 1999; Collett, 2015).

The first step in a STEPP analysis is the choice of the windowing system to use. Given that we focus here only on the patients who experienced recurrence of breast cancer (i.e., cases with `event == 1`), the unit-based sliding window approach would result in sparse events within the overlapping subpopulations and a strong imbalance of events across treatment subpopulations

Table 1: `bigCI` data. Number of patients experiencing the different competing events by treatment group.

	`event`			
`trt`	`0` (censored)	`1` (relapse)	`2` (other events)	*total*
`1` (letrozole)	1239	73	49	1361
`2` (tamoxifen)	1143	123	58	1324
total	2382	196	107	2685

which in turn may cause instability in the estimation of the treatment effects. Therefore, we embrace the event-based sliding window approach. As for the choice of the e_1 and e_2 parameters, we decide to set $e_1 = 5$ and $e_2 = 15$. Clearly, a good suggestion is to perform some sensitivity analysis (not reported here) to assess the impact on the results of the choices for the e_1 and e_2 values. The code reported below sets the stage for the main analysis by generating the corresponding subpopulations:

```
swin <- new(‘‘stwin’’, type = ‘‘sliding_events’’, e1 = 5, e2 = 15)
subp <- new(‘‘stsubpop’’)
subp <- generate(subp, win = swin, covariate = bigCI$ki67,
  coltrt = bigCI$trt, trts = c(1, 2), coltype = bigCI$event)
summary(subp)
```

```
Window type: sliding_events
Number of events per subpopulation (eventspop e2): 15
Largest number of events in common among consecutive subpopulations (mineventspop e1): 5
Number of subpopulations created: 5

Subpopulation summary information
                         Covariate Summary                 Sample       Type 1 Events
 Subpopulation     Median        Minimum       Maximum        Size  Trt Group 1   Trt Group 2
       1             4.00         0.0000        7.0000         964           25            17
       2             9.00         7.0000       11.0000         618           15            16
       3            14.00        11.0000       17.0000         577           28            15
       4            20.00        17.0000       24.0000         399           27            15
       5            28.00        21.0000       90.0000         511           42            22
```

Note that in the case of event-based sliding windows it is not enough to provide the covariate information for generating the subpopulations but it is also required to include the information about the treatment indicator (`coltrt` argument), the list of treatments (`trts` argument) as well as the event type variable (`coltype` argument). Then, the `generate` method automatically uses the value `1` of `coltype` for the event of interest (here, the recurrence of breast cancer), the value `0` for no events and any other value for the competing events.

The output shows that 5 overlapping subpopulations have been generated with varying sample sizes but with a number of events per each treatment group that satisfies the requirements imposed by e_1 and e_2, thus guaranteeing a sufficient number of events in each subpopulation to reliably estimate the treatment effects.

We then proceed with the setup of the problem by defining new `steppes` and `stmodelCI` objects as shown in the following code:

```
res <- new(‘‘steppes’’)
modelCI <- new(‘‘stmodelCI’’, coltrt = bigCI$trt, coltime = bigCI$time,
  coltype = bigCI$event, trts = c(1, 2), timePoint = 4)
```

Next, we fit the STEPP model and compute the permutation test p-value by using the `estimate` and `test` methods as follows, where in the test we use 2500 Monte Carlo replications to achieve a reliable interpretation of the p-value (output partly omitted):

```
res <- estimate(res, subp, modelCI)
set.seed(101)
nperm <- 2500
res <- test(res, nperm)
print(res, estimate = TRUE, cov = FALSE, test = TRUE)
```

```
Sample size in treatment 1: 1361
Sample size in treatment 2: 1324
Total sample size (excluding missing data): 2685

[...]

Cumulative incidence differences at time point 4

trt 1 vs. trt 2
                        Cumulative
                        Incidence
     Subpopulation      Difference       Std. Err.
           1             -0.0246          0.0154
           2              0.0090          0.0175
           3             -0.0429          0.0232
           4             -0.0734          0.0342
           5             -0.0936          0.0356
        Overall          -0.0403          0.0112

Hazard ratio estimates
     Subpopulation        Log HR         Std. Err.      Hazard Ratio
           1             -0.4121          0.3087            0.66
           2              0.0573          0.3595            1.06
           3             -0.5331          0.3056            0.59
           4             -0.8547          0.3111            0.43
           5             -0.7405          0.2502            0.48
        Overall          -0.5635          0.1429            0.57

Supremum test results
trt 1 vs. trt 2
Interaction p-value based on cumulative incidence estimates: 0.0952
Interaction p-value based on hazard ratio estimates: 0.0604

Chi-square test results
Interaction p-value based on cumulative incidence estimates: 0.0968
```

