STATISTICS IN MEDICINE
Statist. Med. (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3524

A small sample study of the STEPP approach to assessing treatment–covariate interactions in survival data

Marco Bonetti[1,*,†], David Zahrieh[2], Bernard F. Cole[3] and Richard D. Gelber[2,4]

[1]*Department of Decision Sciences, Bocconi University, Via Röntgen 1, Milan 20136, Italy*
[2]*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, U.S.A.*
[3]*Department of Mathematics and Statistics, University of Vermont, 16 Colchester Avenue, Burlington, VT 05401, U.S.A.*
[4]*Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, U.S.A.*

SUMMARY

A new, intuitive method has recently been proposed to explore treatment–covariate interactions in survival data arising from two treatment arms of a clinical trial. The method is based on constructing overlapping subpopulations of patients with respect to one (or more) covariates of interest and in observing the pattern of the treatment effects estimated across the subpopulations. A plot of these treatment effects is called a subpopulation treatment effect pattern plot. Here, we explore the small sample characteristics of the asymptotic results associated with the method and develop an alternative permutation distribution-based approach to inference that should be preferred for smaller sample sizes. We then describe an extension of the method to the case in which the pattern of estimated quantiles of survivor functions is of interest.

KEY WORDS: treatment–covariate interaction; clinical trials; permutation-based inference; survival analysis

1. INTRODUCTION

The study of interactions between treatment effect and variables of interest is a fundamental part of the analysis of data arising from clinical trials. Such study may help identify subgroups of patients for whom treatment effect is largest (or smallest, perhaps even negative) with clear implications

*Correspondence to: Marco Bonetti, Department of Decision Sciences, Bocconi University, Via Röntgen 1, Milan 20136, Italy.
†E-mail: marco.bonetti@unibocconi.it

Contract/grant sponsor: The United States National Cancer Institute; contract/grant number: CA-75362
Contract/grant sponsor: The Italian Ministry for Research (MIUR) protocol; contract/grant number: 2007AYHZWC

Received 9 February 2008
Accepted 4 December 2008

for the design of later trials and for clinical practice, since the tailoring of treatment decisions to the individual patient becomes possible. Since many clinical trials are extremely expensive and time consuming, it is also important that the largest amount of information be obtained from them. One could argue that it would indeed not be ethical to do otherwise.

One approach to the study of treatment–covariate interactions that has recently been proposed is based on the analysis of treatment effects across overlapping subpopulations of patients defined with respect to a covariate of interest. The method is called subpopulation treatment effect pattern plot (STEPP) and it was introduced in [1, 2]. It has been applied to clinical trials conducted by the International Breast Cancer Study Group (IBCSG), (see [3–7]) in its implementation based on measuring treatment effect with the difference in survival probabilities between two treatment groups at a fixed time point.

In Section 2 we briefly review the STEPP approach. In Section 3 we present the results of a simulation study designed to explore the small sample properties of the test and the simultaneous confidence band associated with the method. In that section we describe an alternative permutation distribution approach to inference, which should be used for the smaller sample sizes. In Section 4 we describe the details of an implementation of STEPP to the case in which treatment effect is based on a quantile of the survivor function in the two treatment arms. We summarize our findings and recommendations in Section 5.

2. THE STEPP

Consider n patients in a clinical trial in which they are randomized to one of two treatments and suppose that on all patients a baseline covariate $Z \in [z_{\min}, z_{\max}] \subset \Re$ is observed. The STEPP approach has been introduced to explore the possible presence of an interaction effect between treatment and the covariate Z. The approach consists of defining overlapping subpopulations of patients defined with respect to Z and computing an estimate of treatment effect within each subpopulation. Here we focus on the case in which Z is one-dimensional, but this is not strictly necessary. We consider the case in which the subpopulations $(\mathscr{P}_j, j = 1, \ldots, K)$ are constructed according to a 'sliding window' pattern, as this has proved most useful in applications. The sliding window pattern consists of assigning a patient i to subpopulation $\mathscr{P}_j$ when $z_i \in [l_j, u_j]$, where the two non-decreasing sets of numbers $\{l_j\}$ and $\{u_j\}$ are such that $l_j \in [z_{\min}, z_{\max}], u_j = \inf\{u \geqslant l_j | P_n(l_j < Z \leqslant u) \geqslant p)\}$ for some fixed $p \in (0, 1)$, with P_n the empirical distribution of Z in the data. The sets of values $\{l_j\}$ and $\{u_j\}$ can be constructed by assigning the values of two parameters $r_1 < r_2 < n$ and then by defining $\mathscr{P}_1$ as containing patients having values of Z between the smallest observed Z value and the $p = (r_2/n) \times 100$-th percentile of $P_n(z)$, which defines u_1. Equivalently, u_1 is the smallest number such that at least r_2 patients fall in $\mathscr{P}_1$. Subpopulation $\mathscr{P}_2$ is then defined by choosing l_2 as the smallest value such that at most r_1 patients fall between l_2 and u_1. Then, u_2 is defined as the smallest number such that $\mathscr{P}_2$ contains at least r_2 patients. This process is repeated until the last possible population is defined.

Within each subpopulation $\mathscr{P}_j$ an estimate $\widehat{\theta}_j$ of treatment effect is produced and the plot of these estimates with respect to the median value of Z within each subpopulation is a STEPP plot. A simultaneous confidence band can be produced if the joint distribution of the treatment effects can be estimated. If, for example, one can show that the vector of the estimates $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ of the treatment effects $(\theta_1, \ldots, \theta_K)$ is approximately normal with mean $(\theta_1, \ldots, \theta_K)$ and a variance–covariance matrix Σ that can be estimated consistently from the data, then a rectangular

DOI: 10.1002/sim

simultaneous confidence band of level 95 per cent (say) can be constructed by solving numerically the equation in γ $P(\bigcap_{j=1}^{K}\{\theta_j \in \widehat{\theta}_j \pm \gamma(1.96)[\widehat{\mathrm{var}}(\widehat{\theta}_j)]^{1/2}\}) = 0.95$ for a sample of random variables generated from the estimated asymptotic distribution of the estimates. The parameter γ represents the widening of the marginal confidence intervals that is necessary to produce the desired simultaneous coverage of the band.

