Biometrics 59, 859–868
December 2003

# Inference in Spline-Based Models for Multiple Time-to-Event Data, with Applications to a Breast Cancer Prevention Trial

**Kiros Berhane**

Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street
CHP-220, Los Angeles, California, U.S.A.
*email:* kiros@usc.edu

**and**

**Lisa A. Weissfeld**

Department of Biostatistics, University of Pittsburgh, 303 Parran Hall, Pittsburgh, Pennsylvania, U.S.A.
*email:* lweis@imap.pitt.edu

Summary. As part of the National Surgical Adjuvant Breast and Bowel Project, a controlled clinical trial known as the Breast Cancer Prevention Trial (BCPT) was conducted to assess the effectiveness of tamoxifen as a preventive agent for breast cancer. In addition to the incidence of breast cancer, data were collected on several other, possibly adverse, outcomes, such as invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis and/or pulmonary embolism. In this article, we present results from an illustrative analysis of the BCPT data, based on a new modeling technique, to assess the effectiveness of the drug tamoxifen as a preventive agent for breast cancer. We extended the flexible model of Gray (1994, Spline-based test in survival analysis, *Biometrics* **50,** 640–652) to allow inference on multiple time-to-event outcomes in the style of the marginal modeling setup of Wei, Lin, and Weissfeld (1989, Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *Journal of the American Statistical Association* **84,** 1065–1073). This proposed model makes inference possible for multiple time-to-event data while allowing for greater flexibility in modeling the effects of prognostic factors with nonlinear exposure-response relationships. Results from simulation studies on the small-sample properties of the asymptotic tests will also be presented.

Key words: Additive models; Proportional hazards; Ridge regression; Smoothing; Splines; Survival analysis.

## 1. Introduction

The advent of promising chemoprevention agents for the prevention of breast and other cancers has brought both hope and controversy to the scientific world and the general public. Central to the assessment of the usefulness of chemoprevention agents are careful study of the costs, potential benefits, and possible harmful side effects of any drug used for the purpose of disease prevention. Thus, clinical trials must be carefully designed to collect information on all potential outcomes of interest and the analysis must account for both the beneficial and potentially harmful effects of any chemoprevention agent that is used to prevent a disease. Unlike treatment trials, prevention trials are, by nature, designed to monitor multiple outcomes. The outcomes are multivariate in nature and are not subjected to competing risks, since the development of a cardiovascular outcome does not preclude the development of a cancer at a later point in time. In fact, both outcomes are of interest in a prevention study, since the goal is to determine the overall impact of the chemopreventive agent. This leads to the assumption of an independent censoring mechanism for each of the outcomes, making it different from the competing risks problem.

An important example of a chemoprevention trial is the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention Trial, hereafter referred to as the BCPT (Fisher et al., 1996). The goal of this randomized controlled clinical trial was to assess the effectiveness of tamoxifen as a preventive agent for breast cancer. There were several outcomes of interest in this trial, namely, the development of breast cancer, invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep vein thrombosis, and/or pulmonary embolism. Subjects were followed for a minimum of 5 years or until death. Several of these outcomes are of particular interest (invasive endometrial cancer, ischemic heart disease, deep-vein thrombosis and pulmonary embolism), since they are negative outcomes associated with the use of tamoxifen. Thus, in order to assess the overall effectiveness of tamoxifen, these outcomes must be treated in a simultaneous and comprehensive manner. As an example of this, it would be difficult to argue that tamoxifen is of benefit if the drug has little or no effect on the development of breast cancer and a large number of subjects developed a deep-vein thrombosis, pulmonary embolism, or an invasive endometrial cancer. For this reason, it is important to consider these outcomes simultaneously in an analysis. The analytic method should also allow for time-dependent treatment effects, because treatment is likely to be stopped after onset of one outcome, even though subjects are usually followed to monitor for other possible outcomes up to the time of death or termination of the trial.

