STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:2595–2609

A graphical method to assess treatment–covariate interactions using the Cox model on subsets of the data

Marco Bonetti*,† and Richard D. Gelber

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA 02115, U.S.A.

SUMMARY

We introduce the subpopulation treatment effect pattern plot (STEPP) method, designed to facilitate the interpretation of estimates of treatment effect derived from different but potentially overlapping subsets of clinical trial data. In particular, we consider sequences of subpopulations defined with respect to a covariate, and obtain confidence bands for the collection of treatment effects (here obtained from the Cox proportional hazards model) associated with the sequences. The method is aimed at determining whether the magnitude of the treatment effect changes as a function of the values of the covariate. We apply STEPP to a breast cancer clinical trial data set to evaluate the treatment effect as a function of the oestrogen receptor content of the primary tumour. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Clinical trials are often conducted to compare treatments with respect to their effect on survival. Analyses of the data usually focus on the estimation of outcome for the entire study population, to avoid overinterpretation arising from subgroup analyses. On the other hand, some indication of quantitative differences in treatment effect according to subpopulations is extremely useful for designing future studies and for assisting with risk-benefit considerations in selection of therapy today. Because the subsets are usually defined with respect to one or more covariates, such analyses amount to the study of treatment–covariate interactions. A description of the issues involved can be found, among others, in Byar and Green [1] and in Peto [2].

One widely-used approach to quantifying treatment effects in clinical trials is through the Cox proportional hazards (PH) model [3]. Consider the usual setting in failure time data in which we observe $X = \min(T, C)$, with T (failure time) and C (censoring time) independent. Let $\delta = 1(X = T)$,

*Correspondence to: Marco Bonetti, Department of Biostatistical Science, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, U.S.A.

† E-mail: bonetti@jimmy.harvard.edu

Contract/grant sponsor: National Cancer Institute; contract/grant number: CA-06516, CA-75362
Contract/grant sponsor: American-Italian Cancer Foundation; contract/grant number: AICF 101-98 and 101-99
Contract/grant sponsor: International Breast Cancer Study Group

Received July 1998
Accepted November 1999

and let Z be a collection of covariates. The vectors (X_i, δ_i, Z_i), observed on n individuals, are assumed to be i.i.d. The PH model [3] assumes that the hazard function for failure time T for the individual i having covariate vector Z_i is $\lambda(t, Z_i) = \lambda_0(t)\exp(\beta' Z_i)$, $i = 1, 2, \ldots, n$. The effect of a covariate on the treatment effect can be studied by including in the PH model a term for the interaction between treatment and the covariate [3, 4]. This approach requires proportional hazards and an underlying multiplicative structure to maintain the nominal significance level for the interaction test. As an alternative, Schemper [5] extends the non-parametric approach first introduced by Patel and Hoel [6] to the analysis of treatment–covariate interaction in the presence of censoring. The variance estimate for the proposed test statistic is obtained by the jack-knife procedure. Extensions to modelling the non-linearity of the covariate effects have been proposed for the PH model, in particular by using the generalized additive model (GAM) approach introduced by Hastie and Tibshirani [7] (see for example Gray [8] or Sleeper and Harrington [9]). Modelling of the interaction between a dichotomous covariate (for example, treatment) and a continuous covariate (for example, age) within the GAM framework, however, is not as immediate. Gray [8] considers such a situation, but he limits the discussion to the study of whether the effect of a continuous covariate is different within, say, treatment groups. (He suggests fitting a separate spline function to the covariate in each group). However, it seems possible to plot the spline-estimated treatment effect as a function of the covariate values.

If the covariate of interest is not continuous, the interaction with the treatment effect consists of changes in such effect among the subgroups of patients defined by the values of the covariate. Gail and Simon [10] develop a likelihood ratio test to detect interactions between treatment effects and patient subsets when such subsets are disjoint and specified in advance. If the covariate of interest is continuous, then its range may be split into two parts, and differences in treatment effect examined between the two subsets. Koziol and Wu [11] propose a method to determine a cut-off point for dividing the covariate axis into the two categories for assessing the treatment–covariate interaction.

The multiplicity problems associated with subset analyses are typical of frequentist statistical techniques. The Bayesian approach, on the other hand, does not present such issue of the control of error rates, because the conclusions drawn relative to one subset need not depend on whether one will also draw conclusions about other subsets. Dixon and Simon [12] discuss such an approach and apply it to the study of variation in treatment effect among patient subsets. They develop the linear case under the assumption of normality of the prior distributions of the parameters and of exchangeability among the interactions.

We propose an alternative to these approaches; to reduce the risk that individual subgroup analyses might be overinterpreted, we suggest that patterns of treatment response, which might differ according to a continuum of values of a covariate of interest, could be examined. The method that we propose is based on dividing the observations into subgroups defined with respect to the covariate of interest, and fitting the PH model separately on each subpopulation. To increase the number of patients that contribute to each point estimate, we allow the subpopulations to overlap. This increases the precision of the individual estimates. For simplicity we limit our discussion to the case of two treatment groups. We consider the collection of the hazards ratios for the treatment effect as we move across the subpopulations as a way of illustrating the influence of the covariate on the treatment effect itself.

We fit the PH model on each of p subpopulations $\mathscr{P}_l$, $l = 1, \ldots, p$, defined with respect to one or more non-time-varying covariates, and call $\hat{\beta}_l$ the corresponding estimate for each vector of regression coefficients β_l corresponding to the vector Z of covariates in subpopulation $\mathscr{P}_l$. We

then have a collection of PH models $\lambda_l(t, Z_i) = \lambda_0(t)\exp(\beta_l' Z_i)$, $i = 1, 2, \ldots, n_l$, where n_l is the sample size of subpopulation $\mathscr{P}_l$. Let n be the total sample size. In the Appendix we derive the joint asymptotic distribution of the estimators $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. In particular, we are interested here in the case in which one of the components of each vector β_l corresponds to a treatment effect. Because we are interested in having the subpopulations overlap, we further assume that all the β_l are equal. This is necessary to avoid the problem related to the inability to determine which model represents the stochastic mechanism of an observation when the corresponding individual is contained in more than one subpopulation. Under such hypothesis all the $\hat{\beta}_l$, $l = 1, \ldots, p$ are estimates of the same quantity. We are interested in the study of possible deviations from this null hypothesis as an exploratory tool for the identification of treatment–covariate interactions.