To test the null hypothesis of no interaction between the covariate of interest (i.e. across subpopulations) and treatment effect one can use the test statistic

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

where $\widehat{\theta}_{\mathrm{ALL}}$ is the measure of treatment effect computed on all patients in the study. The distribution of T can be estimated by sampling repeatedly from the estimated asymptotic distribution of $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K, \widehat{\theta}_{\mathrm{ALL}})$, and a Monte Carlo p-value can thus be produced.

Bonetti and Gelber [2] discuss the implementation of the approach for the case in which $\widehat{\theta}_j$ consists of the estimated difference in survival at a fixed time point t^* between two arms A and B, i.e. $\widehat{\theta}_j = \widehat{S}_{A,j}(t^*) - \widehat{S}_{B,j}(t^*)$, with $\widehat{S}(t)_{G,j}$ the Kaplan–Meier estimator of survival at time t within treatment group G inside subpopulation $\mathscr{P}_j$.

3. A STUDY OF THE FINITE SAMPLE PROPERTIES, AND AN ALTERNATIVE APPROACH TO INFERENCE

We explored the small sample properties of the survival difference implementation of STEPP. In particular, we evaluated the accuracy in the recovery of the type I error α for the test based on the statistic T, and the coverage of the confidence band around the STEPP plot. For the same T-based test we also estimated the power under a series of alternative hypotheses. The power of the STEPP test was also compared with the power of the test for the presence of a non-zero treatment–covariate interaction using a Cox proportional hazards model that included the terms treatment arm (Tx), Z, and $Z \times Tx$. Calculations were performed using a combination of the R programming language [v2.6.2; available at www.r-project.org] and of the C++ compiler g++ [v2.95.2; available at gcc.gnu.org], on a SPARC solaris 8 machine.

3.1. Recovery of the type I error probability of the test and coverage of the confidence band around the STEPP plot

Under the null hypothesis, patient survival times were randomly generated from an exponential distribution such that $S(4) = 0.1$, 0.5, and 0.9, where $S(.)$ is the survival function, assuming no treatment effect anywhere. Patients entered the study uniformly over 5 years with two additional years of follow up. At 7 years from the opening of accrual, administrative censoring was applied to the survival times. We randomly assigned to each patient one of two treatment groups (A, B) in a 1:1 ratio and a continuous covariate Z, where $Z \sim \mathrm{N}(55, 25)$. Overlapping subpopulations were created with respect to Z using the sliding window approach. Within each subpopulation, survival at 4 years was estimated for each treatment group. In addition, an overall estimate of survival at 4 years across all subpopulations was obtained for each treatment group.

DOI: 10.1002/sim

We estimated the covariance matrix $\Sigma = \Sigma_A + \Sigma_B$ of the vector of treatment effects $[\widehat{\theta}_1, \ldots, \widehat{\theta}_K, \widehat{\theta}_{\mathrm{ALL}}]$, where $\widehat{\theta}_j = \widehat{S}_{A,j}(4) - \widehat{S}_{B,j}(4)$, $j = 1, \ldots, K$ and $\widehat{\theta}_{\mathrm{ALL}} = \widehat{S}_A(4) - \widehat{S}_B(4)$. The matrices Σ_A and Σ_B were estimated as described in [2].

For each of 300 simulations of sample size n, survival data were generated, the subpopulations were constructed based on the parameters r_1 and r_2, and the covariance matrix Σ estimated. For the confidence band, within each simulation 1000 vectors of treatment effects of size K were sampled from the estimated marginal normal distribution of $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$. We then numerically solved for the parameter value γ, thus obtaining the confidence band. We estimated the true probability that the (random) confidence band contained the vector of the true treatment effects by counting the number of bands out of the 300 that did contain the zero vector of length K.

For the test statistic T the vector of treatment effects was transformed linearly to obtain the covariance matrix of the centered treatment effects $\widehat{\theta}_j - \widehat{\theta}_{\mathrm{ALL}}$, $j = 1, \ldots, K$. One thousand vectors were generated from the asymptotic distribution with that covariance matrix and the null distribution of the test statistic T was estimated from them. The proportion of the simulated values of T greater than the T statistic observed in the data estimated the p-value for that simulated data set. The number of times out of 300 simulated data sets for which the p-value was less than α was the estimated effective α level of the test. For some cases (in particular, for the cases $S(4) = 0.90$ and $S(4) = 0.10$ with $n = 100$ and $n = 200$) simulated data sets that did not provide estimates at t^* were discarded. This reflects the fact that in practice one would only conduct a STEPP analysis when the value of $S(t^*)$ can indeed be estimated within the subpopulations.

Tables I and II show the estimated α level of the test T for a treatment–covariate interaction and the estimated coverage of the 95 per cent confidence band around the STEPP plots, under a variety of parameter values $S(4), r_1, r_2, n$, and α. Note that the parameter values ($r_1 = 30$, $r_2 = 40$) produce subpopulations that contain approximately 20 patients in each treatment arm and that estimation

Table I. Estimated α level of the test for interaction based on the T statistic. Results are based on 300 simulations of sample size n.