There are several methods available for the analysis of multivariate survival data, such as that collected in the area of prevention, with the Wei, Lin, and Weissfeld (1989) approach being one of the more general methods. Using this approach, each outcome is modeled separately using a Cox proportional hazards model (Cox, 1972). The variance-covariance matrix of the resulting parameter estimates is then obtained via a sandwich estimator. While this method is quite useful, it may fail to appropriately model exposure-response relationships that may have nonlinear forms. Given the fact that it has already been demonstrated that important prognostic factors (e.g., BMI) have a markedly nonlinear effect on breast cancer survival and/or prognosis (Gray, 1994), there is a need for flexible models that could model nonlinear effects of prognostic factors, but also allow for simultaneous inference on several time-to-event outcomes. Most of the research on flexible models for time-to-event data has concentrated on single time-to-event outcomes (e.g., O'Sullivan, 1988; Hastie and Tibshirani, 1990a; Gray, 1994).

In this article, we propose a new method for inference on multiple time-to-event outcomes by extending the Wei et al. (1989) approach to allowing flexibility in modeling each of the outcomes. This method allows for flexibility through a spline on the covariate space in the style of Gray (1992, 1994). In Section 2, we give background details on the Cox (1972) and Gray (1994) models. In Section 3, we discuss the proposed flexible model for multiple outcomes and we derive the variance covariance matrix and inference for the extension to Wei et al. (1989) based on Gray's model. In Section 4, we present results from an extensive simulation study on the empirical size of the proposed tests in small sample settings. In Section 5, we present results from a detailed analysis of the BCPT data. In Section 6, we discuss various areas for further extensions. Additional technical details on calculations in estimating the variance estimator for inference on multiple outcomes are given in the Appendix.

## 2. Background

### 2.1 *The Cox Model*

Consider a study in which multiple, say, $G$ different, time-to-event outcomes are under consideration. For any one outcome, say the $g$th one, Cox (1972) proposed a proportional hazards model of the form

$$\lambda_{gi}(t) = \lambda_{g0}(t)\, exp\left\{ \sum_j \beta_{jg} Z_{jgi} \right\}, \quad t \geq 0, \tag{1}$$

where $\lambda_{g0}(t)$ is an unspecified baseline hazard function and $\beta_{jg}$, $j = 1, \ldots, p$, denotes the regression parameter associated with the $j$th risk or prognostic factor. Here, one observes data for each of the outcomes that is of the form $(X_{gi}, Z_{gi}, \Delta_{gi})$, where $X_{gi} = min(\tilde{X}_{gi}, C_{gi}), C_{gi}$ is the censoring time, $Z_{gi}(t) = (Z_{1gi}(t), \ldots, Z_{pgi}(t))^T$ and $\Delta_{gi} = 1$ if $X_{gi} = \tilde{X}_{gi}$ and 0 otherwise. Note that under these assumptions each outcome is independently censored by its own censoring time $C_{gi}$. For this fully linear model, the partial likelihood is given as

$$PL_g(\beta) = \prod_{i=1}^{n} \left( \frac{exp\{\boldsymbol{\beta}_{g(T)} Z_{gi}(X_{gi})\}}{\sum_{l \in \mathcal{R}_g(X_{gi})} exp\{\boldsymbol{\beta}_{g(T)} Z_{gl}(X_{gl})\}} \right)^{\Delta_{gi}}, \tag{2}$$

where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{1g}, \ldots, \boldsymbol{\beta}_{pg})^T$ and $\mathcal{R}_g(t) = \{l : X_{gl} \geq t\}$ denotes the set of subjects at risk just prior to time $t$ with respect to the $g$th type of failure. The solution to the score equation $U_g(\boldsymbol{\beta}_g) = \partial log PL_g(\boldsymbol{\beta}_g)/\partial \boldsymbol{\beta}_g = 0, \hat{\boldsymbol{\beta}}_g$, can be shown to be a consistent estimator of $\boldsymbol{\beta}_g$, provided that the model is correctly specified (Anderson and Gill, 1982). Specifically, letting $\boldsymbol{\beta}_{g(T)}$ be the vector of true parameter values for the $g$th outcome, inference is based on the asymptotic normality of the score vector $U_g(\boldsymbol{\beta}_{g(T)})$. Based on this result, $n^{(1/2)}(\hat{\boldsymbol{\beta}}_g - \boldsymbol{\beta}_{g(T)})$ is asymptotically normal with mean $\mathbf{0}$ and variance given as the limit of $nA_g^{-1}$ where

$$A_g\left(\boldsymbol{\beta}_{g(T)}\right) = \frac{-\partial^2 log PL_g\left(\boldsymbol{\beta}_{g(T)}\right)}{\partial \boldsymbol{\beta}_{g(T)} \boldsymbol{\beta}_{g(T)}^T}. \tag{3}$$

For more details on the Cox proportional hazards model, see Cox and Oakes (1984).