In Section 2 we discuss further the study of treatment effects on overlapping subpopulations of the patient cohort, and the strictly related problem of simultaneous inference in subset analyses. In Section 3 we illustrate the method on clinical trial data, both on the study of a treatment–covariate interaction and on a subset analysis. In Section 4 we give a summary. In the Appendix we prove the main result, and show an application of a variation of the method (that focuses on the extreme values of the covariate of interest) to the same clinical trial data.

2. STEPP: SUBPOPULATION TREATMENT EFFECT PATTERN PLOT

Let the subpopulations be defined with respect to an additional continuous or ordered categorical non-time-varying covariate Z^*, and let Z_i^* be the value of such covariate for patient i. When we plot estimated treatment effects corresponding to each subpopulation we obtain what we call 'STEPP', for subpopulation treatment effect pattern plot. Because of its wide use in comparing treatment groups in clinical trials, here we discuss treatment effects quantified by hazards ratios estimated through the PH model.

Our focus is on the practical interpretation of data arising from clinical trials to increase their usefulness, both for patient care purposes and to stimulate future clinical research. With this in mind, two ways are proposed as possible choices for defining subpopulations: the 'sliding-window' pattern and the 'tail-oriented' pattern.

These two different patterns of subpopulations are illustrated in Figure 1, and are indicated by (a) and (b), respectively. The horizontal axis in Figure 1 indexes the various subpopulations for which treatment effects are estimated, and the vertical axis shows the range of the covariate values used to define the cohort of patients included in each subpopulation.

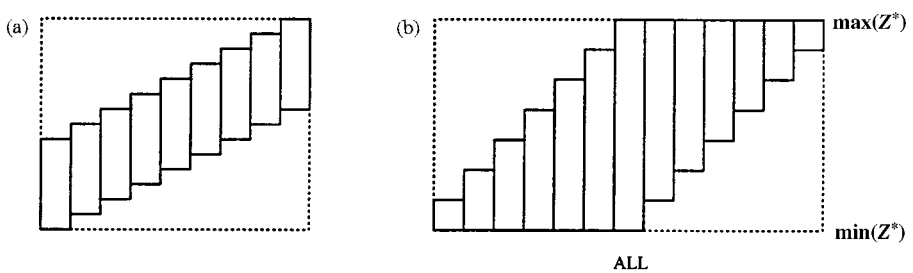

Figure 1. Illustration of the two subpopulation patterns for STEPP: (*a*) sliding window; (*b*) tail oriented.

2.1. Sliding window version

We first define subpopulation $\mathscr{P}_l$ as containing patients having $Z^* \in [\eta_l^{\text{low}}, \eta_l^{\text{upp}})$, where the definition of the values η_l^{low} and η_l^{upp} for $l = 1, \ldots, g$ will now be described. The goals in the selection of the subpopulations are to simultaneously ensure that each subpopulation contains a 'large enough' number of observations, and that the number of the subpopulations is large enough to provide a 'good' resolution over the range of the covariate of interest. These requirements can hardly be formalized, and are best left to the analyst's choice. Another desirable property seems to be that each subpopulation contains roughly the same number of patients.

We propose the following automated procedure. Let $\eta_1, \eta_2, \ldots, \eta_{\max}$ be the ordered different values of Z^* observed in the data; let $\eta_{\max} = \sup\{Z_i^*, i = 1, \ldots, n\}$. For convenience, call $\eta_0 = \eta_1 - 1$. Proceed as follows:

1. Fix two quantities n_1 and n_2, $n_1 < n_2 < n$, where n is the number of individuals in the sample.
2. Identify among the values η_t, $t = 1, \ldots, \max$ the smallest one for which $\sum_{i=1}^{n} 1(Z_i^* \leqslant \eta_t) \geqslant n_2$, and call it η_1^{upp}. Define $\eta_1^{\text{low}} = \eta_0$, and let $b = 2$. The counter b will be used to index the subpopulations.
3. 'Slide' from left to right as follows:
   (a) Identify η_b^{low} as the smallest of the η for which $\sum_{i=1}^{n} 1(Z_i^* \leqslant \eta_{b-1}^{\text{upp}}) 1(Z_i^* > \eta_b^{\text{low}}) \leqslant n_1$.
   (b) Identify η_b^{upp} as the smallest of the η for which $\sum_{i=1}^{n} 1(Z_i^* \leqslant \eta_b^{\text{upp}}) 1(Z_i^* > \eta_b^{\text{low}}) \geqslant n_2$. If there is no such value let $\eta_b^{\text{upp}} = \eta_{\max}$ and stop.
4. Repeat step 3 after increasing b by 1.

Define subpopulation $\mathscr{P}_b$ as containing all patients for whom $\eta_b^{\text{low}} < Z_i^* \leqslant \eta_b^{\text{upp}}$. Given this algorithm, the value chosen for n_2 roughly defines how many patients are included in each subpopulation. Observe that there is a trade-off between resolution over the range of Z^* and variability in the estimates of the regression coefficients. The difference between n_2 and n_1 describes the minimum number of subjects replaced between any two subsequent subpopulations. The choice of the values n_1 and n_2 can generate a variety of different subpopulations, whose number will also change. Windows include many subjects if n_2 is large, and there is little patient turnover from window to window if n_1 is close to n_2; this scenario provides relatively precise estimates of treatment effect and a large number of windows.

The choice of n_1 and n_2 determines the number of subpopulations. Define $n_j = p_j n$, $j = 1, 2$, for two proportions $0 < p_1 < p_2 < 1$. If we assume that there are no ties in the values Z_i^* (that is, if $P(Z_i^* = Z_j^*,\ i \neq j) = 0$), it is easy to show that as $n \to \infty$ the number of subpopulations defined by the algorithm described above tends to the smallest integer greater than or equal to the number $[1 + (1 - p_2)/(p_2 - p_1)]$, with the last subpopulation containing a proportion of the patients at most equal to the desired p_2. If ties are present, however (as is the common case of Z^* discrete), this criterion will necessarily suffer from the discontinuities, so that even as $n \to \infty$ the proportions p_1 and p_2 may never be achieved exactly. Observe how the situation $n_2 = n$ (and n_1 free) produces only one subpopulation, which contains all patients.