				α		
$S(4)$	n	r_1	r_2	0.01	0.05	0.10
0.1	100	30	40	0.03	0.11	0.26
	200	60	80	0.03	0.11	0.19
	500	150	200	0.01	0.11	0.17
	1000	300	400	0.00	0.04	0.07
0.5	100	30	40	0.05	0.12	0.19
	200	60	80	0.01	0.08	0.13
	500	150	200	0.01	0.06	0.12
	1000	300	400	0.00	0.05	0.09
0.9	100	30	40	0.01	0.10	0.20
	200	60	80	0.01	0.05	0.10
	500	150	200	0.01	0.05	0.12
	1000	300	400	0.01	0.05	0.11

The standard errors for the three α levels 0.01, 0.05, and 0.10 are equal to 0.006, 0.013, and 0.017, respectively (based on the normal approximation). The distribution of the covariate of interest Z is N(55, 25). Entry times follow a Uniform(0, 5) distribution, and follow up ends at 7 years from start of accrual.

DOI: 10.1002/sim

Table II. Coverage of the 95 per cent confidence band around point estimates for the population-specific estimated treatment effects $\widehat{\theta}_j$.

			$S(4)$				
n	r_1	r_2	90 per cent	80 per cent	50 per cent	20 per cent	10 per cent
100	30	40	0.99	0.84	0.86	0.91	0.97
200	60	80	0.98	0.92	0.92	0.89	0.96
500	150	200	0.94	0.96	0.94	0.93	0.90
1000	300	400	0.93	0.95	0.95	0.93	0.93

Results refer to the survival estimates $\widehat{S}(t^*)$ (with $t^*=4$ years) and are based on 300 simulations of sample size n. The standard error for each probability is equal to 0.013 (based on the normal approximation). The distribution of the covariate of interest Z is N(55, 25). Entry times follow a Uniform(0, 5) distribution, and follow up ends at 7 years from start of accrual.

is likely to be problematic in that case. In the simulated data that were used to produce Table II, the value of γ that we obtained for the confidence bands was roughly constant and equal to 1.3.

It seems clear from Table I that there is a tendency to an inflation of the type I error probability for n as high as 500. This phenomenon becomes worse for smaller sample sizes and it occurs across the range of null parameters that we considered for the survival distribution. This dangerous behavior can clearly lead to false rejections of the null hypothesis. It is not observed when $n=1000$, in which case α is recovered appropriately.

The coverage of the confidence band reported in Table II is satisfactory for sample sizes equal to 500 or more. In general, the coverage tends to deteriorate when the survival pattern is extreme, as when estimating a very small survival proportion, but these situations can be avoided by choosing the time point t^* appropriately. For sample sizes smaller than 500, the results obtained are not satisfactory.

3.2. Power comparisons

We considered seven different scenarios to gain some information on the power performance of the STEPP test. All scenarios are based on the exponential assumption for the time to event distribution conditional on the value of the continuous covariate Z, i.e. $\lambda(t|Z=z)=\lambda(z)$. We compared the estimated power of the STEPP test for interaction with the test on the interaction parameter in a Cox model. We based such a comparison on the assumption that the latter is probably the most commonly performed analysis when one suspects the presence of a treatment–covariate interaction. In the simulated data, the covariate Z was generated from a normal distribution. Here we used $Z\sim\mathrm{N}(55,49)$ to ensure that results be in useful ranges for the comparison of the power characteristics of the two tests.

In all scenarios, for one of the two arms (arm 1) we assumed the hazard function $\lambda_1(t;z)=\lambda_1(z)=\lambda_1$ for the time to event random variable. It follows that within a subpopulation $\{Z\in[z_\mathrm{L},z_\mathrm{U}]\}$ the conditional survival function at t^* is equal to $S(t^*|Z\in[z_\mathrm{L},z_\mathrm{U}])=\exp\{-\lambda_1 t^*\}$, with λ_1 such that $S_1(t^*)=0.4$ at $t^*=4$ years.

Scenario 1: The first scenario consists of assigning to treatment group 2 the conditional hazard function $\lambda_2(t;z)=\lambda_2(z)=\beta_0+\beta_1 z$. The parameters β_0 and β_1 were chosen to have $\lambda_2(40)=(0.2)\lambda_1$ and $\lambda_2(70)=\lambda_1$. In particular, $\beta_0=-(65/75)\lambda_1$ and $\beta_1=(2/75)\lambda_1$. The conditional survival

DOI: 10.1002/sim

function $S(t^*|Z\in[z_L, z_U])$ at t^* corresponding to the model $\lambda(z)=\beta_0+\beta_1 z$ within each subpopulation $\{z_L, z_U\}$ is equal to

$$P(T>t^*|Z\in[z_L, z_U])=\frac{\exp\{-\beta_0 t^*\}\int_{z_L}^{z_U}\exp\{-\beta_1 t^* u\}f_Z(u)\,du}{F_Z(z_U)-F_Z(z_L)} \tag{2}$$

Scenario 2: In the second scenario the conditional hazard $\lambda_2(t;z)=\lambda_2(z)$ follows a logistic function defined to have $\lim_{z\to\infty}\lambda_2(z)=\lambda_1$ and $\lim_{z\to-\infty}\lambda_2(z)=\lambda_1-\Delta$ for a fixed (and smaller than λ_1) value of Δ. Also, we designed this scenario so that it would be $\lambda_2(60)=\lambda_1-\Delta/2$ and $\lambda_2(62)=\lambda_1-\Delta/10$. One can easily show that these constraints imply that

$$\lambda_2(z)=(\lambda_1-\Delta)+\Delta\frac{\exp\{\alpha_0+\alpha_1 z\}}{1+\exp\{\alpha_0+\alpha_1 z\}} \tag{3}$$

with $\alpha_0=-30\log(9)$, $\alpha_1=\log(9)/2$, and $\Delta=\lambda_1/2$.

Scenarios 3–7: The remaining scenarios consist of modifying the conditional hazard function to have $\lambda_2(t;z)=\lambda_1[1-\beta\phi(z;\mu,\sigma^2)]$ with ϕ the normal density function, so that it has a peak at $z=\mu$ and with $\beta=10$.

By varying the values of the parameters μ and σ^2 we obtain the five alternative models (3) $\mu=55, \sigma=6$; (4) $\mu=60, \sigma=5$; (5) $\mu=60, \sigma=9$; (6) $\mu=58, \sigma=6$; (7) $\mu=58, \sigma=7$.