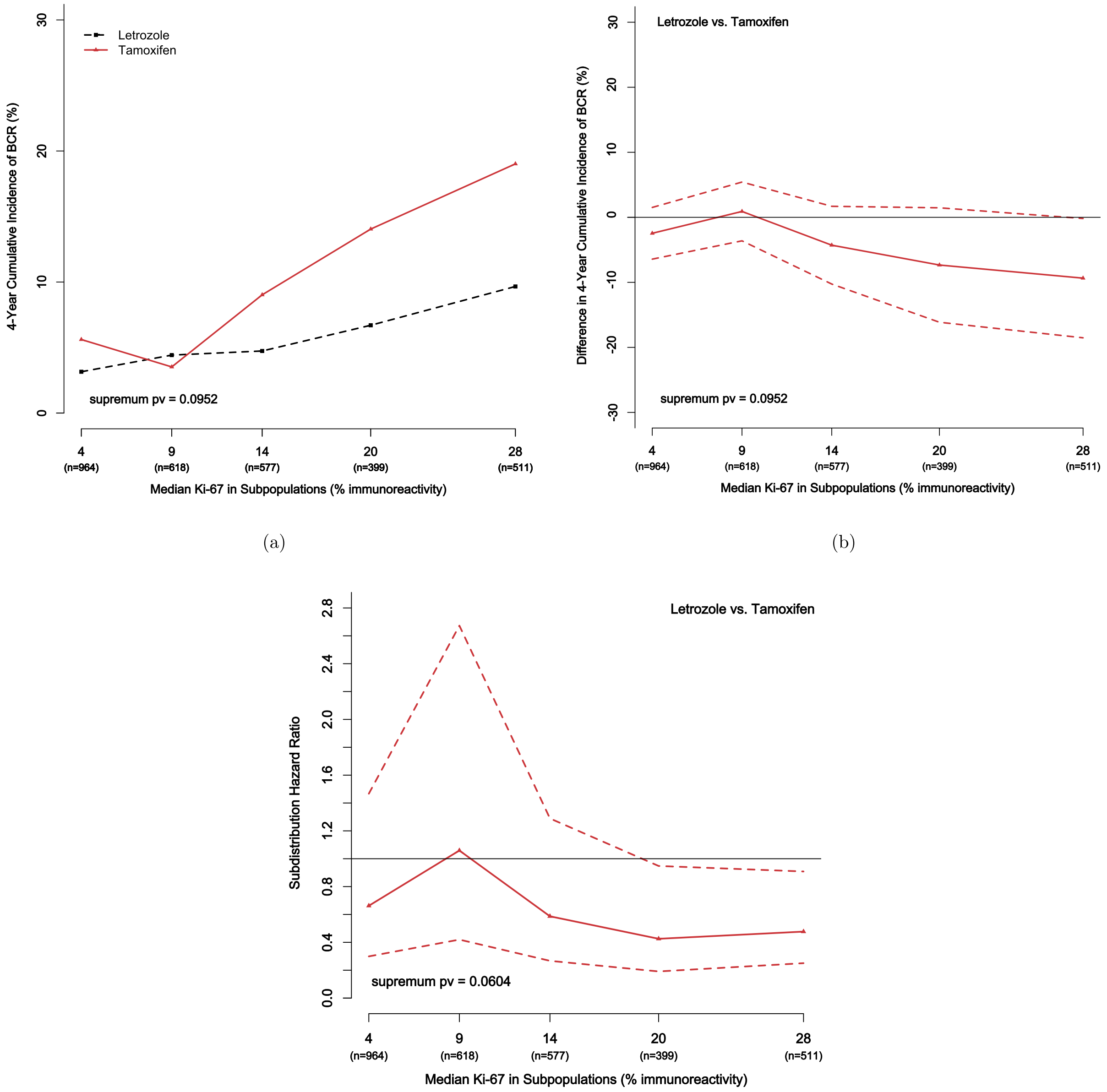
Finally, we produce the three STEPP graphs which report the cumulative incidence and treatment effect estimates together with the corresponding 95% confidence regions (see Figure 3):

```
plot(res, ylabel = ‘‘4-Year Cumulative Incidence of BCR (%)’’,
  xlabel = ‘‘Median Ki-67 in Subpopulations (% immunoreactivity)’’,
  tlegend = c(‘‘Letrozole’’, ‘‘Tamoxifen’’), legendy = 30,
  pointwise = FALSE, ylimit = c(0, 30, -30, 30, 0, 2.8),
  lsty = c(2, 1), marker = c(15, 17), ncex = 0.8, at = 8,
  legend_diff = c(1, 2))
```