2.2. Tail-oriented version

As an alternative to the sliding-window STEPP, the tail-oriented pattern plot concentrates on the influence of extreme values of the covariate on the magnitude of the treatment effect. Consider a set of increasing values of Z^* $\{z_1, z_2, \ldots, z_g\}$; we construct an increasing collection of subpopulations

$\mathscr{P}_l,\ l=1,2,\ldots,g$ by including in $\mathscr{P}_l$ the patients for whom $Z_i^* \leqslant z_l$ (in the notation of the proof in Appendix A, $1_i(\mathscr{P}_l)=1(Z_i^* \leqslant z_l)$). Similarly, we construct the subpopulations $\mathscr{P}_l,\ l=g+1,\ldots,2g-1$ by including in $\mathscr{P}_l$ the patients for whom $Z_i^* > z_{l-g}$. In what follows we call p the resulting total number of subpopulations $(2g-1)$. By taking $z_g = \sup\{Z_i^*,\ i=1,2,\ldots,n\}$ we ensure that $\mathscr{P}_g$ will contain all the patients.

In applications, we suggest choosing the values $\{z_1, z_2, \ldots, z_g\}$ so that they divide the patients in groups of roughly equal sample sizes. Although this procedure is data-dependent in terms of the actual values of the covariate, it does not use any outcome information. In some cases, a large enough number of subjects have the same covariate value (for example, age in years), and cut-offs at each of the observed values of the covariate might be appropriate. This option achieves the maximum possible resolution for the analysis. Lastly, cut-off values may be chosen to correspond with 'usual practice,' when specific subpopulations have historically been used in a particular disease setting. This method has the advantage of facilitating the communication of the results to clinical investigators (who often think in terms of cohorts of patients defined with respect to increasingly larger or smaller values of a covariate) and the comparison of such results with previous studies. Observe that as a result of the definition above, the 'pivotal' subpopulation displayed in the centre of the plot contains all patients.

Considering the construction and interpretation of the tail-oriented version of STEPP and the exploratory nature of the method, we recommend that any testing procedure be applied separately to the left and right parts of the plot. An application of this approach is illustrated in Appendix B.

2.3. *Confidence bands and hypothesis testing*

Let $\hat{\beta}_l^*$ be the component of $\hat{\beta}_l$ corresponding to the treatment effect for subpopulation $\mathscr{P}_l$. The plot of the estimated hazards ratios $\hat{\theta}_l = \exp(\hat{\beta}_l^*)$ will in general be hard to interpret, due to the varying sample sizes in the subpopulations, and especially due to the correlation among the estimates. This is true in particular when the amount of overlapping between the subpopulations is large. A confidence band constructed from the sequence of the hazards ratios $\hat{\theta}_l$ can help in the interpretation, and it can be obtained immediately from a confidence band obtained from the $\hat{\beta}_l^*$. If we let $\sigma_l = [\mathrm{var}(\hat{\beta}_l^*)]^{1/2}$, the marginal asymptotic 95 per cent confidence interval for each β_l^* is given by $\{\beta_l^* \in \hat{\beta}_l^* \pm 1.96\,\sigma_l\}$. We choose to construct a 95 per cent confidence band that is based on rectangular simultaneous confidence intervals for $\beta_1^*, \beta_2^*, \ldots, \beta_p^*$. We define the band as $\{\beta_l^* \in \hat{\beta}_l^* \pm \gamma 1.96\sigma_l,\ l=1,\ldots,p\}$, that is, with widths proportional to the widths of the marginal confidence intervals. The value of γ is such that $P[\bigcap_{l=1}^{p}\{\beta_l^* \in \hat{\beta}_l^* \pm \gamma 1.96\sigma_l\}] = 0.95$, and it can easily be obtained through simulation from the asymptotic distribution of the estimators. Observe that γ can be regarded as a measure of the effect of the simultaneous inference on the width of the confidence intervals.

An *omnibus* test for the equality of the coefficients across subpopulations can be obtained by considering the transformation $\Delta = A\beta^*$ such that $\Delta_j = \beta_{j+1}^* - \beta_j^*, \quad j=1,2,\ldots,p-1$. Under the null hypothesis $H_0 : \Delta = 0$, if we call Σ the estimated asymptotic variance matrix of $\hat{\beta}^*$, the test statistic $G = \hat{\Delta}'(A\Sigma A')^{-1}\hat{\Delta}$ is (approximately) χ^2_{p-1}-distributed.

Knowledge of the asymptotic distribution of the estimators allows one to apply other test statistics that might seem appropriate for a particular alternative hypothesis, since the distribution of any test statistic can be estimated via simulations.

2.4. Subset analysis

The result derived in Appendix A can in general be used when the subpopulations are defined arbitrarily, and in particular according to different covariates. Fitting of the PH model on such subsets of the patients is commonly done in clinical trials. Knowledge of the (asymptotic) distribution of the resulting estimators thus allows one to make the correct *simultaneous* inference on the treatment coefficients obtained from fitting the PH model on each subpopulation. We can therefore capture the additional uncertainty of the estimates due to the multiple subdivisions of the population. In Section 3 we show an example of this using the data set from the IBCSG trial.

3. AN APPLICATION

We have applied STEPP to the International Breast Cancer Study Group Trial VII, a 2×2 factorial clinical trial evaluating chemoendocrine treatment versus endocrine therapy alone for postmenopausal breast cancer patients. A total of 1212 patients from 24 institutions in 9 countries are involved in the trial. The interested reader will find a complete description of the trial in IBCSG [13]. The following covariates were collected as part of the trial: treatment; age; level of oestrogen receptor (ER, in fmol/mg of cytosol protein), and number of positive axillary lymph nodes. We have used the data from the 592 evaluable patients in two of the four arms, which represent the two treatments tamoxifen for five years and tamoxifen for five years plus three cycles (three months) of early chemotherapy given with tamoxifen. Disease-free survival (DFS) was defined as the length of time from the date of randomization to any relapse (including ipsilateral breast recurrence), the appearance of a second primary tumour (including contralateral breast cancer), or death, whichever occurred first. The median follow-up duration was 60 months, and the DFS proportions at 5 years were 55 per cent and 64 per cent for the two treatment arms, respectively. Here we concentrate on the influence of the level of ER on the treatment effect on disease-free survival.

Preliminary analyses showed that ER and nodes should be log-transformed before the PH model could be fit properly, and in what follows this will be implicit. Fitting of the PH model with all the covariates described above on all patients produced an estimated treatment group hazards ratio (HR) of 0.64 favouring tamoxifen plus chemotherapy, with a corresponding 95 per cent confidence interval equal to $[0.49, 0.83]$. The two-sided p-value associated with the HR is 0.001. The overall effect of adding the early cycles of chemotherapy to tamoxifen is thus highly significant. The interaction term between treatment and log(ER) is not significant when it is included in the regression model (p-value $= 0.84$).