For a fixed value $Z=z$ one has that $S(t|z)=\exp\{-\Lambda(t|z)\}$ with $\Lambda(t|z)=\int_0^t\lambda(u|z)\,du$ the cumulative hazard function. In the exponential model being considered here we have immediately that $S(t|z)=\exp\{-\lambda(z)t\}$. Figure 1 shows the hazard functions and the proportion surviving at $t^*=4$ for the two arms as a function of the covariate Z. The three displays refer to scenarios (1), (2), and (3), the last one taken as an example of scenarios (3)–(7).

Table III shows the results of the power study for the seven scenarios and for varying values of n (and of r_1, r_2). For each setting, the empirical power of the two 0.05-level tests of the null hypothesis of no treatment–covariate interaction is reported. The tests are based on the test statistic T for STEPP and on the parameter corresponding to the interaction term for the proportional hazards model.

These power results should be considered reliable since they are based on larger sample sizes. The results suggest that the benchmark Cox model test for the interaction produces higher powers in the two Scenarios (1) and (2), in which the alternative hypothesis consists of a monotone (with respect to Z) hazards ratio between the two arms. However, when the group of patients for whom the treatment effect is largest is somewhere in the middle of the patient population, the STEPP method produces similar or higher powers than the Cox test for all sample sizes considered. It should be pointed out that Scenarios (3)–(7) do not produce an extremely concentrated group of patients for whom the treatment effect is largest, as could be obtained if one for example were to set σ to values smaller that 5. Should that be done, then one would expect even better performance of STEPP relative to the Cox model.

The fifth column in Table III refers to the procedure described later in Section 4. This procedure performs similarly to the T test described above.

Clearly, no simulation exercise can cover all the possible combinations of parameter values and possible alternative models. For example, the number and relative sizes of the subpopulations could be modified (here we used $r_2/n=0.4$ and 7 subpopulations), as could the model chosen for the null hypothesis, the various alternative scenarios and for the power part the form of the alternative model and its implications on the kind of interactions that one considers.

DOI: 10.1002/sim

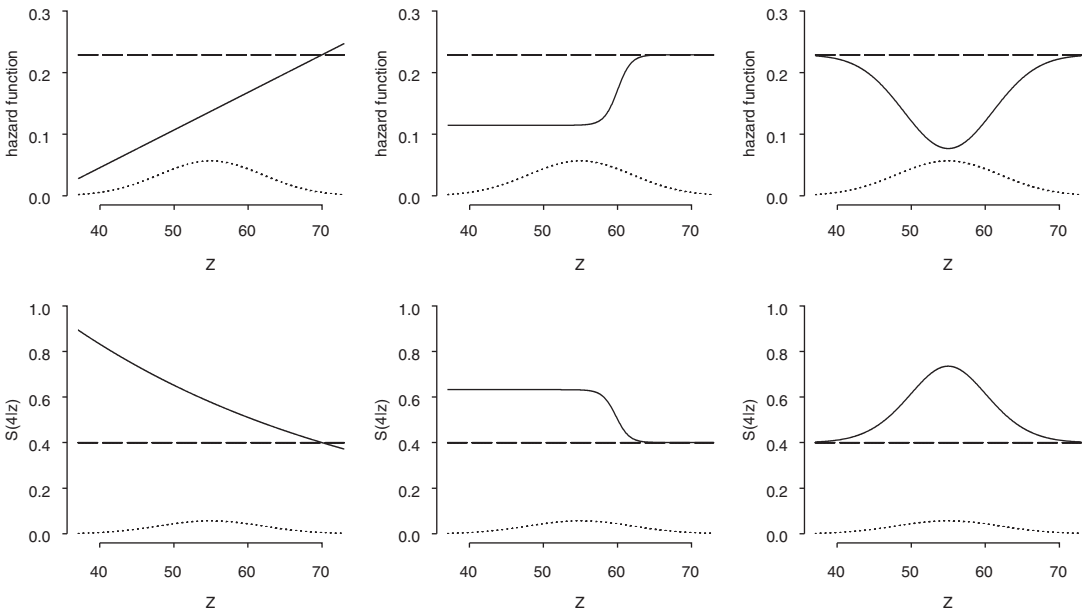

Figure 1. Three representative scenarios for the power comparisons are reported in Table III. The three scenarios are (left to right) (1), (2), and (3), as described in Section 3. The top graphs show the hazard as a function of the covariate Z ($\lambda_1(z)$ and $\lambda_2(z)$), and the bottom graphs the proportion surviving at $t^*=4$, also as a function of Z ($S_1(t^*|z)$ and $S_2(t^*|z)$). Dashed lines refer to the first treatment group and solid curves to the second treatment group. The dotted curve is a sketch of the density function of the covariate Z.

We do feel, however, that the information gained in this small study is sufficient for one to appreciate the fact that STEPP may be effective at detecting interactions of uncommon shapes. However, the study also indicates quite clearly the poor small-sample behavior of the procedures. As a consequence, we suggest using a permutation distribution-based procedure for inference.

3.3. A permutation distribution approach to inference for STEPP

Note that under the null hypothesis of no treatment–covariate interaction, one is allowing for a covariate effect on survival, but that such effect is the same across the two arms with respect to the definition of interaction used here: $S_A(t^*|z)-S_B(t^*|z)$ should be constant (and in particular equal to $S_A(t^*)-S_B(t^*)$). Following the approach taken in [8], this suggests the possibility of using a general permutation approach to inference in which one permutes the covariate values across the patients within each treatment group and then re-computes on the permuted samples the test statistic T, where the variances are estimated from the permuted samples. This produces a sample from the permutation distribution of T to be used for testing. (For a general reference on permutation tests see also [9].)