### 2.2 *Gray's Model*

Gray (1994) proposed a penalized B-spline based model by replacing the linear model form, $\sum_j \beta_{jg} Z_{jgi}$, with the flexible form, $\sum_j f_{jg}(Z_{jg})$, in the proportional hazards model given by (1). In practical applications, the effects of most covariates are known to have some parametric form, while some of them are best modeled via non-parametric smoothers. So, for simplicity of discussion and without loss of generality, we discuss most details for a model with $p$ covariates with parametric forms and one additional covariate with non-parametric function, say $h_g$. We also suppress the dependence of the covariates on $X_{gi}$. We first let

$$\lambda_{gi}(t) = \lambda_{g0}(t)\, exp\left\{ \sum_j \beta_{jg} Z_{jgi} + f_g(h_{gi}) \right\}, \quad t \geq 0, \tag{4}$$

where $j = 1, \ldots, p$. The penalized regression spline approach is used to estimate $f_g(h_{gi})$, i.e.,

$$f_g(h_g) = \gamma_{1g} h_g + \sum_{q=2}^{m+3} \gamma_{qg} B_{qg}(h_g). \tag{5}$$

Note that the constant term has been dropped, since it is accounted for by the baseline hazard, and only $(m + 2)$ of the B-spline basis functions are used for identifiability (DeBoor, 1974). See the Appendix for more details. Following Gray (1994), let $\boldsymbol{\gamma}_g = (\gamma_{g2}, \ldots, \gamma_{g(m+3)})$ and $\boldsymbol{\eta}_g = (\gamma_{1g}, \boldsymbol{\gamma}_g)$. Then, a penalized partial likelihood that includes a penalty function to allow for smoother alternatives would be defined as

$$PL_g^p(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) - 1/2\lambda \int [f_g''(u)]^2 du, \qquad (6)$$

where $\lambda$ controls the amount of smoothing. Note that setting $\lambda = 0$ and $\lambda \rightarrow \infty$ lead to a no-penalty regression spline function and a linear term, respectively. Recognizing that the penalty function given above is quadratic in the parameter vector $\boldsymbol{\eta} = (\gamma_1, \ldots, \gamma_{m+3})$, one could rewrite (6) as

$$PL_g^p(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) - 1/2\lambda_g \boldsymbol{\eta}_g^T \mathbf{K}_g \boldsymbol{\eta}_g, \qquad (7)$$

where $\mathbf{K}$ is a nonnegative definite matrix that is a function of the covariate $h_g$. Note that $\mathbf{K}$ is an $(m+3) \times (m+3)$ matrix with the first row and column being zeros, since the linear function passes unpenalized. Note also that the quadratic form in the penalty matrix $\mathbf{K}$ is due to the accumulation of squared second differences. For more details on the actual steps involved in calculating the penalty matrix $K$, see Green and Silverman (1994, Section 2.1–Section 2.5). The hypotheses of interest with respect to the smooth function are then $\boldsymbol{\eta}_g = \mathbf{0}$ and $\boldsymbol{\gamma}_g = \mathbf{0}$, representing the hypotheses of "no effect" and "linear effect," respectively.