3.1. STEPP

We apply STEPP in its version (a) to this data set. This version consists of fitting the PH model on subsets defined in a sliding-window fashion. (Application of the tail-oriented version is shown in Appendix B). We set $n_1 = 55$ and $n_2 = 60$. The algorithm introduced in Section 2.1 generates a total of 55 subpopulations; each subpopulation contains about 60 patients, and approximately 5 patients (60 minus 55) are exchanged as the window moves along the ER axis. The resulting treatment hazards ratio estimates (tamoxifen plus chemotherapy versus tamoxifen alone) are shown in Figure 2 together with the corresponding 95 per cent confidence band. The dashed horizontal line shows the overall treatment hazards ratio for the entire patient population, and the numbers in

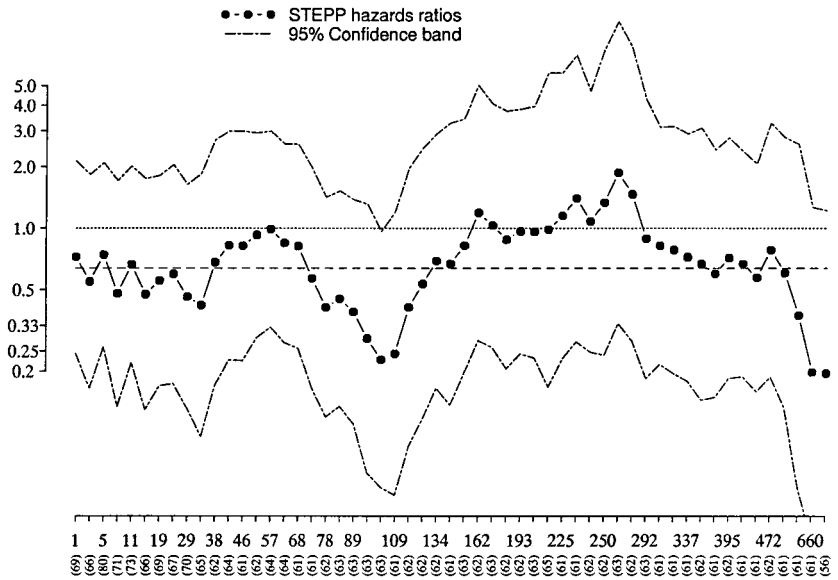

Figure 2. STEPP (sliding-window analysis) for IBCSG Trial VII data according to ER values ($n_1 = 55$, $n_2 = 60$).

Table I. Testing for various choices of n_1 and n_2 ($n = 592$).

n_1	n_2	Number of subpopulations	$\hat{\gamma}$	p-value
40	50	38	1.6375	0.3815
55	60	55	1.6375	0.5842
60	75	26	1.5375	0.2771
70	75	53	1.6125	0.8013
95	100	49	1.5875	0.0044
100	110	31	1.5375	0.4576
105	110	46	1.5625	0.2907
120	130	29	1.5125	0.0740
125	130	45	1.5375	0.2893
140	150	30	1.5125	0.1423
145	150	44	1.5125	0.2789
190	200	25	1.4625	0.0507
195	200	37	1.4625	0.0572
170	180	27	1.4625	0.6894
175	180	40	1.4625	0.4659

parentheses below the x-axis are the numbers of patients in each subpopulation. The label on the x-axis for each subpopulation is the median of the values of ER in that subpopulation. Application of the omnibus test does not reject the hypothesis of no interaction (p-value $= 0.58$).

We experimented with different values for (n_1, n_2), and observed a large variability in the results of the omnibus test as the size of the subpopulations (n_2) and the number of patients exchanged ($n_2 - n_1$) are changed. Table I shows the number of subpopulations obtained from different combinations of values n_1 and n_2, and for each combination the value of the estimate of the parameter γ

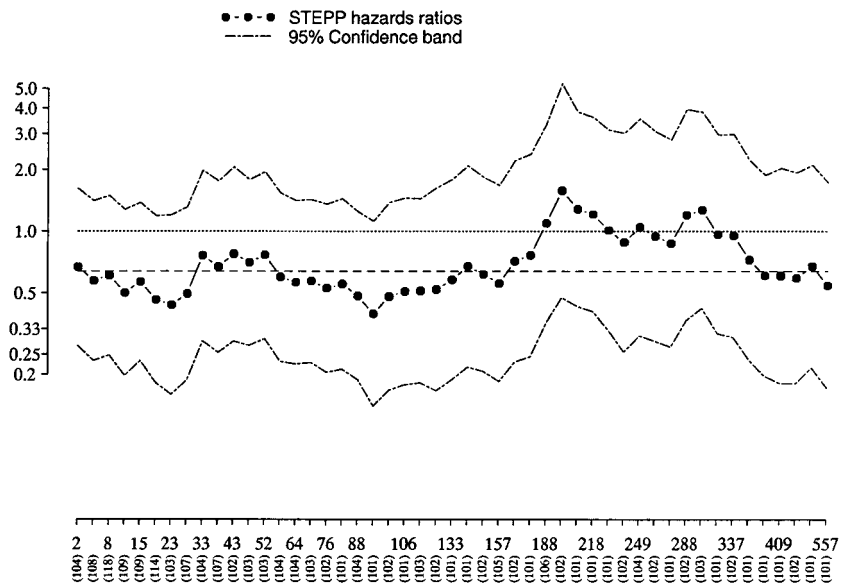

Figure 3. STEPP (sliding-window analysis) for IBCSG Trial VII data according to ER values ($n_1 = 95,\ n_2 = 100$).

and the p-value for the omnibus test. The table shows how STEPP highlights the large uncertainty associated with the process of making inference about interaction without introducing a specific functional assumption about its form. This issue, common to all smoothing techniques, suggests that STEPP should be used mainly for exploratory purposes, and as a useful hypothesis-generating tool. More generally, STEPP addresses the issue of the variability of the results obtained from subgroup analyses, the results of which should always be judged with extreme caution.

The appearance of STEPP, on the other hand, seems to have a remarkable degree of robustness to the selection of the pair (n_1, n_2), and the confidence band also does not seem to change much. Figures 3 and 4 show the plots corresponding to the choices $(95, 100)$ and $(145, 150)$ for the two parameters (n_1, n_2). The STEPP should therefore be used with confidence when trying to identify ranges of the covariate for which treatment effects may behave unusually. In this example there seems to be indication of a strong treatment effect for values of ER around 100, as well as for very large values of ER. On the other hand, for patients with tumours having ER values between the mid-100s and the mid-300s, the magnitude of the effect of adding chemotherapy to tamoxifen appears to be less than the overall estimate. Thus, some patient subgroups might not benefit as much from chemotherapy as other subgroups.