We repeated the simulation experiment reported in Table I using this alternative permutation distribution approach. Three hundred data sets of size n were simulated and within each simulated data set the subpopulations were constructed and the test statistic T calculated. In particular, if

DOI: 10.1002/sim

Table III. Estimated powers of the (0.05 level) test for treatment–covariate interaction.

Scenario	r_1	r_2	n	Est. median surv.	$\widehat{S}(4)$	Cox
1	150	200	500	0.30	0.28	0.72
	300	400	1000	0.65	0.52	0.95
	450	600	1500	0.85	0.73	0.99
2	150	200	500	0.26	0.27	0.41
	300	400	1000	0.47	0.44	0.75
	450	600	1500	0.71	0.60	0.89
3	150	200	500	0.20	0.37	0.06
	300	400	1000	0.47	0.65	0.06
	450	600	1500	0.70	0.83	0.07
4	150	200	500	0.70	0.78	0.77
	300	400	1000	0.94	0.99	0.97
	450	600	1500	1.00	0.99	1.00
5	150	200	500	0.13	0.16	0.20
	300	400	1000	0.20	0.21	0.29
	450	600	1500	0.29	0.31	0.42
6	150	200	500	0.29	0.43	0.31
	300	400	1000	0.72	0.70	0.51
	450	600	1500	0.91	0.91	0.66
7	150	200	500	0.24	0.25	0.21
	300	400	1000	0.53	0.43	0.31
	450	600	1500	0.63	0.59	0.49

Results refer to the test T when treatment effect is based on the survival estimate ('$\widehat{S}(4)$'), to the test of the interaction term in a proportional hazards model ('Cox'), and to the test T when treatment effect is based on the estimated median survival ('Est. median surv.', see Section 4). Results are based on 300 simulations of sample size n. Probability of survival at 4 years is $S(4) = 0.4$. The distribution of the covariate of interest Z is N(55, 49). Entry times follow a Uniform$(0, 5)$ distribution and follow up ends at 7 years from start of accrual.

for any of the simulated data sets the survival time within each treatment arm and/or within each treatment arm and subpopulation combination was not estimable at t^*, another simulated data set was generated. This process continued until the 300 simulated data sets were generated. For each of these 300 simulated data sets, 1000 permutations of survival time and survival status were performed within each arm. (Here, too, another permuted sample was generated if the Kaplan–Meier estimator did not provide an estimate at t^*, until all 1000 permutated data sets were generated for each of the 300 simulated data sets.) For each of the 1000 permuted data sets the test statistic T was calculated, thus obtaining an estimate of its permutation distribution. The critical value for rejection, say at the 0.05 level, is the 95th percentile of these 1000 T statistics. The p-value for the test was calculated as the proportion of the 1000 T statistics greater than the value of T observed on the simulated data set. Results are shown in Table IV and they clearly indicate an improvement in performance compared with the asymptotic results shown in Table I.

4. AN EXTENSION: STEPP FOR QUANTILES OF A SURVIVAL FUNCTION

The survival estimate implementation of STEPP can be extended to the case in which the treatment effect is defined as the difference in an estimated quantile of the survival function between two

DOI: 10.1002/sim

Table IV. Estimated α level of the test for interaction based on the T statistic for the STEPP implementation based on the difference in estimated survival at 4 years.

				α		
$S(4)$	n	r_1	r_2	0.01	0.05	0.10
0.1	100	30	40	0.00	0.04	0.12
	200	60	80	0.02	0.07	0.13
	500	150	200	0.01	0.05	0.08
	1000	300	400	0.00	0.07	0.09
0.5	100	30	40	0.01	0.07	0.11
	200	60	80	0.01	0.07	0.13
	500	150	200	0.01	0.05	0.08
	1000	300	400	0.01	0.05	0.09
0.9	100	30	40	0.03	0.09	0.14
	200	60	80	0.01	0.05	0.10
	500	150	200	0.01	0.06	0.11
	1000	300	400	0.02	0.07	0.12

Results are obtained from 300 simulations of sample size n and they are based on the permutation approach to inference. The standard errors for the three α levels 0.01, 0.05, and 0.10 are equal to 0.006, 0.013, and 0.017, respectively (based on the normal approximation). The distribution of the covariate of interest Z is N(55, 25). Entry times follow a Uniform(0, 5) distribution and follow up ends at 7 years from start of accrual.

arms. In particular, the difference in median survival between two groups is an intuitive and widely used measure of treatment effect. Clearly, the computation of the median survival may not always be the best choice (nor be possible) if the survival curve decreases too slowly relative to the length of follow up available, in which case one may use a higher percentile than the 50th without any major changes in the approach that we now describe.

Consider first treatment group A. Let $\theta_{A,1}, \ldots, \theta_{A,K}$ be the survival medians within each of the K subpopulations. Call $(\tilde{\theta}_{A,1}, \ldots, \tilde{\theta}_{A,K})$ the estimated median survival obtained by inversion of $(\widehat{S}_{A,1}(\cdot), \ldots, \widehat{S}_{A,K}(\cdot))$, the vector of the Kaplan–Meier estimators of the marginal survival functions $S_{A,j}(\cdot)$, $j=1, \ldots, K$. Specifically,

$$\tilde{\theta}_{A,j} = \sup\{t : \widehat{S}_{A,j}(t) > \tfrac{1}{2}\}, \quad j=1, \ldots, K \tag{4}$$

One may apply the resampling method introduced in [10] to obtain an estimate of the joint distribution of the estimated quantiles. Under the assumption of a common median survival η across the K subpopulations, one has that

$$\sqrt{n}\begin{bmatrix} (\widehat{S}_{A,1}(\eta) - 1/2) \\ \vdots \\ (\widehat{S}_{A,K}(\eta) - 1/2) \end{bmatrix} \xrightarrow{\mathrm{d}} N_K(0, \Lambda_A) \tag{5}$$