For model (4), the unpenalized part of equation (7) can be written as

$$PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = \prod_{i=1}^{n} \times \left( \frac{exp\left\{ \sum_{j=1}^{p} Z_{gj}\beta_{gj} + h_g\gamma_{g1} + \sum_{q=2}^{m+3} B_{qg}(h_g)\gamma_{qg} \right\}}{\sum_{s \in \mathcal{R}_g(X_{gi})} exp\left\{ \sum_{j=1}^{p} Z_{gj}\beta_{gj} + h_g\gamma_{g1} + \sum_{q=2}^{m+3} B_{qg}(h_g)\gamma_{qg} \right\}} \right)^{\Delta_{gi}}, \qquad (8)$$

where all components are as defined in Section 2.1, for the $g^{th}$ type of failure. Let $\boldsymbol{\psi}_g = (\boldsymbol{\beta}_g, \ \boldsymbol{\eta}_g)$ and $P_g = (Z_{1g} : \ldots : Z_{pg} : h_g : B_{2g}(h_g) : \ldots : B_{m+3,g}(h_g))$, with $P_{g(r)}$ denoting the $r^{th}$ column vector, $r = 1, \ldots, (m+p+3)$. Letting $\hat{A}_g$ be the unpenalized information matrix, as in (3) for the $g^{th}$ outcome, as a function of $\boldsymbol{\psi}$, it can be shown that

$$\sqrt{n}\big(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_{g(T)}\big) = n(A_g + \lambda_n \tilde{K})^{-1} n^{-1/2} U_g\big(\boldsymbol{\psi}_{g(T)}\big) + o_p(1)$$

where $U_g(\boldsymbol{\psi}_{g(T)})$ is the unpenalized score vector, $\boldsymbol{\psi}_{g(T)}$ is the vector of true parameter values for the $g$th outcome (Gray, 1994), and $\tilde{K}$ is the expanded penalty matrix that augments rows and columns of zeros to $\mathbf{K}$, to account for the unpenalized terms in the model. Then, it follows from the asymptotic normality of $U_g(\boldsymbol{\psi}_{g(T)})$ that $\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_{g(T)})$ is asymptotically normal with mean $\mathbf{0}$ and variance given as the limit of $nV_g$ where

$$V_g = (A_g + \lambda_n \tilde{K})^{-1} A_g (A_g + \lambda_n \tilde{K})^{-1}. \qquad (9)$$

Note that the above asymptotic results assume that the number of terms in the spline function is held fixed as $n \rightarrow \infty$ (Gray, 1994). Gray's model also uses $(A_g + \lambda_n \tilde{K})^{-1}$ in all tests. The reference distribution for the test statistics under $H_0$ is given by a weighted sum of $\chi_1^2$'s, where the weights are given by eigenvalues of the matrix $\lim A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}}(A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} + \lambda\tilde{K})^{-1}$, for the g$^{th}$ outcome, with $A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} = A_{\boldsymbol{\eta\eta}} - A_{\boldsymbol{\eta\psi}} A^{-1}_{\boldsymbol{\psi\psi}} A_{\boldsymbol{\psi\eta}}$. In contrast, test statistics that are based directly on (9) have a $\chi^2_{df}$ reference distribution, with the number of degrees of freedom equal to the rank of the covariate vector under the null hypothesis (Wang and Taylor, 1995). Note that tests that are based on these two approaches may result in different orderings of outcomes in the sample space, because they are based on different quadratic forms (as pointed out by one of the referees). For theoretical developments along the lines of Wei et al. (1989), the test form that is based on (9) is more suitable.

## 3. A Flexible Model for Multiple Outcomes

While making inference on each of the margins is often of interest, this could be done easily by using developments found in Gray (1994). The focus of our interest here is in being able to conduct simultaneous inference on several time-to-event outcomes in models that have nonparametric smooth terms. Once the marginal distributions are modeled, then the methods described in Wei et al. (1989) can be extended to test for trends across parameter estimates and to combine estimates across margins to test for covariate effects of interest.

To develop the simultaneous inferential procedures for several outcomes, we first note that the $\boldsymbol{\psi}_g$'s across the $G$ multiple outcomes (defined in Section 2.2) are generally correlated. Then, analogous to developments in Wei et al. (1989), the asymptotic covariance matrix between $\sqrt{n}(\hat{\boldsymbol{\psi}}_g - \boldsymbol{\psi}_g)$ and $n^{(1/2)}(\hat{\boldsymbol{\psi}}_v - \boldsymbol{\psi}_v)$ can be consistently estimated by

$$\hat{D}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) = \hat{V}_g(\hat{\boldsymbol{\psi}}_g)\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v)\hat{V}_v(\hat{\boldsymbol{\psi}}_v), \qquad (10)$$

