Biostatistics (2004), **5**, 3, *pp*. 465–481
doi: 10.1093/biostatistics/kxh002

Patterns of treatment effects in subsets of patients in clinical trials

MARCO BONETTI, RICHARD D. GELBER
Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute,
44 Binney Street, Boston, MA 02115, USA
bonetti@jimmy.harvard.edu

SUMMARY

We discuss the practice of examining patterns of treatment effects across overlapping patient subpopulations. In particular, we focus on the case in which patient subgroups are defined to contain patients having increasingly larger (or smaller) values of one particular covariate of interest, with the intent of exploring the possible interaction between treatment effect and that covariate. We formalize these subgroup approaches (STEPP: subpopulation treatment effect pattern plots) and implement them when treatment effect is defined as the difference in survival at a fixed time point between two treatment arms. The joint asymptotic distribution of the treatment effect estimates is derived, and used to construct simultaneous confidence bands around the estimates and to test the null hypothesis of no interaction. These methods are illustrated using data from a clinical trial conducted by the International Breast Cancer Study Group, which demonstrates the critical role of estrogen receptor content of the primary breast cancer for selecting appropriate adjuvant therapy. The considerations are also relevant for general subset analysis, since information from the same patients is typically used in the estimation of treatment effects within two or more subgroups of patients defined with respect to different covariates.

Keywords: Breast cancer; STEPP; Subset analysis; Survival data; Treatment-covariate interaction.

1. INTRODUCTION

Consider a randomized phase III clinical trial in which patients are randomized to one of two treatment arms. The main goal of performing such a study is usually the estimation of some overall measure of treatment effect based on all eligible patients enrolled in the study. Given the resources and effort required to conduct a clinical trial, it is current practice in most reported clinical trial analyses to also estimate treatment effects and test hypotheses within different subsets of patients. The use of such subset analyses, however, is controversial. The statistical concerns include the issue of inflation of alpha levels due to repeated testing, i.e. the possibility that significant results will emerge by chance alone when one examines multiple subgroups of patients. The other major concern is the real possibility of lack of power to detect treatment effects within smaller subgroups of patients. This is likely to produce false negative results, and it is particularly relevant since most clinical trials are designed to have enough power for a primary test of treatment effect only for the whole study population. Results from subset analyses therefore should always be treated and presented with caution.

Consider the ISIS-1 study of atenolol vs. control in patients with suspected acute myocardial infarction. In that study an overall mortality reduction of 30% was observed ($p < 0.004$), but a significant reduction of 71% ($p < 0.01$) was observed within the subgroup of patients born under the astrological sign of Leo. Treatment effect was not significant ($p > 0.1$) within each of the remaining 11 astrological birth signs (see ISIS-1 Collaborative Group, 1986). In the ISIS-2 study that followed, a treatment effect on mortality was observed overall ($p < 0.00001$), but not for the subgroup of patients with astrological birth sign Gemini or Libra ($p = 0.5$) (see ISIS-2 Collaborative Group, 1988). Controversies about the results of subsets analyses are quite common. A recent example is the presentation of the preliminary results of an AIDS vaccine clinical trial, which showed no apparent efficacy of the vaccine in reducing the infection rate of the 5009 high-risk individuals enrolled in the study. The infection rates were 5.7 and 5.8% respectively for the vaccine and the placebo group. However, the infection rates for the 314 blacks alone were 2.0 and 8.1% for the vaccine and the placebo groups. When combining all patients who are either black, asian, or in general non-white and non-hispanic, the mortality reduction was statistically significant ($p = 0.003$). The clinical, political, and regulatory ramifications of such findings are easy to imagine, and the discussion focused precisely on the fact that the positive results arise from a subgroup analysis. For a discussion we refer to Cohen (2003).

One illustration of the dangers of false-positive results associated with subset analyses is given in Peto (1982). Let the test statistic for testing the difference between two treatment groups A and B be $\widehat{Z} = (\overline{X}_A - \overline{X}_B)/(2S^2/n)^{1/2}$, where $\overline{X}_A$ and $\overline{X}_B$ are the sample means in the two treatment groups, that we assume contain n patients each. S^2 here is the estimated variance, assumed equal in the two groups. Peto showed that if the overall observed result is $\widehat{Z} = 2$ (two-sided $p = 0.05$), then there is approximately a 33% chance that the test for no treatment difference will be very significant ($p < 0.001$) for a randomly selected subset of half the patients and not close to significant ($p > 0.32$) for the other half. Thus, the results of subset analyses are subject to more statistical variation than is often appreciated.

In spite of these dangers, however, subgroup analyses can be very useful. They may help better link the results of a study to current clinical practice by allowing estimation of the magnitude of treatment effects within relevant patient subpopulations. Increasing the usefulness of clinical trial results for patient care decision-making is the main reason why subgroup analyses are performed and reported in the clinical literature. It is an overarching priority for clinicians to try to identify the best possible treatment for the *individual* patient, so that benefit can be maximized and the negative side effects from unnecessary treatment reduced to a minimum. Along those lines, one could argue that providing only overall results when the estimated size of the treatment effect may differ according to subgroups can actually be harmful to the patients.

As an example of the tension existing between avoiding and utilizing subset analyses, Coates *et al.* (2002) pointed out that the tamoxifen overview data described by the Early Breast Cancer Trialists' Collaborative Group (1998a) is based on a subset of breast cancer patients with endocrine responsive tumors who had enrolled in trials evaluating tamoxifen. While this subset included only 17% of the 48 000 patients enrolled into randomized trials comparing tamoxifen with no tamoxifen, that is the patient group and treatment duration (5 years) most relevant for determining the efficacy of tamoxifen. On the other hand, similar subset analyses have still not taken place within the chemotherapy overview by the Early Breast Cancer Trialists' Collaborative Group (1998b), even though the magnitude of the chemotherapy effect is substantially affected by estrogen-receptor status of the primary tumor, patient age, and the concurrent use of tamoxifen.

Subgroup analyses can also help raise questions for further research. For example, a retrospective analysis in Aebi *et al.* (2000) suggested that the endocrine effects of chemotherapy alone were insufficient for very young patients with endocrine-responsive breast cancer. A subsequent analysis, described in Goldhirsch *et al.* (2001), was performed on data from clinical trials run by three major US co-operative groups, and that analysis showed the same interaction between the effect of age and steroid hormone receptor status of the primary tumor. When feasible, such independent confirmation of results should always be performed to validate the initial findings.

Once one accepts the idea of performing subgroup analyses, this should therefore be done carefully.

Table 1. *Common inflation factor γ necessary to enlarge K marginal 95% confidence intervals to ensure 95% simultaneous coverage*

K	1	2	3	4	5	10	15	20	25	30
γ	1	1.14	1.22	1.27	1.31	1.43	1.49	1.54	1.57	1.6

Accepted strategies typically include the requirement that they be pre-specified in the study protocol, so that they can be considered 'a priori' analyses as opposed to 'post-hoc' (or data-derived) analyses. This, however, may not always be possible, as new scientific knowledge may be available at the time of the analysis that was not available at the time when the study was designed (with amendments to the protocol being only a partial solution to the problem). By contrast to the assessment of results according to astrological sign, subset analyses conducted to independently confirm or refute biologically-plausible hypotheses generated by others should be encouraged. The pre-specification of the subgroup analyses to be performed is meant to protect from the temptation to over-analyze the data in search for statistical significance, or more generally to avoid identifying subgroups after having looked at the data. There is also clear agreement on the fact that subgroups should only be defined according to pre-treatment characteristics and not, for example, on compliance or outcome variables.

In subgroup analyses treatment effects are usually computed within groups of patients defined with respect to the value of a covariate of interest to explore the possible interaction of the magnitude of treatment effect and that covariate. Gail and Simon (1985) studied the case of pre-specified and non-overlapping subsets of patients defined with respect to one or more categorical covariates. The use of non-overlapping subsets has the disadvantage of only allowing the relatively imprecise estimation of treatment effects within each of the subsets, given the usually limited number of patients enrolled in clinical trials (in the hundreds). Also, small subsets may result in poor approximation of asymptotic distributions of treatment effects to the true distributions.