These results show that the letrozole arm (reported as `treatment 1` in the output above) achieves a lower cumulative incidence in all but one subpopulation. In particular, subpopulations with high Ki-67 values show the largest treatment difference, reported as letrozole minus tamoxifen, thus indicating benefit for letrozole compared to tamoxifen (see Figure 3b). Although the differences provide evidence in favor of heterogeneity across subpopulations, the supremum test p-value does not support it. A similar conclusion is also achieved on the relative scale (see Figure 3c): the subdistribution hazard ratio estimates for breast cancer recurrence tend to be less than 1, thus confirming that letrozole is more beneficial than tamoxifen, but the supremum test p-value is still above the usual 0.05 level. Nonetheless, the confidence regions for the ratios provide support for a significant stronger effect of letrozole in the last two subpopulations, that is for subgroups of patients with the largest Ki-67 values. A sensitivity analysis, not reported here but included in the supplementary material, confirmed the same findings (see also Lazar et al., 2016). Therefore, contrary to the results in the official BIG 1-98 reports (Breast International Group (BIG) 1-98 Collaborative Group et al., 2005; Coates et al., 2007), the STEPP analysis allows to conclude that the superiority of letrozole over tamoxifen is more clearly observed for subgroups of women with large Ki-67 values.

5 Discussion

STEPP is a well established exploratory tool for identifying the presence of treatment-covariate interactions. We remark that STEPP is not meant to be used to determine specific cutpoints in the range of values of the covariate of interest, but rather to provide some indication regarding the ranges of values of the covariate of interest for which the treatment effect might have a particular behavior. The permutation test p-value indicating the statistical significance of treatment heterogeneity should always be presented together with the graphical representation of STEPP to avoid over-interpretation of the results. Notably, STEPP makes little or no modeling or distributional assumptions while estimating the treatment effects across values of the covariate, so that it is essentially non-parametric in nature. Although STEPP addresses the multiple testing issues in subpopulation analysis, as heterogeneity is evaluated globally with an omnibus statistical test, it only does so for one covariate. One still needs to address the multiple testing issue if several different covariates are examined. In addition, the STEPP approach does not consider the issue of post-hoc analysis as opposed to pre-specified analysis as well as issues of confounding if the analysis is based on retrospective exposure assessments as opposed to randomized treatments. As is the case with any exploration of subgroup treatment effects, hypothesis generating analyses should be distinguished from those intended to evaluate pre-specified hypotheses. Finally, we remark that the STEPP idea is not directly connected with the kernel conditional density estimator as illustrated in Hyndman et al. (1996) and implemented in the *hdrcde* R package (Hyndman et al., 2022), even if both approaches exploit the idea of generating subgroups based on the values of a covariate of interest.