where $\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v) = n^{-1}\sum_{i=1}^{n} W_{gi}(\hat{\boldsymbol{\psi}}_g) W_{vi}(\hat{\boldsymbol{\psi}}_v)^T, \hat{V}_g(\hat{\boldsymbol{\psi}}_g)$ is an evaluation of equation (9) at $\hat{\boldsymbol{\psi}}_g$. Based on the results from the Appendix, the covariance matrix of $(\hat{\boldsymbol{\psi}}_1, \ldots, \hat{\boldsymbol{\psi}}_G)$ can be consistently estimated by

$$\hat{Q} = n^{-1} \begin{pmatrix} \hat{D}_{11}(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_1) & \ldots & \hat{D}_{1G}(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_G) \\ \vdots & \ddots & \vdots \\ \hat{D}_{G1}(\hat{\boldsymbol{\psi}}_G, \hat{\boldsymbol{\psi}}_1) & \ldots & \hat{D}_{GG}(\hat{\boldsymbol{\psi}}_G, \hat{\boldsymbol{\psi}}_G) \end{pmatrix}. \qquad (11)$$

Note that $W_{gi}$ and $W_{vi}$ in $\hat{C}_{gv}(\hat{\boldsymbol{\psi}}_g, \hat{\boldsymbol{\psi}}_v)$ are defined in terms of the unpenalized score contributions, because the penalty contributions are asymptotically negligible under the null hypothesis, as discussed in the Appendix. The penalty terms, however, could prove to be important in extensions of any existing finite sample correction methods for the Cox regression (e.g., Fay and Graubard, 2001). Such finite sample correction extensions are beyond the scope of this article.

### 3.1 *Testing Statistical Hypotheses*

For the nonparametric term, one could conduct simultaneous inference on the "overall" effect and/or "linearity" of $h$ across failure types. Let $\hat{\boldsymbol{\eta}}_g$ denote the components of $\hat{\boldsymbol{\psi}}_g$ that correspond to the relevant components of the nonparametric term $h_g$ and $\hat{\Gamma}$ denote the relevant submatrix of $\hat{Q}$ corresponding to $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_G)$. Then, one could use the quadratic form

$$(\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_G)\hat{\Gamma}^{-1}(\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_G)^T \sim \chi^2_{\nu}, \qquad (12)$$

under $H_0$, (where $\nu$ is the number of terms in $\boldsymbol{\eta}$) to conduct a joint test on the null hypotheses given by $H_0 : \boldsymbol{\eta}_g = \mathbf{0}$, $g = 1, \ldots, G$. Note that the tests for "overall" significance or "linearity" are done in the above setup by choosing the last $(m + 3)$ and $(m + 2)$ elements of $\boldsymbol{\psi}_g$, respectively. Note that (12) is based on a direct application of (9). A different testing procedure, as discussed in Section 2.2 and described in Wang and Taylor (1995), could also be given by using $(A_g + \lambda_n \tilde{K}_g)^{-1}$ and $(A_v + \lambda_n \tilde{K}_v)^{-1}$ in (11) instead of $V_g$ and $V_v$, respectively. Under the null hypothesis, this modified Wald test statistic would then have an asymptotic distribution of the form

$$\sum_{g=1}^{G} \sum_{j} \lambda_{gj} \phi_j^2$$

where the $\phi_j$'s are independent standard normal random variables, and the $\lambda_{gj}$'s are the eigenvalues of the matrix $\lim A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}}(A_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} + \lambda \tilde{K})^{-1}$, for the $g$th outcome. The arguments that lead to this form are given in Gray (1994) for a single outcome. The extensions to multiple margins are straightforward. Note that the use of penalized B-splines, as opposed to fully nonparametric smoothers such as smoothing splines, makes the computation of the $\lambda_{gj}$'s possible.