DOI: 10.1002/sim

and that the covariance matrix Λ_A can be estimated consistently (see [2]). Call the estimated covariance matrix $\widehat{\Lambda}_A$. One can generate a large number M of multivariate samples $(u^l_{A,1},\ldots,u^l_{A,K})$, $l=1,\ldots,M$ from the distribution $N_K(0,\widehat{\Lambda}_A/n)$, and for each sample solve the equations

$$\begin{aligned}\widehat{S}_{A,1}(\theta^l_{A,1})-\tfrac{1}{2}-u^l_{A,1}&=0\\ &\vdots\\ \widehat{S}_{A,K}(\theta^l_{A,K})-\tfrac{1}{2}-u^l_{A,K}&=0\end{aligned}\qquad(6)$$

to obtain the solutions $(\theta^l_{A,1},\ldots,\theta^l_{A,K})$. Following the argument in Appendix 2 of [10], one can then use the distribution of $(\theta^l_{A,1}-\tilde{\theta}_{A,1},\ldots,\theta^l_{A,K}-\tilde{\theta}_{A,K})$ to approximate the distribution of $(\tilde{\theta}_{A,1}-\theta_{A,1},\ldots,\tilde{\theta}_{A,K}-\theta_{A,K})$ (which after rescaling is also asymptotically a mean zero multivariate normal). In particular, from the generated sample of the $(\theta^l_{A,1},\ldots,\theta^l_{A,K}), l=1,\ldots,M$, one can estimate the covariance matrix of the estimators $\tilde{\theta}_{A,j}$, $j=1,\ldots,K$.

By repeating the process for treatment group B one can then immediately obtain an estimate of the asymptotic variance–covariance matrix of the (asymptotically normal) vector of the treatment effects $\widehat{\theta}_j=\tilde{\theta}_{A,j}-\tilde{\theta}_{B,j}$, $j=1,\ldots,K$. As we have seen above, an additional $(K+1)$st subpopulation is needed for the definition of the test statistic T that contains all patients in the study.

Testing of the null hypothesis of no treatment–covariate interaction and the construction of a confidence band around the estimated treatment effects then follows as described in Section 2.

Here, too, one is concerned about being too far from the asymptotic distribution for realistic sample sizes. Trouble is likely since this median survival implementation is based on the same asymptotic result used for the survival estimate implementation. Table V illustrates this point by showing a comparison of the α level recovery ability of the asymptotic versus the permutation distribution approach to inference. For the asymptotic approach we estimated the median survival within each subpopulation and across subpopulations for each treatment arm, as well as the covariance matrix of the estimated treatment effects (after subtracting from each the overall treatment effect for T). In particular, the covariance matrix estimation required that within each of the 300 simulations we also generate 1000 samples $(u_{A1},\ldots,u_{AK})$ and $(u_{B1},\ldots,u_{BK})$ from the distributions $N_K(0,\widehat{\Sigma}_A)$ and $N_K(0,\widehat{\Sigma}_B)$ within the two treatment arms and that we solve equations (6). The α level of the T test statistic for this implementation of STEPP was then estimated similarly to what was done for the survival estimate case. As far as the permutations-based method is concerned, its specialization to this quantile implementation is straightforward. Here, too, in some cases ($S(4)=0.1, n=100$ and $S(4)=0.5, n=100$) simulated and permuted data sets that did not allow the estimation of the median survival were discarded and a replacement data set (or permuted sample) generated until the target number was obtained. The results shown in Table V suggest that the permutation distribution approach should probably be preferred also for this implementation of STEPP.

Note that the permutation construction that we have discussed above actually consists of the absence of a covariate effect on survival within each arm, while allowing for different survival levels in the two arms. Similar to what was pointed out in [8, p. 196], one can expect the test to detect alternative hypotheses in which a treatment–covariate interaction does exist because the subpopulation-specific treatment effects will in fact vary. On the other hand, in situations of no interaction but with (equal) covariate effect one would expect adequate recovery of the alpha

DOI: 10.1002/sim

Table V. Estimated α level of the test for interaction based on the T statistic for the median survival implementation of STEPP.

				Asymptotic			Permutation		
				α			α		
	n	r_1	r_2	0.01	0.05	0.10	0.01	0.05	0.10
$S(4)=0.1$	100	30	40	0.00	0.02	0.06	0.01	0.07	0.12
	200	60	80	0.01	0.04	0.06	0.01	0.04	0.09
	500	150	200	0.01	0.04	0.08	0.01	0.06	0.10
	1000	300	400	0.01	0.04	0.06	0.00	0.06	0.11
$S(4)=0.5$	100	30	40	0.01	0.04	0.10	0.03	0.08	0.12
	200	60	80	0.01	0.03	0.07	0.01	0.04	0.10
	500	150	200	0.00	0.01	0.06	0.00	0.04	0.10
	1000	300	400	0.01	0.02	0.04	0.01	0.05	0.09

Results are obtained from 300 simulations of sample size n and they are based on the asymptotic approach and on the permutation approach to inference, respectively, for the left and right side of the table. The standard errors for the three α levels 0.01, 0.05, and 0.10 are equal to 0.006, 0.013, and 0.017, respectively (based on the normal approximation). The distribution of the covariate of interest Z is N(55, 25). Entry times follow a Uniform(0, 5) distribution and follow up ends at 7 years from start of accrual.