3.2. *Subset analysis*

Consider fitting the PH model on two subsets of patients $\mathscr{P}_1$ and $\mathscr{P}_2$, where $\mathscr{P}_1$ contains patients having age at least equal to 60 and $\mathscr{P}_2$ contains patients having ER level at least equal to 10. A total of 268 of the 596 patients are common to these two subpopulations, which have sizes 338 and 458, respectively. The maximum partial likelihood estimates of the treatment coefficients in the subpopulations $\mathscr{P}_1$ and $\mathscr{P}_2$ are $\hat{\beta}^*_{\mathscr{P}_1} = -0.37$ and $\hat{\beta}^*_{\mathscr{P}_2} = -0.51$, and their corresponding estimated

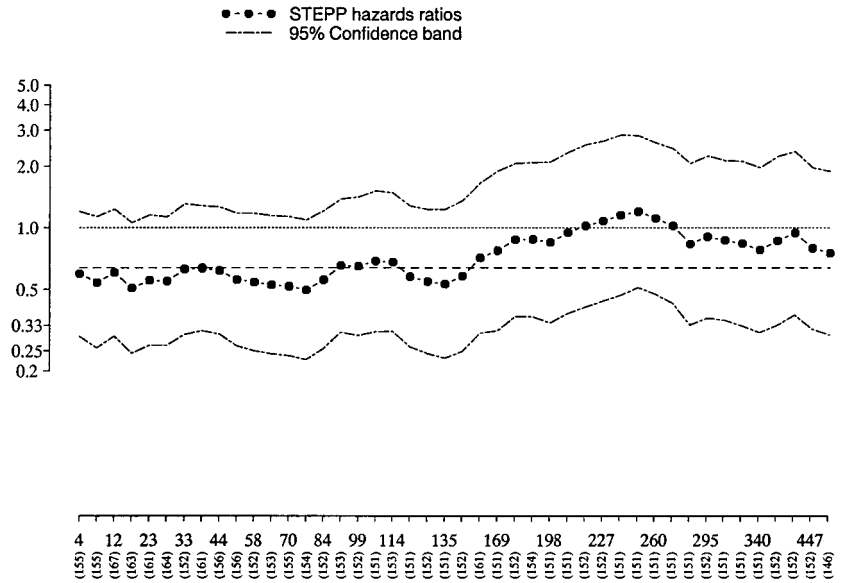

Figure 4. STEPP (sliding-window analysis) for IBCSG Trial VII data according to ER values ($n_1 = 145$, $n_2 = 150$).

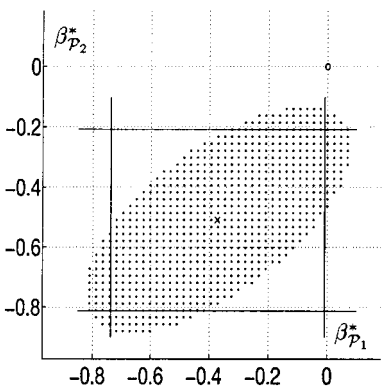

Figure 5. Joint 95 per cent confidence interval for the treatment coefficients corresponding to fitting a proportional hazards regression model on the two subpopulations of patients $\mathscr{P}_1$ (having age at least equal to 60) and $\mathscr{P}_2$ (having ER level at least equal to 10).

asymptotic covariance matrix $\hat{\Sigma}_{\hat{\beta}^*}$ can be obtained from the result described in the Appendix. A 95 per cent joint confidence interval for $\hat{\beta}^*_{\mathscr{P}_1}$ and $\hat{\beta}^*_{\mathscr{P}_2}$ can thus be obtained as shown in Figure 5. For comparison, the two marginal 95 per cent confidence intervals are also shown in Figure 5. Observe the high correlation existing between the two estimators ($r = 0.63$), which was to be expected because of the extensive overlapping of the two subpopulations.

A test for the hypothesis $H_0 : \beta^*_{\mathscr{P}_1} = \beta^*_{\mathscr{P}_2} = 0$ is rejected at the 0.05 level since the corresponding 95 per cent confidence interval does not include the point $(0, 0)$. Alternatively, one may be interested

in whether the two subpopulations have the same treatment coefficient, that is, the hypothesis $H_0': \beta^*_{\mathscr{P}_1} = \beta^*_{\mathscr{P}_2}$. Letting $a = [1, -1]'$, the value for the test statistic $\tilde{Z} = (\hat{\beta}^*_{\mathscr{P}_1} - \hat{\beta}^*_{\mathscr{P}_2})/[a'\hat{\Sigma}_{\hat{\beta}^*}a]^{1/2}$ is 0.94, and since $\tilde{Z}$ is approximately normally distributed under H_0', we conclude that we cannot reject H_0' at the 0.05 level of significance.

4. DISCUSSION

STEPP considers a sequence of treatment effects estimated on potentially overlapping subpopulations in order to understand the influence of covariates on treatment effect. STEPP is especially useful when trying to identify subpopulations of patients for whom the overall trial results may be less representative.

In this article we have discussed a form of STEPP based on quantifying treatment effects through a regression parameter in the PH model. One referee pointed out that it is impossible to clearly specify the stochastic mechanism of an observation when the corresponding individual falls in two or more overlapping subpopulations on which *different* PH models are estimated. For this reason we stress the exploratory nature of the method (rather than its use for formal estimation) when the PH model is used in STEPP. Observe that this problem is also an issue in subsets analysis. In Section 3.2, for example, a patient having both age at least equal to 60 and ER level at least equal to 10 would belong to both subpopulations $\mathscr{P}_1$ and $\mathscr{P}_2$, and it is not clear which PH model (if any) would describe the disease-free survival for this patient. Such subsets analyses are very common, and this concern is largely ignored in practice. However, further examination of this issue will be considered for future research.

When compared to the usual subset analysis approach, STEPP allows one to examine the patterns of treatment effects according to a sequence of subpopulations, rather than through the comparison of the treatment effects between only two subpopulations defined by a more or less arbitrary cut-off value. We have seen how the choice of the cut-off points that define the subgroups can have a major impact on inference.

The sample size needed for a clinical trial to be able to investigate treatment–covariate interactions without imposing parametric assumptions on the form of the interaction is quite large, and applying STEPP to a larger database from a meta-analysis would increase the precision of the estimates.