To overcome this problem, overlapping subgroups of patients can be constructed so that they contain patients with increasingly larger (or smaller) values of the covariate of interest. This practice clearly produces correlated estimates of treatment effect. The need for the application of proper adjustments to the results from each subset analysis to obtain the correct simultaneous inference can be illustrated as follows. Suppose that K confidence intervals $CI_j = [\widehat{\theta}_j - z_{1-\alpha/2}\sigma_j, \widehat{\theta}_j + z_{1-\alpha/2}\sigma_j]$, $j = 1, \ldots, K$ are constructed from the treatment effect estimators $\widehat{\theta}_j \sim N(\mu_j, \sigma_j^2)$, with $z_{1-\alpha/2}$ being the $(1-\alpha/2)$th percentile from the standard normal distribution. For simplicity we assume that the null hypothesis of interest is that $\mu_j = 0$ for all $j = 1, \ldots, K$. One can increase the width of each confidence interval CI_j by multiplication of the half-width by a factor γ so that overall, $P(\cup_{j=1}^{k}\{\{0\} \notin [\widehat{\theta}_j - \gamma z_{1-\alpha/2}\sigma_j, \widehat{\theta}_j + \gamma z_{1-\alpha/2}\sigma_j]\}) = \alpha$. If we assume the estimators $\widehat{\theta}_j$ to be independent (such as when they are computed separately on a set of disjoint subsets of patients), then $\gamma z_{1-\alpha/2} = z_{1-\alpha^*/2}$ with $\alpha^* = 1 - (1-\alpha)^{1/k}$. This corresponds to the values of γ shown in Table 1 for various numbers of subgroups K, using $\alpha = 0.05$. It should be noted that the adjustment under this independence case is similar to that obtained by application of a Bonferroni correction to each confidence interval (i.e. based on $\alpha^* = \alpha/K$): calling A_j the event $\{\{0\} \in CI_j\}$, Boole's inequality $P\left(\cup_{j=1}^{K}\overline{A}_j\right) \leqslant \sum_{j=1}^{K} P(\overline{A}_j)$ is quite tight under independence and when the probabilities, $P(\overline{A}_j)$, are small.

On the other extreme, if one assumes perfect (positive or negative) correlation among the estimators $\widehat{\theta}_j$, one has $P\left(\cup_{j=1}^{K}\overline{A}_j\right) = 1 - P\left(\cap_{j=1}^{K}A_j\right) = 1 - P(A_i)P(A_2|A_1)P(A_3|A_1 \cap A_2)\cdots P(A_K|A_1 \cap \cdots \cap A_{K-1})$. Since all the conditional probabilities are equal to one, this is equal to $1 - P(A_1) = \alpha$, so that no adjustment by γ is necessary (or equivalently, $\gamma = 1$). The general case will fall somewhere between these two extreme situations, with the treatment effect estimates being correlated due to the fact that an overall model is fitted to the data, or because individual estimates are produced on overlapping subgroups of patients (possibly defined with respect to different variables as is commonly done in the clinical literature). For example, if $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with zero mean, and we let $\overline{X}_p = (1/p)\sum_{i=1}^{p} X_i$ and $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$ be the means computed on the first p observations and on all n observations, respectively, then the correlation coefficient between the two means is equal to $(p/n)^{1/2}$. Thus the correct reduction in the individual α levels needs to be determined.

Subset analysis approaches share a similar motivation with the method of the maximally selected chi-square, that consists of testing for outcome differences between groups of patients defined by a changing cutoff value of a covariate, to find the cutoff that produces the maximum difference between the two groups (see Miller& Siegmund (1982) and Betensky and Rabinowitz (1999)). Smoothing techniques have also been applied to the study of treatment–covariate interactions. For example, Kim and Truong (1998) used linear smoothers, and Gray (1992) used splines in additive models for the analysis of survival data under the assumption of proportional hazards. A procedure of double smoothing to construct an estimator of a percentile of the survival distribution as a function of one or two covariates is described in Bowman and Wright (2000).

Methods based on the construction of subsets of patients are used extensively because they present three appealing advantages over more traditional smoothing techniques: (i) they are easily described to and understood by clinical audiences; (ii) they estimate treatment effect within each subset by traditional statistics (e.g. by Kaplan–Meier estimates of survival); and (iii) they allow for the clear definition of the group of patients used to estimate the treatment effect for each subset. The drawback of these analyses can be the use of inappropriate inferential procedures, if no adjustment is made for the fact that patient subgroups may be overlapping. In this report we formalize these approaches, focusing on the case of subsets of patients defined with respect to one covariate.

We identify below two specific families of subpopulations that can be used, and we call the plot of the resulting treatment effects a STEPP, or subpopulation treatment effect pattern plot. A first implementation of STEPP was introduced in Bonetti and Gelber (2000) in the context of studying deviations from the hypothesis of no treatment–covariate interaction in the Cox proportional hazards model. In Section 2 we introduce our approach, and describe a general result for the joint asymptotic distribution of treatment effects estimated over possibly overlapping subsets of patients. The results are specialized to inference for the difference in survival at a fixed time point between two arms (described in Section 3).

From October 1988 to August 1999, 1715 postmenopausal patients with lymph node-negative, operable breast cancer were entered in a clinical trial (Trial IX) run by the International Breast Cancer Study Group (IBCSG). Patients were randomly assigned to receive either tamoxifen for 5 years or 3 cycles of chemotherapy (CMF) at months 1, 2 and 3 after randomization followed by tamoxifen for 57 months. Disease-free survival (DFS) was defined as the length of the time from randomization to any relapse, the appearance of a second primary cancer, or death, whichever occurred first. In Section 4 we apply our methods to the examination of a possible interaction between the estrogen receptor (ER) level of the patient's tumor and the treatment effect measured in terms of the difference of DFS at 5 years between the two arms of the Trial. Clearly, only plausible interactions should in general be investigated. The motivation for the study of such an interaction in this clinical trial arises from the fact that substantial data exist suggesting that tamoxifen is more effective in patients with estrogen receptor-positive tumors (see Early Breast Cancer Trialists' Collaborative Group, 1998a and Ludwig Breast Cancer Study Group, 1984). In Trial IX the randomization was stratified according to ER status (negative and positive), and the intention to perform separate analyses according to ER status was specified in the original protocol. Our analysis below will be an extension of these analyses to consideration of the numerical value of ER level (as opposed to only the status being positive or negative).

Finally, we close with some discussion in Section 5.

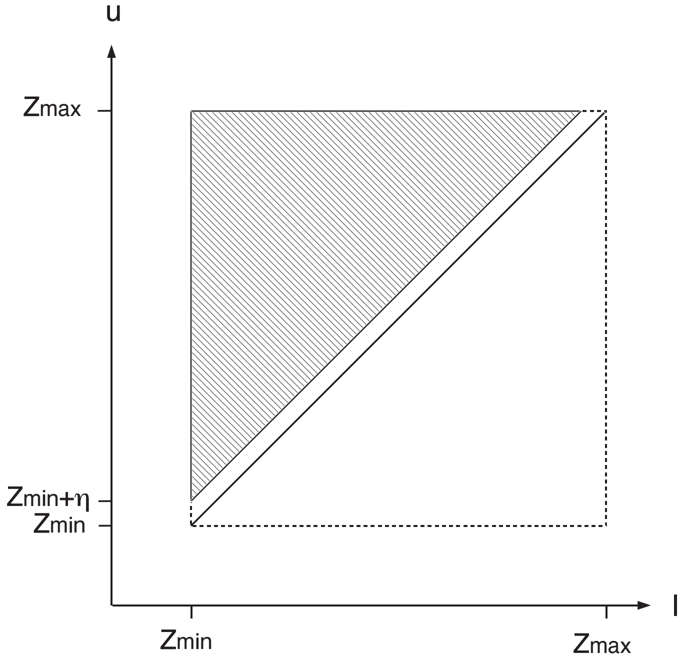

Fig. 1. The shaded region in the figure is the set $\mathcal{W} = \{l \in [Z_{\min}, Z_{\max} - \eta], u \in [l + \eta, Z_{\max}]\}$.

2. Definition of subpopulations and STEPP

Consider the population of n patients in a clinical trial, where a covariate $Z \in [Z_{\min}, Z_{\max}] \subset \Re$ is collected on all individuals together with each patient's treatment assignment, and with some measure X of outcome. Here we will focus on the outcome measure defined as the Kaplan–Meier estimate $\widehat{S}(t^*)$ of survival at the fixed time point t^* within each subgroup of patients. For the IBCSG Trial IX, Z is the ER expression level of the tumor, and outcome is disease-free survival at 5 years. The treatment effect of interest within each subgroup of patients is the difference $\widehat{S}_A(t^*) - \widehat{S}_B(t^*)$ between the two independent treatment groups A and B, so that without loss of generality we discuss one arm. In this setting one has $X = (T \wedge U, \delta = 1(X = T))$, where T and U are independent failure time and censoring time random variables, respectively. Assume for now that the marginal distribution F_Z of Z is absolutely continuous over its support $[Z_{\min}, Z_{\max}]$.