level and to check this we have considered the two additional scenarios shown in Figure 2. The scenarios consist of a constant (but different across the two arms) survival probability at $t^*=4$ (Scenario A) and of a decreasing survival probability at t^* as Z increases, but such that the survival difference between the two arms is constant (Scenario B), so that in both cases non-treatment–covariate interaction exists. In particular, also with survival time distributed exponentially, these scenarios are: (A) No effect of Z on survival, but presence of a non-zero treatment effect $S_A(t^*|z)-S_B(t^*|z)=\Delta$, which is however not a function of Z. In particular, $\lambda_A(t|z)=\lambda_A(t)=\lambda_A$ and $\lambda_B(t|z)=\lambda_B(t)=\lambda_B=-(1/t^*)\log[S_A(t^*)-\Delta]=-(1/t^*)\log[\exp\{-t^*\lambda_A\}-\Delta]$; ($B$) Effect of Z on survival and presence of a non-zero treatment effect $S_A(t^*|z)-S_B(t^*|z)=\Delta$, which is however not a function of Z. In particular, $\lambda_A(t|z)=\lambda_A(z)=g_A(z)=\beta_0+\beta_1 z$ and $\lambda_B(t|z)=\lambda_B(z)=g_B(z)=-(1/t^*)\log[\exp\{-t^*(\beta_0+\beta_1 z)\}-\Delta]$. It is easy to check that these expressions do indeed correspond to the situation $S_A(t^*|z)-S_B(t^*|z)=\Delta$. In ($A$) we used $S_A(t^*)=S_A(4)=0.6$, to which correspond the hazard $\lambda_A=0.1277064$. Using $\Delta=0.2$ yields $\lambda_B=0.2290727$. In (B) we follow the structure of arm B in the power scenario 1: in particular, for arm A we choose $\beta_0=-(65/75)(0.2290727)$ and $\beta_1=(2/75)(0.2290727)$. These values are such that the hazard rate in this arm increases linearly as a function of Z, so that survival at $t^*=4$ decreases (non-linearly) as Z increases. To achieve a constant decrease Δ in survival when we move from arm A to arm B we calculate for arm B the conditional hazard function $\lambda_B(t|z)$ given above. In Figure 2 we show the hazard functions and the survival percentages at $t^*=4$ for the two scenarios. The results for these additional scenarios are reported in Table VI, which shows accurate recovery of the alpha level also for these cases by both implementations of STEPP. Note that for the median implementation Scenario B is not shown, as it is designed to satisfy the lack of interaction as defined by a constant difference in survival proportions at $t^*=4$ and not a constant difference in medians.

DOI: 10.1002/sim

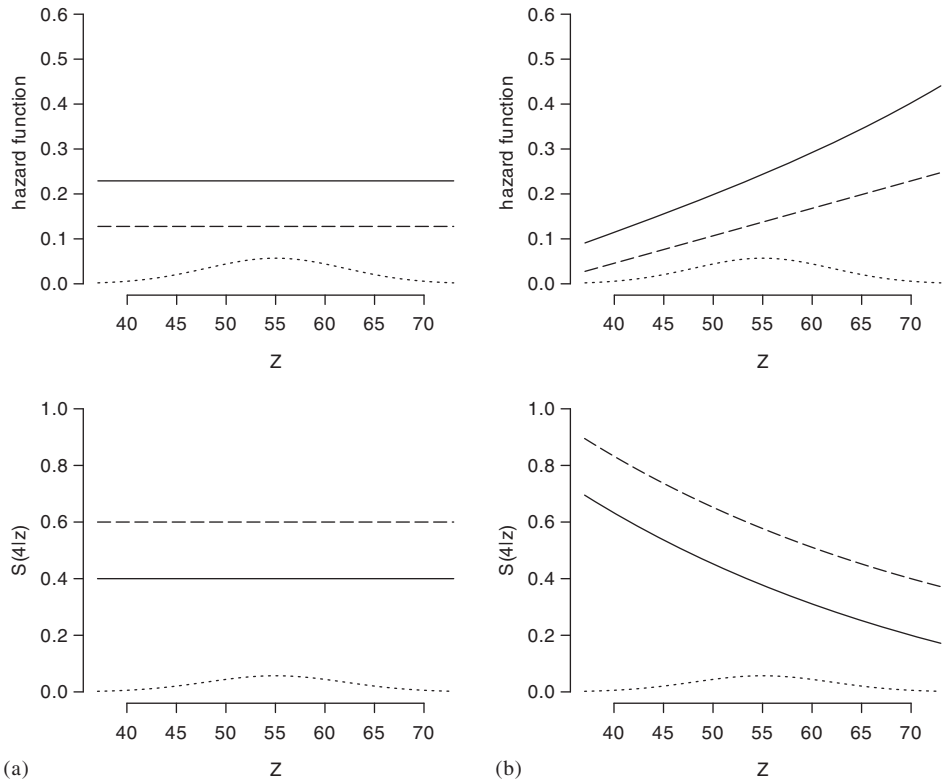

Figure 2. Two additional scenarios for the alpha recovery simulations reported on in Table V. The two scenarios are (A) and (B) as described in Section 3. The top graphs show the hazard as a function of the covariate Z ($\lambda_1(z)$ and $\lambda_2(z)$), and the bottom graphs the proportion surviving at $t^*=4$, also as a function of Z ($S_1(t^*|z)$ and $S_2(t^*|z)$). Dashed lines refer to the first treatment group and solid curves to the second treatment group. The dotted curve is a sketch of the density function of the covariate Z.

5. DISCUSSION

STEPP is an exploratory tool with graphical features that makes it easier for clinicians to interpret the results of the analysis. The method provides an opportunity to detect interactions beyond those that may be apparent based on regression models (such as Cox models). Positive results should be confirmed using results from other data sets investigating similar treatment comparisons. It should also be clear that STEPP is not meant to determine specific cutpoints in the range of values of the covariate of interest, but rather to provide some indication on ranges of values where the treatment effect might have a particular behavior. As pointed out in [1, 2], checking for robustness of the analysis to the choice of the parameters that define the subpopulations is recommended.

In this paper, we explored the performance of STEPP for detecting some cases of heterogeneity in treatment effect with respect to a covariate; we studied the goodness of the asymptotics-based inference and we introduced a new implementation of STEPP that uses the difference in

DOI: 10.1002/sim

Table VI. Estimated α level of the test for interaction based on the T statistic for the STEPP implementation based on the difference in estimated survival at 4 years and on the difference in median survival. Results refer to scenarios A and B as described in the text.