Examination of the plot can help detect deviations from the proportionality assumption if a multiplicative interaction term is used in the PH model. More in general, STEPP can be used to assess an appropriate parametric form for the treatment–covariate interaction.

The two versions of STEPP can be used together, since the two approaches complement each other. While the sliding-window approach is suitable for the exploration of interactions about which no *a priori* information is available, the tail-oriented pattern is designed to be sensitive to interactions that are likely to impact the treatment effect at extreme values of the covariate of interest. The tail-oriented version also presents the primary treatment comparison based on the entire patient population as the 'pivotal' element in the centre of the plot.

Another example of the use of smoothing techniques is provided in Thaler [14], where a non-parametric estimate of the hazard ratio function is developed. While that work does not provide any inference machinery, it bears some similarities with our sliding-window approach, in particular in its use of overlapping time intervals. As was suggested by one referee, STEPP can be extended to study time–covariate interactions, and we plan to explore that possibility in future work.

In conclusion, the use of STEPP helps in appreciating the inherent variability of subgroup analyses, and in trying to identify real shifts in treatment effect magnitude as a function of covariate values.

APPENDIX A: ASYMPTOTIC DISTRIBUTION OF THE ESTIMATORS

We make use of the multiplicative intensity model first introduced by Aalen [15]. The general form of the intensity function of the counting process $N(s)$ is $\lambda(s|Z)=\lambda_0(s)g(Z)Y(s)$, where $\lambda_0(s)$ is an unknown function, $g(Z)$ corresponds to the covariate effect, and $Y(s)$ is a stochastic process which together with $N(s)$ can be observed over the time interval of interest (see Aalen [15]). The proportional hazards (PH) model is obtained from this by specializing $g(Z)=\exp(\beta' Z)$, by defining for the censoring process $N_C(s)=I(X\leqslant s,\delta=1)$, and by setting $Y_C(t)=I(X\geqslant t)$. For such a model the process $M_C(t)=N_C(t)-A_C(t)=N_C(t)-\int_0^t \lambda_0(s)\exp(\beta' Z)Y_C(s)\,\mathrm{d}s$ is a local martingale (that is, $A_C(t)$ is the compensator of $N_C(t)$). The full development of the PH model as a version of the multiplicative intensity model can be found in Andersen and Gill [16].

Let $1_i(\mathscr{P})$ be the indicator function 1(patient $i\in\mathscr{P}$), and $\mathscr{P}$ a subpopulation defined with respect to a non-time-varying covariate. Call $M_i(t)$ the process $M(t)$ associated with the ith patient, Z_i the corresponding vector of covariates, and $Y_i(u,\mathscr{P})=I(X_i\geqslant u, 1_i(\mathscr{P})=1)$. Following the development in Fleming and Harrington (Reference [17] pp. 148–150), the score function $U(\beta,\mathscr{P})$ associated with our model for the generic subpopulation $\mathscr{P}$ is the value at $t=\infty$ of the process

$$U(\beta,t,\mathscr{P})=\sum_{i=1}^{n}\int_0^t \{Z_i-R(\beta,u,\mathscr{P})\}1_i(\mathscr{P})\,\mathrm{d}N_i(u) \tag{A1}$$

where

$$R(\beta,u,\mathscr{P})=\frac{S^{(1)}(\beta,u,\mathscr{P})}{S^{(0)}(\beta,u,\mathscr{P})}=\frac{n^{-1}\sum_{i=1}^{n} Z_iY_i(u,\mathscr{P})\exp(\beta' Z_i)}{n^{-1}\sum_{i=1}^{n} Y_i(u,\mathscr{P})\exp(\beta' Z_i)}. \tag{A2}$$

Observe that in (A1) we implicitly define a new set of covariates $Z_i(\mathscr{P})=Z_i1_i(\mathscr{P})$, and that $Z_i(\mathscr{P})1_i(\mathscr{P})=Z_i1_i(\mathscr{P})I(X_i\geqslant u)1_i(\mathscr{P})=Z_i1_i(\mathscr{P})$. Integration with respect to the processes $N_i(t)$ over the range $[0,\sup\{T_i,i=1,2,\ldots,n\}]$ is equivalent to expressing the score function as

$$U(\beta,\mathscr{P})=\sum_{i=1}^{n}\delta_i 1_i(\mathscr{P})\left(Z_i-\frac{\sum_{j=1}^{n}[I(X_j\geqslant X_i)\exp(\beta' Z_j)Z_j1_j(\mathscr{P})]}{\sum_{j=1}^{n}[I(X_j\geqslant X_i)\exp(\beta' Z_j)1_j(\mathscr{P})]}\right).$$

We can write the integrals in (A1) with respect to the processes $M_i(t)$ instead of $N_i(t)$:

$$\begin{aligned}U(\beta,t,\mathscr{P}) &= \sum_{i=1}^{n}\int_0^t \{Z_i-R(\beta,u,\mathscr{P})\}1_i(\mathscr{P})\,\mathrm{d}N_i(u)\\ &= \sum_{i=1}^{n}\int_0^t \{Z_i-R(\beta,u,\mathscr{P})\}1_i(\mathscr{P})\,\mathrm{d}M_i(t)\\ &\quad+\sum_{i=1}^{n}\int_0^t \{Z_i-R(\beta,u,\mathscr{P})\}\lambda_0(u)\exp(\beta' Z_i)Y_i(u,\mathscr{P})\,\mathrm{d}u\end{aligned}$$

where the last term can be shown to be zero. $U(\beta, t, \mathscr{P})$ can thus be expressed in the form $\sum_{i=1}^{n} \int_0^t H_i(s; \mathscr{P}) \mathrm{d}M_i(s)$ with

$$H_i(s; \mathscr{P}) = \left(Z_i - \frac{\sum_{j=1}^{n} [I(X_j \geqslant t) \mathrm{e}^{\beta' Z_j} Z_j 1_j(\mathscr{P})]}{\sum_{j=1}^{n} [I(X_j \geqslant t) \mathrm{e}^{\beta' Z_j} 1_j(\mathscr{P})]} \right) 1_i(\mathscr{P})$$

and is therefore also a martingale with respect to t (see Fleming and Harrington, Reference [17], Theorem 2.4.1). The solution to the system $U(\beta, \mathscr{P}) = 0$ is the p-dimensional maximum partial likelihood estimator (MPLE) $\hat{\beta}(\mathscr{P})$. Taylor expansion of $U(\hat{\beta}(\mathscr{P}), \mathscr{P})$ yields

$$(\hat{\beta}(\mathscr{P}) - \beta(\mathscr{P})) \simeq - \left[\frac{\partial U(\beta(\mathscr{P}), \mathscr{P})}{\partial \beta(\mathscr{P})} \right]^{-1} U(\beta(\mathscr{P}), \mathscr{P}). \tag{A3}$$