A linear contrast could be constructed to test hypotheses with respect to a group of parameters (e.g., all parameters to a spline term on each margin) across outcomes. For example, one could test the hypothesis that $\boldsymbol{\eta}_1 = \cdots = \boldsymbol{\eta}_G = \boldsymbol{\eta}$ and then estimate the common $\boldsymbol{\eta}$ by constructing a linear combination of the $\boldsymbol{\eta}_g$'s in a way that takes the appropriate variance-covariance matrix into account. For linear terms, it may also be of interest to obtain a common across-outcomes estimate of the regression parameter, say, $\eta_g$, via $\sum_{g=1}^{G} c_g \hat{\eta}_g$ with $\sum_{g=1}^{G} c_g = 1$, where weights $c_g$'s that have the smallest asymptotic variance among all of the linear estimators (Wei et al., 1989) are chosen as $\mathbf{c} = (c_1, \ldots, c_G)^T = (\mathbf{e}^T \hat{\Gamma}^{-1} \mathbf{e})^{-1} \hat{\Gamma}^{-1} \mathbf{e}$ and $\mathbf{e} = (1, \ldots, 1)^T$. However, spline terms usually involve multiple parameters and the multicollinearity among them should be taken into account when taking the linear combinations via the off-diagonal covariance terms. Trends in regression effects across margins could also be examined via sequential multiple testing procedures, as in Wei and Stram (1988).

A suite of Splus functions, along with supporting FORTRAN programs for conducting simultaneous inference on several outcomes, will be available at `http://hydra.usc.edu/berhane`. These programs use previous developments by Robert Gray that have been kindly disseminated to the research community via the STATLIB archive. We also plan to put the complete set of software on such popular online statistical libraries as STATLIB.

### 3.2 *Choice of Smoothing Parameters, Degrees of Freedom, and Placement of Knots*

In the above setup, we assume that the amount of smoothing (i.e., the value of the smoothing parameter) is fixed by the analyst via prior knowledge or through a grid search. It is also possible to develop automatic procedures for selecting the smoothing parameters by using criteria such as cross validation. While this could lead to optimal estimation of the functional forms, its implications for hypothesis testing are not obvious. Operationally, one specifies the degrees of freedom for each nonparametric term and the corresponding value of the smoothing parameter is then calculated. As a general operating guide, we use a relatively small number of degrees of freedom (Gray, 1994). The number of the knots that determine the B-spline basis functions are generally set to be at least twice the number of the degrees of freedom, so as to avoid wild fluctuation in the smooth function estimates, usually set between 10–15 per outcome. We will discuss the potential effects of various choices of number of knots in our simulation studies. In this paper, we follow Gray (1994) in putting the knots at locations that yield approximately equal numbers of failure observations between knots. The calculation of degrees of freedom is analogous to that given in Gray (1994) and Wei et al. (1989). For example, to test whether all parameters in a spline model are equivalent across $G$ outcomes, the degrees of freedom are computed as $\sum_{g=1}^{G} df_g$, where

$$df_g = \operatorname{trace}\left\{ \lim A^{(g)}_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} \left( A^{(g)}_{\boldsymbol{\eta\eta}|\boldsymbol{\psi}} + \lambda_g \tilde{K}_g \right)^{-1} \right\}.$$

## 4. Simulation Study

Extensive simulation studies were conducted to examine the performance of the proposed procedures for conducting simultaneous inference on several time-to-event outcomes. We focused on the bivariate case, where two time-to-event outcomes are considered under various levels of dependence. To generate data, the family of bivariate exponential distributions of Gumbel (1960) was used. Consider two marginal distributions, say, $F_1$ and $F_2$, from the univariate exponential, with hazard rates given by $\exp(\beta_1 Z)$ and $\exp(\beta_2 Z)$, respectively. Then, the distribution function of the bivariate exponential distribution is of the form

$$F(x_1, x_2) = F_1(x_1) F_2(x_2)[1 + \theta\{1 - F_1(x_1)\}\{1 - F_2(x_2)\}].$$

The quantity $\theta/4$ measures the correlation between the two event times, where $-1 \le \theta \le 1$. In the above models, $Z$ denotes any vector of covariates that may include binary indicators, or covariate effects that assume various functional forms.