				$S(4)$ implementation			Median survival implementation		
				α			α		
	n	r_1	r_2	0.01	0.05	0.10	0.01	0.05	0.10
Scenario A	100	30	40	0.01	0.05	0.10	0.01	0.06	0.12
	200	60	80	0.01	0.05	0.09	0.01	0.03	0.06
	500	150	200	0.01	0.06	0.11	0.02	0.07	0.12
	1000	300	400	0.01	0.04	0.08	0.01	0.04	0.08
Scenario B	100	30	40	0.01	0.04	0.09	—	—	—
	200	60	80	0.01	0.03	0.08	—	—	—
	500	150	200	0.02	0.05	0.08	—	—	—
	1000	300	400	0.01	0.05	0.09	—	—	—

Note: Results for Scenario B are only shown for the survival proportion implementation, as the scenario is not a no interaction scenario for the median difference definition of treatment effect.

an estimated percentile of the survival function between the two treatment groups. Because of limitations of the asymptotic inference, we proposed an alternative permutation-based inference, which has clearly been shown to be preferable. Note that this approach could also be exploited for future implementations of STEPP.

A comment on the meaning of interaction is useful. The interaction between a covariate value and the magnitude of treatment effect depends on the measure of treatment effect being used. For example, a no interaction model looking at relative risk is an interaction model looking at absolute differences in 5-year survival if the baseline survival risk differs across subpopulations. This fact should be kept in mind when examining the results from the power comparison of the STEPP test with the test on the interaction coefficient in a Cox model described in Section 3.2. Also, as in all smoothing methods, the choice of the parameters that determine the amount of smoothing may have an important impact on the results of the analysis. When using STEPP one should in particular be careful while experimenting with different values of the two parameters r_1 and r_2.

STEPP is designed to investigate patterns of treatment effect across subpopulations defined according to a covariate of interest. Situations might arise in which treatment appears to be particularly effective in a subpopulation, but not in another subpopulation, that is, overlapping with it. In such cases we suggest that investigations using other data sets be performed to further study the behavior of treatment effect in such regions.

As a last remark, note that the STEPP methodology addresses the well-known problem of multiplicity that arises when one conducts several subgroup analyses, as heterogeneity is evaluated in an overall way with an omnibus statistical test. However, this is only done with respect to the covariate being studied, so that if one performs additional analyses (STEPP or other) on other subgroups or other variables, the multiplicity concerns remain and should be considered. Finally, the usual caveats regarding pre-specified versus post-hoc analyses also remain, as do the general considerations on the appropriate reporting of subgroup analyses (see for example [11]).

DOI: 10.1002/sim

ACKNOWLEDGEMENTS

Partial support was provided by the United States National Cancer Institute Grant CA-75362 and by the Italian Ministry for Research (MIUR) protocol 2007AYHZWC *Statistical methods for learning in clinical research*. We thank L. J. Wei for his suggestions on the extension presented in Section 4. We also thank two anonymous referees for their thoughtful comments to an earlier version of the manuscript.

REFERENCES

1. Bonetti M, Gelber RD. A graphical method to assess treatment–covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine* 2000; **19**:2595–2609.
2. Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 2004; **5**(3):465–481.
3. Colleoni M, Li S, Gelber RD, Coates AS, Castiglione-Gertsch M, Price KN, Lindtner J, Rudenstam C-M, Crivellari D, Collins J, Pagani O, Simoncini E, Thürlimann B, Murray E, Forbes J, Eržen D, Holmberg SB, Veronesi A, Goldhirsch A for the International Breast Cancer Study Group. Timing of CMF chemotherapy in combination with tamoxifen in postmenopausal women with breast cancer: role of endocrine responsiveness of the tumor. *Annals of Oncology* 2005; **16**:716–725.
4. Colleoni M, Litman HJ, Castiglione-Gertsch M, Sauerbrei W, Gelber RD, Bonetti M, Coates AS, Schumacher M, Bastert G, Rudenstam C-M, Schmoor C, Lindtner J, Collins J, Thürlimann B, Holmberg S, Crivellari D, Beyerle C, Neumann RLA, Goldhirsch A for the International Breast Cancer Study Group and the German Breast Cancer Study Group. Duration of adjuvant chemotherapy for breast cancer: a joint analysis of two randomised trials investigating 3 versus 6 courses of CMF (cyclophosphomide, methotrexate, and 5-fluorouracil). *British Journal of Cancer* 2002; **86**:1705–1714.
5. Crivellari D, Price K, Gelber RD, Castiglione-Gertsch M, Rudenstam C-M, Lindtner J, Fey MF, Senn HJ, Coates AS, Collins J, Goldhirsch A for the IBCSG. Adjuvant endocrine therapy compared with no systemic therapy for elderly women with early breast cancer: 21-year results of International Breast Cancer Study Group trial iv. *Journal of Clinical Oncology* 2003; **21**:4517–4523.
6. International Breast Cancer Study Group. Endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node negative breast cancer: a randomized trial. *Journal of the National Cancer Institute* 2002; **94**:1054–1065.
7. International Breast Cancer Study Group. Adjuvant chemotherapy followed by goserelin versus either modality alone for premenopausal lymph node negative breast cancer: a randomized trial. *Journal of the National Cancer Institute* 2003; **95**:1833–1846.
8. Potthoff RF, Peterson BL, George SL. Detecting treatment-by-centre interaction in multi-centre clinical trials. *Statistics in Medicine* 2001; **20**:193–213.
9. Pesarin F. *Multivariate Permutation Tests with Applications in Biostatistics*. Wiley: Chichester, 2001.
10. Keaney KM, Wei LJ. Interim analyses based on median survival times. *Biometrika* 1994; **81**:279–286.
11. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* 2007; **357**:2189–2194.

DOI: 10.1002/sim