We thus observe the i.i.d. sample $\{Y_i = (X_i, Z_i),\ i = 1, \ldots, n\}$. Subsets of observations defined with respect to the covariate Z can be constructed, with the subset containing observations having $Z \in [l, u]$ being identified by the pair of values (l, u), with $Z_{\min} \leqslant l < u \leqslant Z_{\max}$.

The subset approaches described in Section 1 consist of selecting a finite collection of points $\{(l_j, u_j), j = 1, \ldots, K\}$ from the shaded set $\mathcal{W}$ in Figure 1 (i.e. a finite collection of subsets of patients) and of estimating the treatment effect within each subset. Note that for technical reasons we do not include in $\mathcal{W}$ pairs (l, u) with $u < l + \eta$, where η is some positive constant (see Appendix, Section A1), and that collections of non-overlapping subsets of patients can also be extracted from $\mathcal{W}$. In Section A1 we show (Theorem 1) that the collection of treatment effects $\widehat{S}_A(t^*|l < Z < u) - \widehat{S}_B(t^*|l < Z < u)$ over all subpopulations in the shaded region converges to a Gaussian process indexed by the family of

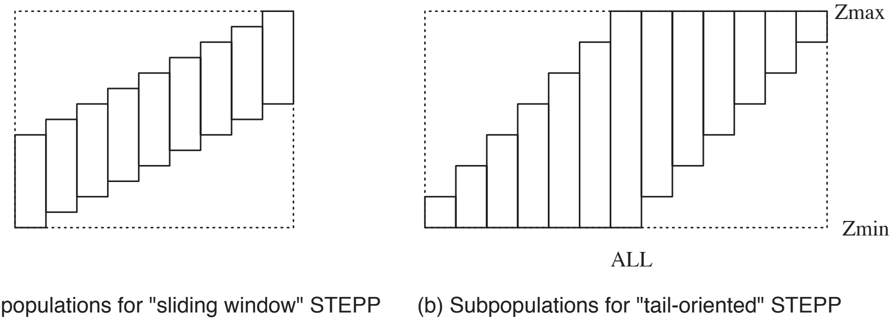
Zmax
Zmin
ALL
(a) Subpopulations for "sliding window" STEPP
(b) Subpopulations for "tail-oriented" STEPP

Fig. 2. Illustration of the two subpopulation patterns.

subpopulations. Note in the proof of Theorem 1 how Z needs not be one-dimensional. In fact, the result is valid (with minor modifications such as the actual value of the VC-index) if one considers rectangular intervals $[l, u]$ in higher dimensions. From that result it follows in particular that the joint asymptotic distribution of treatment effects defined over a finite colleciton of subpopulations is multivariate normal.

Among the possible choices of patterns of subpopulations we consider two specific choices, defined for $j = 1, \ldots, K$ by non-decreasing sets $\{l_j\}$ and $\{u_j\}$ such that

(1) $\{l_j \in [\mathrm{Z}_{\min}, \mathrm{Z}_{\max}], u_j = \mathrm{Z}_{\max}\}$ or $\{l_j = \mathrm{Z}_{\min}, u_j \in [\mathrm{Z}_{\min}, \mathrm{Z}_{\max}]\}$ ('tail-oriented' pattern).
(2) $\{l_j \in [\mathrm{Z}_{\min}, \mathrm{Z}_{\max}], u_j = \inf\{u \geqslant l_j \mid P(l_j < Z \leqslant u) \geqslant p\}\}$ for some $p \in (0, 1)$ ('sliding-window' pattern).

Within each of the two regions (1) and (2) a total of K subpopulations are therefore defined through the corresponding collection of cutoff values (l_j, u_j), so that patient with index i having covariate $Z = z_i$ belongs to subpopulation $\mathcal{P}_j$ if $l_j < z_i \leqslant u_j$. (And each subpopulation is therefore in a one–one correspondence with the set of indices of such patients). Inference is conditional on the cutoffs, which may be chosen according to the empirical distribution of Z.

The tail-oriented pattern thus consists of setting $l_j = \mathrm{Z}_{\min}$ for all j and defining the set of cutoff points $u_1 < u_2 < \cdots < u_K = \mathrm{Z}_{\max}$. Subpopulation $\mathcal{P}_j$ then contains patients with $Z_i \leqslant u_j$. In this fashion the subpopulations are nested as an increasing sequence of sets, with the last subpopulation containing all patients. Similarly, one could also start with all patients and define $u_j = \mathrm{Z}_{\max}$ for all j, and then define the cutoff points $l_1 = \mathrm{Z}_{\min} < l_2 < \cdots < l_K$, to consider subpopulations containing patients with $Z_i \geqslant l_j$. The choice of the values l_j and u_j is usually based on different criteria. For example, (a) a roughly constant number of patients may be added (or subtracted) from one subpopulation to the next; (b) 'usual practice' cutoff values may be used that are commonly used in clinical reports for the specific disease; or (c) cutoffs can be set equal to each of the observed values of the covariate if it is discrete.

The sliding window approach can be implemented by fixing two parameters r_1 and $r_2, r_1 < r_2 < n$, and then defining the first subpopulation $\mathcal{P}_1$ as containing patients having Z ranging between the smallest observed value of Z ($l_1 = \mathrm{Z}_{\min}$) and the $(r_2/n \times 100)$th percentile of the observed distribution of Z, taken as u_1. Should such exact percentile not exist, u_1 is defined so that at least r_2 patients fall in $\mathcal{P}_1$. To define the next subpopulation $\mathcal{P}_2$ the value l_2 is found so that at most r_1 patients fall between l_2 and u_1. Then, u_2 is defined so that $\mathcal{P}_2$ contains at least r_2 patients. This process is repeated until the last possible population is defined. Note that after setting $r_1 = np_1$ and $r_2 = np_2$ for some $0 < p_1 < p_2 < 1$, as $n \to \infty$ the proportion of patients in each subpopulation (except for the last one) converges to p_2 (assuming Z continuous). Also, the resulting number of subpopulations converges to the smallest integer greater or equal to the number $[1 + (1 - p_2)/(p_2 - p_1)]$.

Figure 2 summarizes the two constructions (sliding window and tail-oriented) described above. The y-axis represents the values of the covariate Z, and each rectangle describes the set of values of Z defining each subopulation. Note how Figure 2(b) shows the tail-oriented pattern both for increasing and for decreasing values of Z. The label 'ALL' indicates the subpopulation containing all patients. In what follows we will show an example of the sliding window pattern, but all of the discussion also applies to the tail-oriented approach (as well as to other choices of regions within $\mathcal{W}$ that the investigator may choose).

The plot of the estimates $\widehat{\theta}_j$ with the confidence band around them is a STEPP plot, in which the treatment effect estimate for each subpopulation $\mathcal{P}_j$ is plotted versus the x-axis coordinate equal to the median Z covariate value within that subpopulation. Figure 3 shows an example of a sliding window STEPP plot for the IBCSG Trial IX for the DFS difference between the two arms, constructed with respect to the ER expression of the tumor.

3. Inference

From Theorem 1 the estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ of the treatment effects $\theta_1, \ldots, \theta_K$ corresponding to the K subpopulations $\mathcal{P}_1, \ldots, \mathcal{P}_K$ are such that

$$\sqrt{n}\begin{bmatrix} \widehat{\theta}_1 - \theta_1 \\ \vdots \\ \widehat{\theta}_K - \theta_K \end{bmatrix} \xrightarrow{\text{d}} N_K\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \Sigma\right]$$

for some covariance matrix Σ that can be estimated consistently. One can then construct simultaneous confidence bands around the collection of estimators. A 95% rectangular confidence band can be defined as $\{\theta_j \in \widehat{\theta}_j \pm \gamma\ 1.96\,\widehat{\sigma}_j,\ j = 1, \ldots, K\}$, where $\widehat{\sigma}_j = [\widehat{\text{var}(\widehat{\theta}_j)}]^{1/2}$ is a consistent estimator of $[\text{var}(\widehat{\theta}_j)]^{1/2}$, and where the value of γ can be obtained by solving numerically

$$\text{P}\left[\bigcap_{j=1}^{K}\{\theta_j \in \widehat{\theta}_j \pm \gamma\ 1.96\,\widehat{\sigma}_j\}\right] = 0.95$$

for a sample of random variables generated from the asymptotic distribution of the estimators. The parameter γ thus measures the effect of performing joint inference as opposed to marginal inference. An omnibus test for the equality of the treatment effects can be constructed as a quadratic form based on the asymptotic normality of the estimators. However, such a test may be quite sensitive to the particular choice of the subpopulations (see for example Bonetti and Gelber, 2000), and we thus do not recommend its use. An alternative test statistic for the null hypothesis $H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ can be defined as

$$T = \sup\left\{\frac{|\widehat{\theta}_j - \widehat{\theta}_{ALL}|}{\widehat{\sigma}_j^*}, j = 1, \ldots, K\right\},$$

where $\widehat{\theta}_{ALL}$ is the overall treatment effect estimate computed on the subpopulation $\mathcal{P}_{ALL}$ containing all patients, and $\widehat{\sigma}_j^*$ is a consistent estimator of $\sigma_j^* = [\text{var}(\widehat{\theta}_j - \widehat{\theta}_{ALL})]^{1/2}$. Observe that $\mathcal{P}_{ALL}$ corresponds to the choice $(l, u) = (Z_{\min}, Z_{\max})$ in Figure 1. The asymptotic null distribution of T can be estimated by generating samples 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 $H_0 : \theta_1 = \cdots = \theta_K = \theta_{ALL} = \theta_0$, and a p-value for H_0 can then be easily produced.