In the simulations that test for overall significance, we set the covariate values in the two margins to be equal. Specifically, the null hypothesis is $\boldsymbol{\eta}_g = \mathbf{0}$, as defined in Section 2.2, and the test statistic is based on the Wald test, as described in (12). Censoring indicators were generated independently using uniform distributions gauged to depict various percentages of censoring (30%, 50%). Empirical sizes of the spline based tests, based on 2000 runs, were examined under various specifications of sample sizes ($n$ = 200, 300, 400), degrees of freedom ($df$ = 3, 5), number of knots (10, 15, and 20) and levels of dependence between the margins ($\theta$ = 0.5, 1.0). Note that the degree of correlation between the two outcomes is given by $\theta/4$, and $\theta = 1$ is the maximum correlation allowed by the bivariate model of Gumbel (1960).

Table 1 gives results from the simulation with low levels of dependence ($\theta = 0.5$) between the outcomes. The results indicate that the empirical size is reasonably close to the corresponding nominal values only when the sample size is at least 200 per margin. This relatively poorer performance is probably due to the fact that we are dealing with spline-based models when the outcomes are correlated. Based on these simulation results and similar observations in Gray (1994), it would be advisable to use a smoother that has relatively small

**Table 1**
*Empirical sizes of robust inference on marginally correlated ($\theta = 0.5$) bivariate time-to-event outcomes*

| | | | $n = 200$<br>Nominal level | | | $n = 300$<br>Nominal level | | |
|---|---|---|---|---|---|---|---|---|
| Censoring prob. | Deg. of freedom | Number of knots | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| 0.3 | 3 | 10 | 0.012 | 0.038 | 0.069 | 0.018 | 0.055 | 0.092 |
| | | 15 | 0.022 | 0.070 | 0.121 | 0.029 | 0.079 | 0.130 |
| | | 20 | 0.047 | 0.112 | 0.167 | 0.035 | 0.084 | 0.134 |
| | 5 | 10 | 0.030 | 0.068 | 0.114 | 0.022 | 0.071 | 0.121 |
| | | 15 | 0.052 | 0.129 | 0.184 | 0.027 | 0.089 | 0.146 |
| | | 20 | 0.103 | 0.200 | 0.270 | 0.051 | 0.137 | 0.206 |
| 0.5 | 3 | 10 | 0.013 | 0.051 | 0.096 | 0.013 | 0.051 | 0.089 |
| | | 15 | 0.032 | 0.098 | 0.151 | 0.023 | 0.073 | 0.130 |
| | | 20 | 0.074 | 0.163 | 0.238 | 0.041 | 0.120 | 0.185 |
| | 5 | 10 | 0.016 | 0.042 | 0.081 | 0.008 | 0.031 | 0.061 |
| | | 15 | 0.029 | 0.080 | 0.124 | 0.015 | 0.046 | 0.083 |
| | | 20 | 0.068 | 0.152 | 0.216 | 0.035 | 0.078 | 0.123 |

number of degrees of freedom, with the number of knots not exceeding 15 for most practical applications.

Table 2 gives results from simulation with high levels of dependence ($\theta = 1.0$) between the outcomes. Here, due to the added level of dependence between the margins, the empirical sizes for $n = 200$ were still unacceptably high (results not shown). But, the empirical sizes for $n = 300$, 400 gave more reasonable results. Once again, the use of a large sample size is advised for most practical applications. The results from both Tables 1 and 2 indicate that the number of knots should be kept between 10 and 15. Specifically, the results for 10 knots and 15 knots provided empirical sizes that are reasonably close to the nominal sizes for models that use 3 and 5 degrees of freedom, respectively. The simulation results also indicate that the models performed better when the correlation between outcomes was marginal (i.e., $\theta = 0.5$).

## 5. Analysis of the BCPT Data

The Breast Cancer Prevention Trial, hereafter referred to as BCPT, (Fisher et al., 1998) was initiated in 1992 enrolling 13,388 women that were at increased risk for breast cancer due to their relatively old age (≥60 years of age), relatively high 5-year predicted risk for breast cancer (a risk of at least 1.66% for those 35–59 years of age) and/or history of lobular carcinoma *in situ*. Subjects were then randomly classified into placebo and treatment groups (6707 subjects into a placebo group and 6681 subjects receiving 20 mg/day of tamoxifen