We now extend our notation to include the processes defined by the indicator functions $1_i(\mathscr{P}_l)$, where $\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_p$ are subpopulations defined with respect to some non-time-varying covariate. For $l = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$, call $\hat{\beta}_l = \hat{\beta}(\mathscr{P}_l)$, $\beta_l = \beta(\mathscr{P}_l)$, and $U_l(\beta_l) = U(\beta_l, \mathscr{P}_l)$. For definiteness we assume that the parameters in all subpopulations be all equal, but we keep the notation $\beta_1, \ldots, \beta_p$ for clarity. Call $N_i(t) = I(X_i \leqslant t,\ \delta_i = 1)$ and $Y_{li}(t) = I(X_i \geqslant t,\ 1_i(\mathscr{P}_l) = 1)$. We can repeat the expansion in (A3) for each population $\mathscr{P}_l$, and write

$$\begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \\ \vdots \\ \hat{\beta}_p - \beta_p \end{bmatrix} \simeq - \begin{bmatrix} [\frac{\partial U_1(\beta_1)}{\partial \beta_1}] & 0 & \ldots & 0 \\ 0 & [\frac{\partial U_2(\beta_2)}{\partial \beta_2}] & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & [\frac{\partial U_p(\beta_p)}{\partial \beta_p}] \end{bmatrix}^{-1} \begin{bmatrix} U_1(\beta_1) \\ U_2(\beta_2) \\ \vdots \\ U_p(\beta_p) \end{bmatrix}.$$

We are interested in the asymptotic distribution of

$$\begin{bmatrix} \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \\ \vdots \\ \hat{\beta}_p - \beta_p \end{bmatrix} \simeq nM^{-1} \begin{bmatrix} U_1(\beta_1) \\ U_2(\beta_2) \\ \vdots \\ U_p(\beta_p) \end{bmatrix} \tag{A4}$$

where M is block-diagonal, and we call each block on the diagonal $\hat{A}_l(\beta_l) = (1/n)\partial U_l(\beta_l)/\partial \beta_l$, $l = 1, 2, \ldots, p$. By Theorem 4.2 in Andersen and Gill [16], each term $\hat{A}_l(\beta_l)$ converges in probability to a non-singular deterministic matrix, which we call $A_l(\beta_l)$, which can be consistently estimated by $\hat{A}_l(\hat{\beta}_l)$ as follows:

$$\hat{A}_l(\hat{\beta}_l) = \frac{1}{n} \sum_{j=1}^{n} \delta_j 1_j(\mathscr{P}_l) \left[\frac{\sum_{i=1}^{n} Y_{li}(X_j) Z_i^{\otimes 2} \exp\{\hat{\beta}_l' Z_i\}}{\sum_{i=1}^{n} Y_{li}(X_j) \exp\{\hat{\beta}_l' Z_i\}} - \left(\frac{\sum_{i=1}^{n} Y_{li}(X_j) Z_i \exp\{\hat{\beta}_l' Z_i\}}{\sum_{i=1}^{n} Y_{li}(X_j) \exp\{\hat{\beta}_l' Z_i\}} \right)^{\otimes 2} \right]$$

where for a vector a, $a^{\otimes 2} = aa'$.

By Theorem 8.2.1 in Fleming and Harrington (Reference [17], p. 290), the normalized (by $n^{-1/2}$) score process converges to a Gaussian process, and there exists a consistent estimator for the corresponding covariance function. In particular, assume that conditions (2.1)–(2.6) in

Reference [17] pp. 289–290, hold. This guarantees that Theorem 5.3.5 also holds, and if we define the processes $H_{i,l}^{(n)}$ as being $H_{i,l}^{(n)} = [Z_i - E_l(\beta_0, x)]1_i(\mathscr{P}_l)$, simple modifications in Theorem 8.2.1 then establish the asymptotic distribution. Consistency of the MPLE of the coefficients β_l can easily be shown, and by Slutsky's theorem the result about the convergence of the score process is transferred to the estimators.

The asymptotic covariance matrix between any two terms $n^{1/2}(\hat{\beta}_l)$ and $n^{1/2}(\hat{\beta}_h)$ can then be estimated consistently by the quantity $\hat{D}_l(\hat{\beta}_l, \hat{\beta}_h) = \hat{A}_l^{-1}(\hat{\beta}_l)\hat{B}_{lh}(\hat{\beta}_l, \hat{\beta}_h)\hat{A}_h^{-1}(\hat{\beta}_h)$, where $\hat{B}_{lh}(\hat{\beta}_l, \hat{\beta}_h) = n^{-1}\sum_{j=1}^{n} W_{lj}(\hat{\beta}_l)W_{hj}(\hat{\beta}_h)'$, and

$$W_{lj}(\hat{\beta}_l) = \delta_j 1_j(\mathscr{P}_l)\left\{Z_j - \frac{S_l^{(1)}(\hat{\beta}_l, X_j)}{S_l^{(0)}(\hat{\beta}_l, X_j)}\right\} - \sum_{m=1}^{n} \frac{\delta_m 1_m(\mathscr{P}_l) Y_{lj}(X_m)\exp(\hat{\beta}_l' Z_j)}{nS_l^{(0)}(\hat{\beta}_l, X_m)} \times \left\{Z_m - \frac{S_l^{(1)}(\hat{\beta}_l, X_m)}{S_l^{(0)}(\hat{\beta}_l, X_m)}\right\}.$$

Observe that one could define a separate counting process $N_{il}(t) = I(X_i \leqslant t,\ \delta_i = 1,\ 1_i(\mathscr{P}_l) = 1)$ for each individual for each population, and express the score functions as integrals with respect to these counting processes. This approach, however, would not allow derivation of the result, since these counting processes do not constitute a proper multivariate counting process. (In fact, the condition that no two processes jump at the same time is clearly violated.)