(a)

(b)

(c)

Figure 3: `bigCI` data. Plot of: (a) cumulative incidence estimates for recurrence of breast cancer, (b) absolute treatment effect measured as the difference in the cumulative incidence estimates (letrozole minus tamoxifen; a value below zero suggests that letrozole is better), (c) relative treatment effect measured as the subdistribution hazard ratio (letrozole vs. tamoxifen; a value less than one suggests that letrozole is better). Panels (b) and (c) also report the corresponding 95% confidence regions.

In this paper, we presented the *stepp* packages that allow researchers to perform an analysis according to the STEPP methodology. The package is available for both the R and Stata software. Note that the current implementation of the packages has some limitations. In particular, it restricts the analysis to the comparison of two treatment groups. Furthermore, it allows for the study of only one covariate of interest. The extension to more than one covariate is not straightforward and may be conducted following different strategies. However, a simple approach to deal with the case of multiple covariates involves the creation of a composite score by combining the covariates of interest through some dimension reduction techniques such as principal component analysis. Indeed, the vast majority of the applications of STEPP to analyses of randomized trials have focused on comparing treatments with respect to absolute differences in recurrence risk with respect to a composite risk score incorporating multiple covariates. The most recent published example is the STEPP analysis for the Aphinity trial (Gelber et al., 2022), where the score was calculated using a Cox regression model that included some prespecified clinical characteristics. Other references of the use of composite risk indexes are Viale et al. (2008, 2011) and Regan et al. (2016).

Supplementary Material

The R and Stata scripts containing the code related to the examples discussed in the paper and a second application that involves a binary outcome are available in the supplementary material on the journal website.

Acknowledgments

We thank the editor and an associate editor for the useful comments that significantly improved the paper.

References

Bonetti M, Gelber RD (2000). A graphical method to assess treatment–covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine*, 19(19): 2595–2609.

Bonetti M, Gelber RD (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3): 465–481.

Bonetti M, Zahrieh D, Cole BF, Gelber RD (2009). A small sample study of the STEPP approach to assessing treatment–covariate interactions in survival data. *Statistics in Medicine*, 28(8): 1255–1268.

Thürlimann B, Keshaviah A, Coates AS, Mouridsen H, Mauriac L, et al. (Breast International Group (BIG) 1-98 Collaborative Group) (2005). A comparison of letrozole and tamoxifen in postmenopausal women with early breast cancer. *The New England Journal of Medicine*, 353(26): 2747–2757.

Clahsen P, van de Velde C, Duval C, Pallud C, Mandard AM, Delobelle-Deroide A, et al. (1999). The utility of mitotic index, oestrogen receptor and Ki-67 measurements in the creation of novel prognostic indices fornode-negative breast cancer. *European Journal of Surgical Oncology*, 25(4): 356–363.

Coates AS, Keshaviah A, Thürlimann B, Mouridsen H, Mauriac L, Forbes JF, et al. (2007). Five years of letrozole compared with tamoxifen as initial adjuvant therapy for postmenopausal women with endocrine-responsive early breast cancer: Update of study BIG 1-98. *Journal of Clinical Oncology*, 25(5): 486–492.

Collett D (2015). *Modelling Survival Data in Medical Research.* CRC Press.

Cox DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34(2): 187–202.

Fine JP, Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446): 496–509.

Foster JC, Taylor JMG, Rubert SJ (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30: 2867–2880.

Gelber RD, Wang XV, Cole BF, Cameron D, Cardoso F, Tjan-Heijnen V, et al. (2022). Six-year absolute invasive disease-free survival benefit of adding adjuvant pertuzumab to trastuzumab and chemotherapy for patients with early her2-positive breast cancer: a subpopulation treatment effect pattern plot (stepp) analysis of the aphinity (big 4-11) trial. *European Journal of Cancer*, 166: 219–228.

Gerdes J, Schwab U, Lemke H, Stein H (1983). Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. *International Journal of Cancer*, 31(1): 13–20.

Gray RJ (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, 16(3): 1141–1154.

Hayes AF (2017). *Introduction to mediation, moderation, and conditional process analysis*. Guilford Press.