We now focus on the case of treatment effect defined as the difference in survival at a fixed time point t^*. Let $\{(T_i, U_i), i = 1, \ldots, n\}$ be an i.i.d. sample from independent failure time and censoring time random variables, and the observed data consist of $(X_i = T_i \wedge U_i, \delta_i = 1(X_i = T_i))$. Let $F(t)$ be the cdf of T, and define the survival function $S(t) = P(T_i > t) = 1 - F(t)$. Let $\widehat{S}(t)$ be the Kaplan–Meier estimator of the survival function (Kaplan and Meier, 1958).

Consider the two conditional survival distributions $S_j(t) = S(t|\mathcal{P}_j)$, $j = 1, 2$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are two subpopulations of patients. Let $n_j = \sum_{i=1}^{n} 1(i \in \mathcal{P}_j)$. In Section A2 we show that as $n \overset{n\to\infty}{\longrightarrow} \infty$, if $n_j/n \overset{n\to\infty}{\longrightarrow} p_j$ for $j = 1, 2$,

$$\begin{bmatrix} \sqrt{n_1}(\widehat{S}_1(t) - S_1(t)) \\ \sqrt{n_2}(\widehat{S}_2(t) - S_2(t)) \end{bmatrix} \overset{\mathrm{d}}{\to} N_2(0, A\Sigma A), \text{ with } A = \begin{bmatrix} S_1(t)/\sqrt{p_1} & 0 \\ 0 & S_2(t)/\sqrt{p_2} \end{bmatrix}$$

and $\Sigma = E((\eta_1\ \eta_2)'(\eta_1\ \eta_2))$. The matrix A can be estimated consistently by replacing $\widehat{S}_j(t)$ and n_j/n in its expression. The matrix Σ can also be estimated consistently by the quantity

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} (\widehat{\eta}_{1i})^2 & (\widehat{\eta}_{1i}\widehat{\eta}_{2i}) \\ (\widehat{\eta}_{1i}\widehat{\eta}_{2i}) & (\widehat{\eta}_{2i})^2 \end{bmatrix}$$

where

$$\widehat{\eta}_{ji} = 1(i \in \mathcal{P}_j)\left(\frac{n_j \delta_i 1(X_i \leqslant t)}{\sum_{h=1}^{n} 1(X_h \geqslant X_i)1(h \in \mathcal{P}_j)} - \sum_{w=1}^{n}\left(\frac{1(w \in \mathcal{P}_j)1(X_i \geqslant X_w)\delta_w 1(X_w \leqslant t)}{\left[\sum_{h=1}^{n} 1(X_h \geqslant X_w)1(h \in \mathcal{P}_j)\right]^2}\right)\right).$$

The generalization to k subpopulations $\mathcal{P}_j,\ j = 1, \ldots, K$ is straightforward. Finally, let $\widehat{\Sigma}_A$ and $\widehat{\Sigma}_B$ by the estimates of the asymptotic variance–covariance matrices of the scaled survival estimates $\sqrt{n}\,\widehat{S}_{A,j}(t^*)$ and $\sqrt{n}\,\widehat{S}_{B,j}(t^*)$ (for $j = 1, \ldots, K$) in two treatment groups A and B. Then the scaled vector of the estimated treatment effects $\{\sqrt{n}\,\widehat{\theta}_j = \sqrt{n}\,[\widehat{S}_{A,j}(t^*) - \widehat{S}_{B,j}(t^*) - (S_{A,j}(t^*) - S_{B,j}(t^*))],\ \ j = 1, \ldots, K\}$ is asymptotically normal with the variance–covariance matrix that can be estimated consistently by $\widehat{\Sigma} = \widehat{\Sigma}_A + \widehat{\Sigma}_B$.

In Section A2 we describe a resampling method that can also be used to estimate the same variance–covariance matrix with some savings in terms of computing time.

4. APPLICATION

A controversy exists concerning the proper role of chemotherapy for postmenopausal women with breast cancer. Some argue that chemotherapy provides an insufficient benefit to justify its unpleasant side effects and cost and because overall results of the Early Breast Cancer Trialists' Collaborative Group (1998b) estimate only a 2–3% gain in survival associated with chemotherapy use. Others look to positive results from clinical trials that include women with high risk disease (see Colleoni *et al.*, 1998) and women over 49 years old (regardless of menopausal status, see Fisher *et al.*, 1997) to justify the use of chemotherapy for all postmenopausal women. The controversy is to some extent related to reliance on overall results, without proper consideration of biologically plausible patient subpopulations for which the magnitude of chemotherapy effects may differ.

To illustrate how the STEPP methodology might help to resolve this controversy, we apply the sliding window pattern of subpopulation to data from the IBCSG randomized clinical trial IX that we introduced in Section 1. For a complete description of the trial and its results refer to International Breast Cancer Study Group (2002).

For 1176 of the 1715 patients randomized the quantitative ER measurement on the primary tumor (obtained using the ligand binding method with results expressed in units of femtomols [fmol] per milligram of cytosol protein) was available. The median follow-up for this subset of patients was 72 months, and the DFS proportions at 5 years were 79% and 83% for the tamoxifen alone arm and for the CMF followed by tamoxifen arm, respectively. The p-value for the two-sided log-rank test comparing the two disease-free survival functions was 0.045, and the test for interaction between treatment and the two prospectively defined levels of ER ($<$ 10 fmol/mg or ER-negative; $\geqslant$ 10 fmol/mg or ER-positive) from fitting a Cox proportional hazards model was equal to 0.01 (see International Breast Cancer Study Group, 2002). We use a STEPP plot to confirm the presence and visualize the shape of such interaction.

Figure 3(a) shows the DFS estimates at 5 years for CMF followed by tamoxifen and for tamoxifen alone within subpopulations defined by ER value (plotted on the log scale on the x-axis). Figure 3(b) shows the corresponding STEPP plot of the difference in 5-year DFS between the two treatment groups, with marginal 95% confidence intervals and a 95% confidence band. The subpopulations were constructed using the parameter values $r_1 = 190$ and $r_2 = 200$. As described in Section 2, this produces subpopulations with approximately 200 patients (r_2) in each and slides by dropping and adding about 10 patients ($r_2 - r_1$) for each new subpopulation.

The difference in DFS comparing CMF followed by tamoxifen vs. tamoxifen alone is much larger for smaller values of ER than for larger values of ER. The p-value for the T test statisic is equal to 0.01. In particular, Figure 3(b) illustrates that the DFS at 5 years for the two treatment groups is similar for all subpopulations having median values of ER larger than approximately 15, suggesting that CMF may not be adding much to the effectiveness of tamoxifen for such patients. The STEPP analysis thus confirms the existence of the interaction between treatment and ER status, and it illustrates the magnitude of the impact of this interaction in a way that is clinically useful.

The resampling method described at the end of the Appendix was also applied to these data, and it produced very similar results to the ones obtained by using the asymptotic distribution directly, with the advantage of a slightly faster execution time for the resampling approach.

The study of treatment–covariate interactions via subset analyses requires that a subjective choice be made: namely, the choice of the subset of the region $\mathcal{W}$ to be used. It thus follows that all inferences should be interpreted with such choice in mind. In particular, the test T for no interaction is dependent on the selection of points (l_j, u_j) as defined in Section 2. For the sliding window STEPP the test depends on the values chosen for the two parameters r_1 and r_2, and it seems desirable to perform a simple sensitivity analysis to assess the stability of the conclusions of the analysis. As an example, Table 2 shows the number of subpopulations and the p-value obtained from various choices of these two parameters for the difference in DFS estimate at 5 years in the Trial IX example. The p-values are consistently highly significant whenever at least 15% of the patients are included in each subpopulation (i.e. when $r_2 \geqslant 180$). For smaller numbers of patients in each subpopulation the results become less stable, but they still indicate at least marginal significance (with only two cases resulting in p-values greater than 0.1).