APPENDIX B: TAIL-ORIENTED STEPP

We now apply the tail-oriented version of STEPP [(Figure 1(*b*)] to the IBCSG clinical trial data. Using 17 subpopulations (eight for decreasing ER values, one for all patients, and eight for increasing values of ER), we obtained the results shown in Figure A1, where we show the value for treatment hazards ratios (tamoxifen plus chemotherapy versus tamoxifen alone), the overall 95 per cent confidence band, and the hazards ratios for the non-overlapping subpopulations $\mathscr{P}_l^* = \mathscr{P}_l \cap \mathscr{P}_{l+1}$ for $l = 1, \ldots, p$. The latter estimates are shown in the figure as open diamonds, and they illustrate how little precision in the estimates is available when non-overlapping subpopulations are used. The dashed horizontal line shows the treatment hazards ratio for the entire patient population (labelled as 'ALL' on the *x*-axis) and the numbers in parentheses below the *x*-axis are the numbers of patients included in the subpopulations. The plot also shows, for comparison, the individual 95 per cent confidence intervals for all patients ('ALL'), and for the two subpopulations 'ER < 10' and 'ER $\geqslant 10$'. These two subpopulations represent a subgroup analysis usually performed on breast cancer adjuvant therapy clinical trials data; for this trial these subgroups were prospectively stratified at randomization. For this example, we have chosen the cut-off points for subpopulations shown in Figure A1 according to the 'usual practice' criterion, after consultation with an expert medical oncologist.

The plot in Figure A1 illustrates the pattern of treatment effect sizes relative to the 'ALL' patient value according to subpopulations defined by ER value, both for decreasing (to the left) or increasing (to the right) values of ER and for independent and non-overlapping subgroups (open diamonds). The results from this analysis indicate the possibility of a difference in the hazards ratios when we look at the subpopulations having large values of ER. Testing can be performed on each half of the plot. Testing for equality in the coefficients on the left side gives a *p*-value of 0.43,

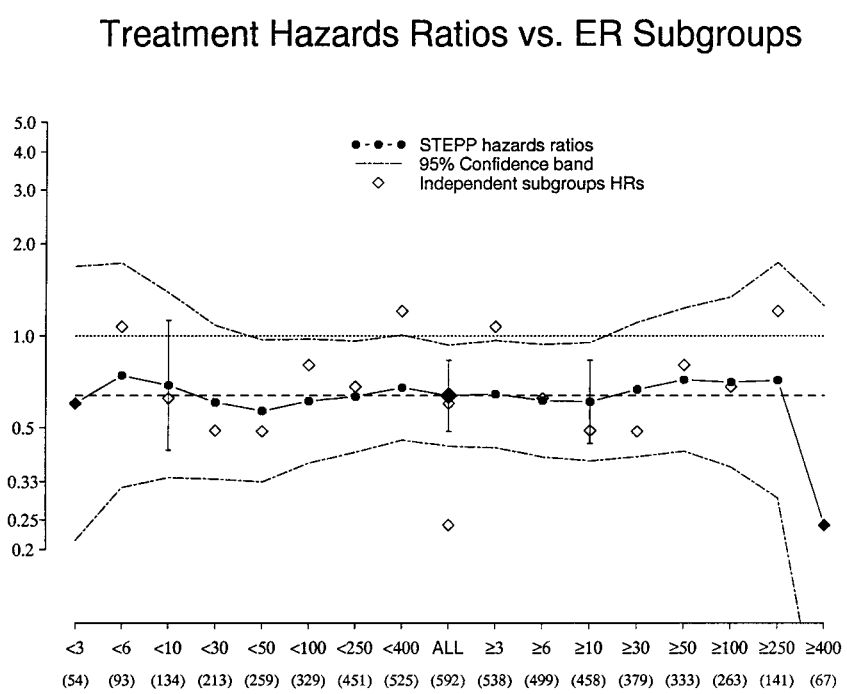

Figure A1. STEPP (tail-oriented analysis) for IBCSG Trial VII data according to ER values.

and testing the right side alone gives a p-value of 0.0004. These results indicate the possibility that for very high levels of ER the addition of chemotherapy to tamoxifen could be very effective. The confidence band for the STEPP is 43 per cent wider than the individual confidence intervals ($\gamma = 1.43$).

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 Australia-New Zealand Breast Cancer Trials Group, grants 880513, 910420 and 940892, the Australian Cancer Society, the Frontier Science and Technology Research Foundation, the Swiss Group for Clinical Cancer Research). We thank the IBCSG for their permission and encouragement to use the data from IBCSG Trial VII as an example for describing the STEPP method. We especially appreciate the help of Dr Aron Goldhirsch, who motivated the development of STEPP to provide clinically relevant analyses of data from randomized clinical trials, and of Dr Monica Castiglione-Gertsch, who provided the study co-ordination and medical review for Trial VII. We also thank the two referees who provided stimulating comments that led to increased appreciation of issues concerning subgroup analyses.

REFERENCES

1. Byar D, Green S. The choice for treatment for cancer patients based on covariate information: Application to prostate cancer. *Bulletin du Cancer* 1980; **67**:477–490.

2. Peto R. Statistical aspects of cancer trials. *Treatment of Cancer*, In Halnan KE (ed.). Chapman and Hall: London, 1982, pp. 867–871.
3. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1972; **34**:187–220.
4. Shuster J, van Eys J. Interaction between prognostic factors and treatment. *Controlled Clinical Trials* 1983; **4**:209–214.
5. Schemper M. Non-parametric analysis of treatment–covariate interaction in the presence of censoring. *Statistics in Medicine* 1988; **7**:1257–1266.
6. Patel KM, Hoel DG. A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association* 1973; **68**:615–620.
7. Hastie T, Tibshirani R. Generalized additive models (c/r: P310-318). *Statistical Science* 1986; **1**:297–310.
8. Gray RJ. Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**:942–951.
9. Sleeper LA, Harrington DP. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* 1990; **85**:941–949.
10. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**:361–372.
11. Koziol JA, Wu SH. Changepoint statistics for assessing a treatment-covariate interaction. *Biometrics* 1996; **52**:1147–1152.
12. Dixon DO, Simon R. Bayesian subset analysis (corr: 94v50 p322). *Biometrics* 1991; **47**:871–881.
13. International Breast Cancer Study Group. Effectiveness of adjuvant chemotherapy in combination with tamoxifen for node-positive postmenopausal breast cancer patients. *Journal of Clinical Oncology* 1997; **15**:1385–1394.
14. Thaler HT. Nonparametric estimation of the hazard ratio. *Journal of the American Statistical Association* 1984; **79**:290–293.
15. Aalen O. Nonparametric inference for a family of counting processes. *Annals of Statistics* 1978; **6**:701–726.
16. Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study (com: P1121-1124). *Annals of Statistics* 1982; **10**:1100–1120.
17. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.