Hyndman RJ, Bashtannyk DM, Grunwald GK (1996). Estimating and cvisualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5: 315–336.

Hyndman RJ, Einbeck J, Wand MP (2022). hdrcde: Highest density regions and conditional density estimation. R package version 3.4.

Imai K (2017). *Quantitative social science. An introduction*. Princeton University Press.

Lagakos S (2006). The challenge of subgroup analysis – Reporting without distorting. *The New England Journal of Medicine*, 354: 1667–1669.

Lazar A, Bonetti M, Cole BF, Yip WK, Gelber RD (2016). Identifying treatment effect heterogeneity in clinical trials using subpopulations of events: STEPP. *Clinical Trials*, 13(2): 169–179.

Lazar A, Cole BF, Bonetti M, Gelber RD (2010). Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: Subpopulation treatment effect pattern plot. *Journal of Clinical Oncology*, 28(29): 4539–4544. PMID: 20837942.

Li J, Zhao L, Tian L, Cai T, Claggett B, Callegaro A, et al. (2015). A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative, controlled clinical studies. *Biometrics*, 72: 877–887.

Pesarin F (2001). *Multivariate permutation tests*. John Wiley & Sons.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, et al. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer*, 35(1): 1–39.

Pocock S (2008). More on subgroup analysis in clinical trials. *The New England Journal of Medicine*, 358: 2076–2077.

Regan MM, Francis PA, Pagani O, Fleming GF, Walley BA, Viale G, et al. (2016). Absolute benefit of adjuvant endocrine therapies for premenopausal women with hormone receptor–positive, human epidermal growth factor receptor 2–negative early breast cancer: text and soft trials. *Journal of Clinical Oncology*, 34(19): 2221–2231. PMID: 27044936.

Royston P, Altman DG (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 43: 429–467.

Royston P, Ambler G (1998). sg81: Multivariable fractional polynomials. *Stata Technical Bulletin*, 43: 24–32.

Royston P, Sauerbrei W (2004). A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine*, 23: 2509–2525.

Royston P, Sauerbrei W (2007). Multivariable modeling with cubic regression splines: A principled approach. *Stata Journal*, 7(1): 45–70.

Royston P, Sauerbrei W (2008). *Multivariable model-building.* John Wiley & Sons.

Royston P, Sauerbrei W (2009). Two techniques for investigating interactions between treatment and continuous covariates in clinical trials. *Stata Journal*, 9(2): 230–251.

Simon R (2002). Bayesian subset analysis: Application to studying treatment-by-gender interactions. *Statistics in Medicine*, 21 2909–2916.

Simon R, Dixon DO, Freidlin BA (1995). *A Bayesian model for evaluating specificity of treatment effects in clinical trials.* Kluwer Academic Publications.

VanderWeele TJ (2015). *Explanation in causal inference.* Oxford University Press.

Viale G, Giobbie-Hurder A, Regan MM, Coates AS, Mastropasqua MG, Dell'Orto P, et al. (2008). Prognostic and predictive value of centrally reviewed ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: Results from breast international group trial 1-98 comparing adjuvant tamoxifen with letrozole. *Journal of Clinical Oncology*, 26(34): 5569–5575.

Viale G, Regan M, Dell'Orto P, Mastropasqua M, Maiorano E, Rasmussen B, et al. (2011). Which patients benefit most from adjuvant aromatase inhibitors? Results using a composite measure of prognostic risk in the big 1-98 randomized trial. *Annals of Oncology*, 22(10): 2201–2207.

Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM (2007). Statistics in medicine—Reporting of subgroup analyses in clinical trials. *The New England Journal of Medicine*, 357(21): 2189–2194.

Yip WK, Bonetti M, Cole BF, Barcella W, Wang XV, Lazar A, et al. (2016). Subpopulation treatment effect pattern plot (STEPP) analysis for continuous, binary, and count outcomes. *Clinical Trials*, 13(4): 382–390.

Zhao L, Tian L, Cai T, Claggett B, Wei LJ (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539.