We compared our results with those obtained by fitting the survival models described in Gray (1992), which use a penalized smoothing spline to model the hazard as a function of continuous covariates in the Cox proportional hazards model (we used software kindly provided by Dr Gray). Figure 4 shows the fitted splines as a function of ER within the two treatment arms with bands, showing pointwise ±2 standard deviations. The curves represent the estimated log-hazard ratio of log(ER+1) with respect to the mean covariate value for the two arms. Examination of the two curves suggests a higher hazard ratio for lower values for ER within the tamoxifen alone treatment arm, while the hazard ratio appears to be quite constant as a function of ER in the chemotherapy followed by tamoxifen arm. The statistical test for the corresponding treatment–covariate interaction (i.e. for the equality of the parameters of the two splines) returned a p-value of 0.23, and it thus did not reject the null hypothesis. The two splines were based on 10 knots, with the p-value being quite stable as the number of knots was changed from 5 to 20 (knots:p-value

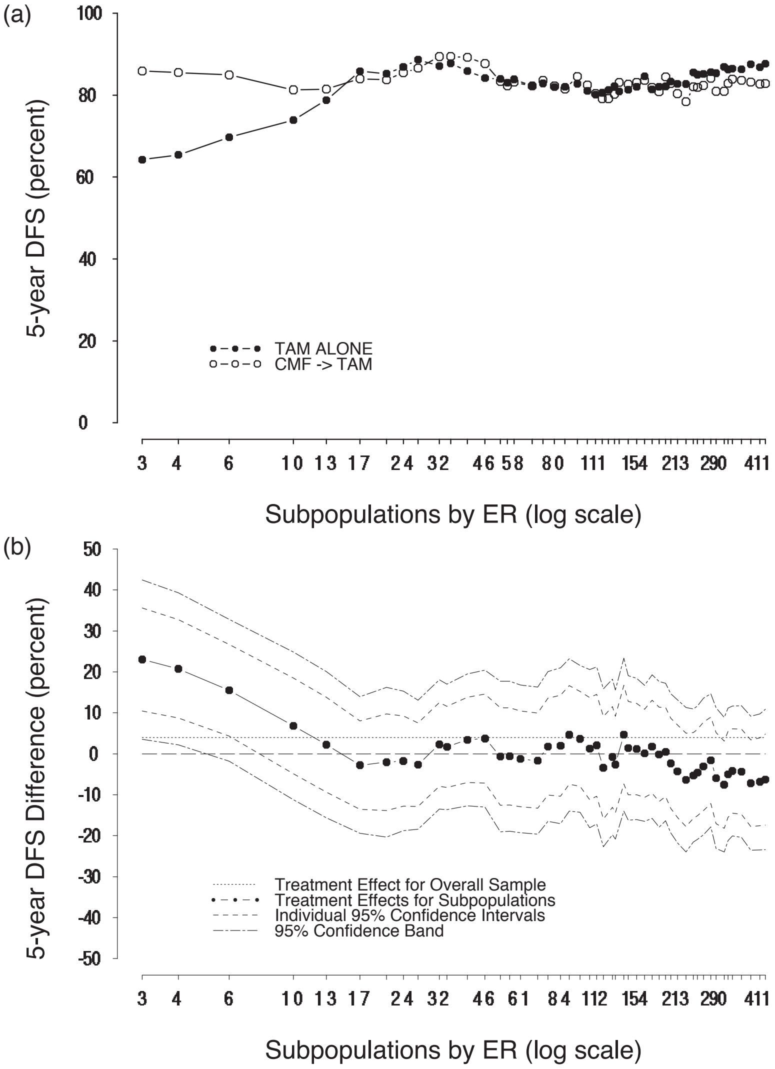

Fig. 3. STEPP plot for IBCSG Trial IX data—ER subgroups: (a) 5-year disease-free survival (DFS) percentages for CMF followed by tamoxifen and for tamoxifen alone; (b) 5-year DFS difference (CMF followed by tamoxifen minus tamoxifen alone).

$= 5 : 0.24, 10 : 0.23, 15 : 0.25, 20 : 0.22$). Despite lack of statistical significance, the results from this spline analysis suggest an increased treatment effect for lower values of ER, which agrees with the more clinically accessible STEPP analysis.

The results from subset analyses could be affected by bias caused by differences in the distribution of important covariates between the two arms across subpopulations. In randomized clinical trials such as the one discussed here one typically relies on randomization (possibly stratified) to achieve overall balancing in covariates across treatment arms. Nevertheless, an assessment of actual balancing is warranted. For the IBCSG Trial IX we examined summaries of the distribution of important covariates (tumor size, age, and type of surgery) across the subpopulations of patients constructed for the STEPP analysis, separately for

Table 2. *Sensitivity analysis of the test T of no treatment–covariate interaction for various choices of r_1 and r_2 $(n = 1176)$ in the IBCSG application. For each (r_1, r_2) combination we report the number of subpopulations (first number in parentheses) and the p-value for the test T (second number in parentheses)*

	r_2					
r_1	120	140	160	180	200	220
100	(34,0.0582)	(21,0.0418)	(15,0.1354)	(12,0.0074)	(10,0.0054)	(9,0.0048)
110	(53,0.0664)	(25,0.0508)	(17,0.1420)	(13,0.0076)	(10,0.0044)	(9,0.0046)
120		(33,0.0530)	(20,0.0370)	(14,0.0060)	(12,0.0050)	(10,0.0032)
130		(51,0.0648)	(25,0.0368)	(16,0.0094)	(13,0.0056)	(10,0.0046)
140			(33,0.0620)	(19,0.0068)	(14,0.0054)	(11,0.0058)
150			(51,0.0708)	(24,0.0088)	(16,0.0070)	(13,0.0072)
160				(31,0.0094)	(19,0.0070)	(14,0.0058)
170				(48,0.0112)	(24,0.0086)	(16,0.0066)
180					(31,0.0106)	(19,0.0098)
190					(48,0.0100)	(23,0.0110)
200						(31,0.0142)
210						(46,0.0126)

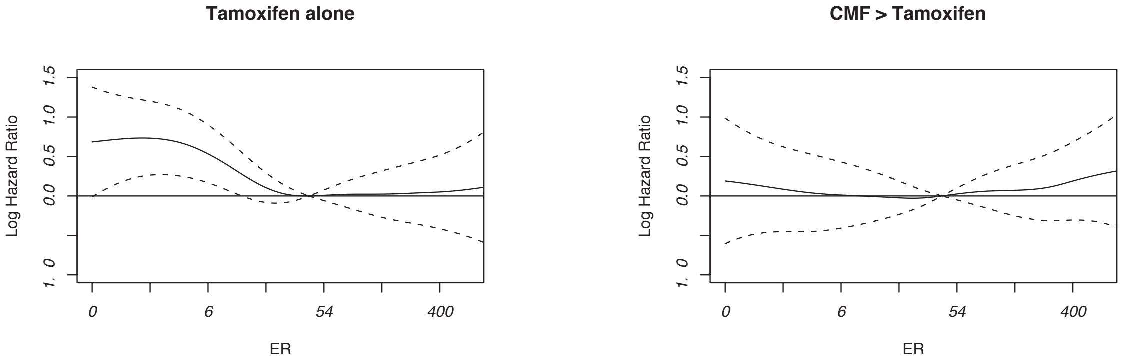

Fig. 4. Results from fitting a spline-based Cox proportional hazards model to the IBCSG Trial IX data.

the two treatment arms (results not shown). We found no indication of differences in these distributions between the two arms. Differential compliance to treatment between the two arms might also explain trends in the treatment effect estimates computed on patient subpopulations, and we examined for that possibility. We defined a patient to be noncompliant to the tamoxifen treatment if she received fewer than 51 months of tamoxifen. The overall tamoxifen noncompliance rate was 7%. No significant trend was observed across the STEPP subpopulations. For completeness, chemotherapy noncompliance in the CMF arm (defined as a patient's receiving fewer than the three cycles) was quite constant at about 9% across the STEPP subpopulations.

5. Discussion

We have provided a framework for the common practice of performing treatment effect estimation on overlapping subsets of patients enrolled in clinical trials. That intuitive practice can represent a useful tool to study treatment–covariates interactions, as the STEPP plots highlight patterns of treatment effect differences without deviating from the conceptual simplicity of subset analysis. Among the advantages deriving from this similarity, the nonparametric nature of the method allows flexibility without imposing strong functional assumptions. In comparison to other approaches (such as the spline method described in Gray, 1992), our subgroup approach is also characterized by the fact that treatment effects are estimated on clearly defined patients subgroups, and use clinically familiar quantities (e.g. the Kaplan–Meier estimator of survival).

One can also plot the observed pattern of treatment effects over a background of reference curves to help identify particular characteristics of the observed path. Such paths can be simulated from the estimated finite-dimensional asymptotic distribution, and plotted under the assumption that all treatment effects are identical to the overall treatment effect. Figure 5 shows an example of such simulated paths for the difference in DFS at 5 years for the IBCSG Trial IX data. Each of the top six graphs shows the observed path of treatment effect (the thicker line) and a reference curve generated from the estimated multivariate normal distribution. The graph at the bottom of the figure shows the observed curve (thick) and 20 such reference curves. The techniques in Bowman and Wright (2000) can also be used to produce reference bands, but only under the hypothesis of independence between survival time and the covariate of interest within each treatment group, and under the hypothesis of a proportional hazards structure. Neither of these assumptions are necessary to the subset approach that we discussed here.

It should be stressed that the common practice of reporting treatment effect estimates within subgroups of patients defined with respect to *different* covariates is a particular case of the methods discussed here. Treatment effect estimates computed on ER-positive patients and on patients older than 40 years of age (say) are clearly correlated, since they are based on overlapping groups of patients. The joint estimation and interpretation of such subgroup treatment effects and of subgroup hypothesis tests requires distributional results such as those presented here, as a Bonferroni adjustment is likely to be too conservative. The results presented here apply immediately, if one defines the subpopulations accordingly.

The approach that we have taken here can be applied to other measures of treatment effect. For example, STEPP can be used to study the difference in median survival between treatment arms (details of that implementation will be presented elsewhere). Also, one can use model-based measures of treatment effect as was done in Bonetti and Gelber (2000), where the Cox proportional hazards model was used to estimate treatment effects. In that context, however, for definiteness one must assume that the model is the same in all subpopulations, and that no treatment–covariate interaction is present. As discussed in Bonetti and Gelber (2000), the analysis in that case is therefore aimed at testing such null hypothesis and at suggesting the form of possible deviations from it, rather than at estimating the magnitude of treatment effects.

As a last comment, note how above we focused on data arising from clinical trials. However, subgroup analyses are also common in observational studies, and the methods are also relevant for that setting, with the addition of the usual caveat of the need for dealing carefully with possible confounding. As we have seen in Section 4, issues of selection bias are relevant more in general (also for randomized studies), since the identification of the causal effects of interest is complicated if randomization is not available (or effective) to reduce confounding and compliance rates differ in the two arms across the subpopulations.

ACKNOWLEDGEMENTS

This work was partially supported by grants from the National Cancer Institute (CA-06516 and CA-75362), the American-Italian Cancer Foundation (AICF 101-98 and 101-99), and by the International Breast Cancer Study Group (IBCSG—funding provided by The Swiss Cancer League, The Cancer League of Ticino, The Swedish Cancer League, The Australian New Zealand Breast Cancer Trials Group

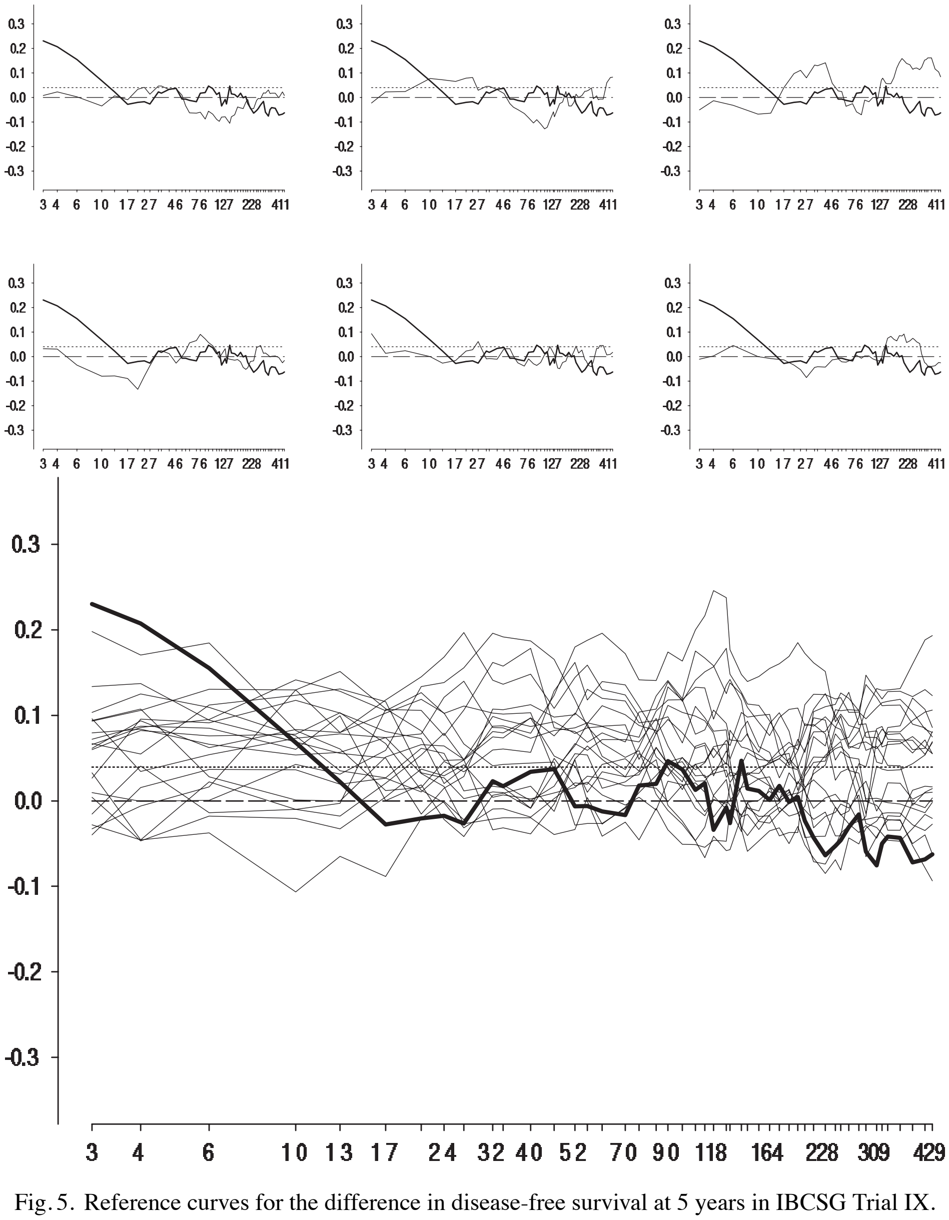

Fig. 5. Reference curves for the difference in disease-free survival at 5 years in IBCSG Trial IX.

[NHMRC grants # 890028, 920876, 950328, 980379, and 141711], The Cancer Council Australia, The Frontier Science and Technology Research Foundation, The Swiss Group for Clinical Cancer Research). We thank the IBCSG (especially Aron Goldhirsch and Monica Castiglione-Gertsch) for permission and encouragement to use the data from IBCSG Trial IX. We thank L. J. Wei for suggesting this implementation of the STEPP method, and David Schoenfeld and Rebecca Betensky for suggesting a version of the test statistic T described in Section 3. Lastly, we are grateful to the reviewers for their thoughtful comments, that have led to improvements to the manuscript.

APPENDIX A

A.1 *Asymptotic distribution of the collection of estimators of treatment effect*

Let $P_n = (1/n)\sum_{i=1}^{n} 1_{\delta_{Y_i}}$ be the empirical measure corresponding to the sample $Y_i = (X_i, Z_i)$, $i = 1, \ldots, n$ (with 1_{δ_y} indicating the probability distribution that is degenerate at y). Also, let $F_n(y) = (1/n)\sum_{i=1}^{n} 1(Y_i \leqslant y) = \sum_{i=1}^{n} 1(Z_i \leqslant z, X_i \leqslant x)$ be the corresponding empirical cdf, and $T(F_n)$ some functional of F_n. Let $F(y)$ be the population cdf of Y.

THEOREM 1 Let $T(\cdot)$ be a Hadamard-differentiable functional. Define the conditional empirical distribution function

$$F(x; l, u) = \frac{\int 1(X \leqslant x)1(l < Z \leqslant u)\,\mathrm{d}F(Y)}{\int 1(l < Z \leqslant u)\,\mathrm{d}F(Y)},$$

where $(x, l, u) \in \mathcal{W} = \Re \times \{(l, u) \in \Re^2 | Z_{\min} < l \leqslant Z_{\max}, u - l \geqslant \eta\}$ for some $\eta > 0$. Then $\sqrt{n}[T(F_n(x; l, u)) - T(F(x; l, u))]$ converges weakly to a Gaussian process indexed by $(x, l, u) \in \mathcal{W}$. We need to consider empirical processes indexed by sets (see for example van der Vaart and Wellner, 1996). It is easy to show that the two collections of sets $\mathcal{C}_1 = \{(l, u] \subset \Re\}$ and $\mathcal{C}_2 = \{(-\infty, x] \subset \Re\}$ both have finite VC-index. Each element C of their product $\mathcal{C} = \mathcal{C}_1 \times \mathcal{C}_2$ can be identified by the triplet (x, l, u), and since it is the product of two VC-classes, it is also a Donsker class. Thus the empirical process $G_n = \sqrt{n}(P_n - P)$ converges weakly to a tight limit G in $l^{\infty}(\mathcal{C})$, where $\mathcal{C}$ can be identified with the subset of $\Re^3$ $\{(l, u, x) | Z_{\min} \leqslant l < u \leqslant Z_{\max}, x \in \Re\}$ and where $P_n(C) = P_n(x, l, u) = (1/n)\sum_{i=1}^{n} 1(X_i \leqslant x)1(l < Z_i \leqslant u)$. Any collection of sets $(C_1, \ldots, C_k)$ of $\mathcal{C}$ can therefore be identified by the triplets $((x, l, u)_1, \ldots, (x, l, u)_k)$. By the square-integrability of the indicator functions of the sets in $\mathcal{C}$ one has that for any such finite collection of sets the multivariate central limit theorem guarantees that the asymptotic distribution of the vector $(G_n C_1, \ldots G_n C_k) \leftrightarrow (G_n(x, l, u)_1, \ldots G_n(x, l, u,)_k)$ is multivariate normal. Coupled with the tightness result above, this fact proves the convergence to a Gaussian process, and the result is still valid when one restricts $\mathcal{C}$ to contain only sets such that $P(u < Z \leqslant l) \geqslant \epsilon > 0$ (since the VC-class property is preserved). Note that if one assumes that the support of Z is some interval $[Z_{\min}, Z_{\max}]$, then the latter restriction is equivalent to the requirement that $u - l \geqslant \eta > 0$, as shown in Figure 1. The map $P(x, l, u) \mapsto F(x; l, u) = P(x, l, u)/\int_{x \in \Re} \mathrm{d}P(x, l, u)$ is therefore Hadamard-differentiable (since the denominator is bounded away from zero). The differentiability of the map follows from the linearity of the integral and from the chain rule theorem (see Theorem 20.9 in van der Vaart, 1998). Application of the functional delta method then completes the proof. The Kaplan–Meier estimator of survival is Hadamard-differentiable (see for example pp. 372 and 392 in van der Vaart and Wellner, 1996), where one has to specialize X to be the pair $(S \wedge C, 1(S < U))$ with S survival time and U an independent censoring time.

A.2 *Asymptotic distribution of estimates of $S(t^*)$*

Recall the following notation from Section 4. In addition, define the cumulative hazard function $\Lambda(t) = \int_0^t (1 - F(u^-))^{-1}\,\mathrm{d}F(u)$, the counting process $N_i(t) = 1(X_i \leqslant t, \delta_i = 1)$, and the compensator $A_i = \int_0^t Y_i(s)\,\mathrm{d}\Lambda(s)$ (where $Y_i(t) = 1(X_i \geqslant t)$) and $M_i(t) = N_i(t) - A_i(t)$. If we indicate the sums over all observations of the $Y(t)$, $N_i(t)$, $A_i(t)$, and $M_i(t)$ by $\overline{Y}(t)$, $\overline{N}(t)$, $\overline{A}(t)$, and $\overline{M}(t)$ respectively, then it is well known that $M(t) = N(t) - \int_0^t \overline{Y}(s)\,\mathrm{d}\Lambda(s) = \sum_{i=1}^{n}(N_i(s) - \int_0^s 1(X_i \geqslant u)\,\mathrm{d}\Lambda(y))$. The quantity $\widehat{S}(t) - S(t)$ can be approximated by the quantity

$$S(t)\int_0^t \frac{\widehat{S}(s^-)}{S(s)} \frac{1(\overline{Y}(s) > 0)}{\overline{Y}(s)}\,\mathrm{d}M(s) \qquad \text{(A.1)}$$

(see for example Fleming and Harrington, 1991, p. 37). By the uniform convergence of $\widehat{S}(t)$ to $S(t)$ and the Glivenko–Cantelli theorem, the integrand in the expression above (multiplied by n) converges uniformly to the function $h(s) = (P(X \geqslant s))^{-1}$. One thus has

$$\widehat{S}(t) - S(t) \approx \frac{1}{n} S(t) \int_0^t h(s)\,\mathrm{d}M(s) = S(t) \frac{1}{n} \sum_{i=1}^{n} \eta_i,$$

with $\eta_i = \int_0^t h(s)\,\mathrm{d}(N_i(s) - \int_0^s 1(X_i \geqslant u)\,\mathrm{d}\Lambda(u))$ being only a function of i, and thus the collection of the η_i being i.i.d. random variables. Now, consider the two conditional survival distributions $S_j(t) = S(t|\mathcal{P}_j)$, $j = 1, 2$, let $n_j = \sum 1(i \in \mathcal{P}_j)$, and extend the notation above to $h_j(s)$, $\Lambda_j(s)$, $N_j(s)$ and η_{ji} to indicate the corresponding quantities within the two subpopulations (set $\eta_{ji} = 0$ if $i \notin \mathcal{P}_j$). One then has

$$\begin{bmatrix} \sqrt{n_1}(\widehat{S}_1(t) - S_1(t)) \\ \sqrt{n_2}(\widehat{S}_2(t) - S_2(t)) \end{bmatrix} \approx \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\frac{n}{n_1}} S_1(t) \sum_{i=1}^{n} \eta_{1i} \\ \sqrt{\frac{n}{n_2}} S_2(t) \sum_{i=1}^{n} \eta_{2i} \end{bmatrix},$$

and assuming that $n_j/n \overset{n \to \infty}{\longrightarrow} p_j$ for $j = 1, 2$, by Slutsky's theorem and by the multivariate central limit theorem the left-hand side converges in distribution to the $N_2(0, A\Sigma A)$ distribution, with

$$A = \begin{bmatrix} S_1(t)/\sqrt{p_1} & 0 \\ 0 & S_2(t)/\sqrt{p_2} \end{bmatrix} \text{ and } \Sigma = E((\eta_1\ \eta_2)'(\eta_1\ \eta_2)).$$

Note how the asymptotic normality of the collection of estimators also follows from Theorem 1, but not the explicit expression of their asymptotic variance–covariance matrix. The matrix A can be estimated consistently by replacing $\widehat{S}_j(t)$ and n_j/n in its expression. The matrix Σ can also be estimated consistently by the quantity

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} (\widehat{\eta}_{1i})^2 & (\widehat{\eta}_{1i}\widehat{\eta}_{2i}) \\ (\widehat{\eta}_{1i}\widehat{\eta}_{2i}) & (\widehat{\eta}_{2i})^2 \end{bmatrix}$$

where the $\widehat{\eta}_{ji}$ can be obtained by plugging in $\widehat{h}_j(s) = n_j / \sum_{i=1}^{n}(1(X \geqslant s)1(i \in \mathcal{P}_j))$ for $h_j(s)$ and by replacing $\Lambda_j(s)$ by Nelson's estimator $\widehat{\Lambda}_j = \int_0^s \frac{\mathrm{d}\overline{N}_j(u)}{\overline{Y}_j(u)}$. Since the stochastic process $\overline{N}_j(s)$ has jumps of size $\delta_i 1(i \in \mathcal{P}_j)$ at the X_i and the process $\widehat{\Lambda}_j(s)$ has jumps of size $\delta_i 1(i \in \mathcal{P}_j)/\sum_{h=1}^{n} 1(X_h \geqslant X_i)1(h \in \mathcal{P}_j)$ at the X_i, after some manipulation the final expression for $\widehat{\eta}_{ji}$ is

$$\widehat{\eta}_{ji} = 1(i \in \mathcal{P}_j) \left(\frac{n_j \delta_i 1(X_i \leqslant t)}{\sum_{h=1}^{n} 1(X_h \geqslant X_i)1(h \in \mathcal{P}_j)} - \sum_{w=1}^{n} \left(\frac{1(w \in \mathcal{P}_j)1(X_i \geqslant X_w)\delta_w 1(X_w \leqslant t)}{\left[\sum_{h=1}^{n} 1(X_h \geqslant X_w)1(h \in \mathcal{P}_j)\right]^2} \right) \right).$$

An alternative resampling approach can also be used to obtain samples from the asymptotic distribution described above. For simplicity, consider the case of two overlapping subgroups. As in (A.1) above, one can write

$$\sqrt{n}\left(\widehat{S}(t) - S(t)\right) \approx -S(t) \int_0^t \frac{\widehat{S}(u^-)\sqrt{n}}{S(u)\overline{y}(u)}\,\mathrm{d}\overline{M}(u) = -S(t) \sum_{i=1}^{n} \int_0^t \frac{\widehat{S}(u^-)\sqrt{n}}{S(u)\overline{y}(u)}\,\mathrm{d}M_i(u),$$

where the quantity $H_i = (\widehat{S}(u^-)\sqrt{n})/(S(u)\overline{y}(u))$ is a predictable process (since left-continuous). It is well known that the right-hand side is a martingale with respect to t, and that it asymptotically follows a normal distribution. Also, because of the strong consistency of $\widehat{S}(t)$, the asymptotic distribution does not change if one substitutes the terms $G_i \, \mathrm{d}N_i$ for the $\mathrm{d}M_i$, where the G_i, $i = 1, \ldots, n$ are standard normal random variables (see Tsiatis, 1981). After approximating $S(u)$ by $\widehat{S}(u^-)$, and by the definition of the counting processes N_i, one thus has

$$\sqrt{n}\left(\widehat{S}(t) - S(t)\right) \approx -S(t) \sum_{i=1}^{n} \left\{ \int_0^t \frac{\widehat{S}(u^-)\sqrt{n}}{S(u)\overline{y}(u)} \, \mathrm{d}N_i(u) \right\} G_i \approx -\widehat{S}(t) \sum_{i=1}^{n} \left\{ \frac{\sqrt{n}\delta_i 1(Y_i \leqslant t)}{\overline{y}(X_i)} \right\} G_i .$$

The asymptotic variance of the left-hand side can then be estimated by generating many samples of n independent standard normal random variables. Observe that the procedure is clearly not necessary if one is only interested in the asymptotic variance of the Kaplan–Maier estimator computed for *one* subpopulation (or for non-overlapping subpopulations), since Greenwood's formula can be used directly in those cases. If we now call $\mathcal{P}_1$ and $\mathcal{P}_2$ two (possibly overlapping) sets of indices that define two patient populations, then (with obvious notation) we can use the procedure above to estimate the covariance matrix of the bivariate vector of the survival estimates in the two subgroups. In fact,

$$\begin{bmatrix} \sqrt{n_1}\left(\widehat{S}_1(t) - S_1(t)\right) \\ \\ \sqrt{n_2}\left(\widehat{S}_2(t) - S_2(t)\right) \end{bmatrix} \approx \begin{bmatrix} -S_1(t) \sum_{i=1}^{n_1} \int_0^t \frac{\widehat{S}_1(u^-)\sqrt{n_1}}{S_1(u)\overline{y}_1(u)} \, \mathrm{d}M_i(u) \\ \\ -S_2(t) \sum_{i=1}^{n_2} \int_0^t \frac{\widehat{S}_2(u^-)\sqrt{n_2}}{S_2(u)\overline{y}_2(u)} \, \mathrm{d}M_i(u) \end{bmatrix},$$

where $\overline{y}_1(u) = \sum_{k=1}^{n} 1(X_k \geqslant u) 1(k \in \mathcal{P}_1)$, and similarly for $\mathcal{P}_2$. Or finally,

$$\sqrt{n} \begin{bmatrix} \left(\widehat{S}_1(t) - S_1(t)\right) \\ \\ \left(\widehat{S}_2(t) - S_2(t)\right) \end{bmatrix} \approx \begin{bmatrix} -\widehat{S}_1(t) \sum_{i=1}^{n} \frac{\sqrt{n}\delta_i 1(Y_i \leqslant t) 1(i \in \mathcal{P}_1)}{\overline{y}_1(X_i)} G_i \\ \\ -\widehat{S}_2(t) \sum_{i=1}^{n} \frac{\sqrt{n}\delta_i 1(Y_i \leqslant t) 1(i \in \mathcal{P}_2)}{\overline{y}_2(X_i)} G_i \end{bmatrix}.$$

REFERENCES

AEBI, S., GELBER, S., CASTIGLIONE-GERTSCH, M. *et al.* (2000). Is chemotherapy alone adequate for young women with oestrogen-receptor-positive breast cancer? *The Lancet* **355**, 1869–1874.

BETENSKY, R. A. AND RABINOWITZ, D. (1999). Maximally selected χ^2 statistics for $k \times 2$ tables. *Biometrics* **55**, 317–320.

BONETTI, M. AND GELBER, R. (2000). A graphical method to assess treatment–covariate interactions using the Cox model on subsets of the data. *Statistics in Medicine* **19**, 2595–2609.

BOWMAN, A. AND WRIGHT, E. (2000). Graphical exploration of covariate effects on survival data through nonparametric quantile curves. *Biometrics* **56**, 563–570.

COATES, A. S., GOLDHIRSCH, A. AND GELBER, R. D. FOR THE INTERNATIONAL BREAST CANCER STUDY GROUP (2002). Overhauling the breast cancer overview: are subsets subversive? *The Lancet Oncology* **351**, 1451–1467.

COHEN, J. (2003). AIDS vaccine trial produces disappointment and confusion. *Science* **299**, 1290–1291.

COLLEONI, M., COATES, A., PAGANI, O. AND GOLDHIRSCH, A. (1998). Combined chemoendocrine adjuvant therapy for patients with operable breast cancer: still a question? *Cancer Treatment Reviews* **24**, 15–26.

EARLY BREAST CANCER TRIALISTS' COLLABORATIVE GROUP (1998a). Tamoxifen for early breast cancer: an overview of the randomised trials. *The Lancet* **351**, 1451–1467.

EARLY BREAST CANCER TRIALISTS' COLLABORATIVE GROUP (1998b). Polychemotherapy for early breast cancer: an overview of the randomised trials. *The Lancet* **352**, 930–942.

FISHER, B., DIGNAM, J., WOLKMARK, N. *et al*. (1997). Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor positive breast cancer. *Journal of the National Cancer Institute* **89**, 1673–1682.

FLEMING, T. R. AND HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

GAIL, M. AND SIMON, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 361–372.

GOLDHIRSCH, A., GELBER, R. D., YOTHERS, G. *et al*. (2001). Adjuvant therapy for very young women with breast cancer: need for tailored treatments. *Journal of the National Cancer Institute Monographs* **30**, 44–51.

GRAY, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942–951.

INTERNATIONAL BREAST CANCER STUDY GROUP (2002). Endocrine responsiveness and adjuvant therapy for postmenopausal node-negative breast cancer: an IBCSG randomized trial. national cancer institute. *Journal of the National Cancer Institute* **94**, 1054–1065.

ISIS-1 COLLABORATIVE GROUP (1986). Randomized trial of intravenous atenolol among 16027 cases of suspected acute myocardial infarction: ISIS-1. *The Lancet* **2**, 57–66.

ISIS-2 COLLABORATIVE GROUP (1988). Randomized trial of intravenousstreptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. *The Lancet* **2**, 349–360.

KAPLAN, E. L. AND MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

KIM, H. T. AND TRUONG, Y. K. (1998). Nonparametric regression estimates with censored data: local linear smoothers and their applications. *Biometrics* **54**, 1434–1444.

LUDWIG BREAST CANCER STUDY GROUP (1984). Randomized trial of chemo-endocrine therapy, endocrine therapy, and mastectomy alone in postmenopausal patients with operable breast cancer and axillary node metastasis. *The Lancet* **1**, 1256–1260.

MILLER, R. AND SIEGMUND, D. (1982). Maximally selected chi square statistics. *Biometrics* **38**, 1011–1016.

PETO, R. (1982). Statistical aspects of cancer trials. In Halnan, K. E. (ed.), *Treatment of Cancer*, London: Chapman and Hall, pp. 867–871.

TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *The Annals of Statistics* **9**, 93–108.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

[*Received February 10, 2003; first revision July 10, 2003; second revision February 5, 2004; accepted for publication February 12, 2004*]