**Table 2**
*Empirical sizes of robust inference on marginally correlated ($\theta = 1.0$) bivariate time-to-event outcomes*

| | | | $n = 300$<br>Nominal level | | | $n = 400$<br>Nominal level | | |
|---|---|---|---|---|---|---|---|---|
| Censoring prob. | Deg. of freedom | Number of knots | 0.01 | 0.05 | 0.10 | 0.01 | 0.05 | 0.10 |
| 0.3 | 3 | 10 | 0.015 | 0.062 | 0.122 | 0.009 | 0.037 | 0.076 |
| | | 15 | 0.033 | 0.092 | 0.156 | 0.012 | 0.051 | 0.086 |
| | | 20 | 0.056 | 0.140 | 0.210 | 0.016 | 0.061 | 0.096 |
| | 5 | 10 | 0.028 | 0.085 | 0.144 | 0.012 | 0.045 | 0.081 |
| | | 15 | 0.048 | 0.119 | 0.174 | 0.016 | 0.066 | 0.112 |
| | | 20 | 0.078 | 0.166 | 0.237 | 0.024 | 0.073 | 0.131 |
| 0.5 | 3 | 10 | 0.022 | 0.085 | 0.172 | 0.004 | 0.025 | 0.051 |
| | | 15 | 0.044 | 0.125 | 0.206 | 0.007 | 0.030 | 0.057 |
| | | 20 | 0.066 | 0.171 | 0.263 | 0.010 | 0.049 | 0.086 |
| | 5 | 10 | 0.013 | 0.052 | 0.095 | 0.024 | 0.077 | 0.119 |
| | | 15 | 0.023 | 0.078 | 0.136 | 0.029 | 0.086 | 0.156 |
| | | 20 | 0.040 | 0.123 | 0.198 | 0.040 | 0.096 | 0.170 |

**Table 3**
*Marginal proportional hazards models on breast cancer, ischemic heart disease, and endometrial cancer*

| Outcome | Covariate | Estimate | Test statistic | df | P-value |
|---|---|---|---|---|---|
| Invasive breast cancer | TRT | −0.69 | 28.08 | 1.00 | <0.01 |
| | LCIS | 0.19 | 0.40 | 1.00 | 0.53 |
| | AGE (overall) | | 2.89 | 4.00 | 0.61 |
| | AGE (linearity) | | 2.78 | 3.00 | 0.44 |
| | PR5YR (overall) | | 17.26 | 4.00 | <0.01 |
| | PR5YR (linearity) | | 6.94 | 3.00 | 0.05 |
| Ischemic heart disease | TRT | 0.13 | 0.54 | 1.00 | 0.47 |
| | LCIS | −0.95 | 2.00 | 1.00 | 0.16 |
| | AGE (overall) | | 73.3 | 3.99 | <0.01 |
| | AGE (linearity) | | 3.54 | 3.00 | 0.30 |
| | PR5YR (overall) | | 5.33 | 4.00 | 0.24 |
| | PR5YR (linearity) | | 2.96 | 3.00 | 0.40 |
| Endometrial cancer | TRT | 0.88 | 8.23 | 1.00 | <0.01 |
| | LCIS | 0.60 | 0.32 | 1.00 | 0.57 |
| | AGE (overall) | | 4.32 | 3.99 | 0.36 |
| | AGE (linearity) | | 3.84 | 3.00 | 0.26 |
| | PR5YR (overall) | | 5.19 | 4.00 | 0.25 |
| | PR5YR (linearity) | | 2.50 | 3.00 | 0.50 |

for up to 5 years). The main aim was to examine the effectiveness of tamoxifen in preventing the possible occurrence of invasive breast cancer in high-risk women. Data were also collected on other outcomes (some of them unwanted adverse side effects), such as invasive endometrial cancer, ischemic heart disease, transient ischemic attack, deep-vein thrombosis, and pulmonary embolism. The treatment regimen was terminated when any one of the outcomes was observed, but subjects were followed up to the end of the trial, to collect information on the other outcomes.

Analysis of data from the BCPT has shown (Fisher et al., 1998) that there was a 49% reduction in the risk of invasive breast cancer in those high-risk women that received tamoxifen treatment (for up to 5 years), compared to those that received placebo. However, the benefits of tamoxifen were tempered by adverse side effects that significantly increased the risk of endometrial cancer, deep-vein thrombosis, pulmonary embolism, and some other cardiac effects. In fact, the issue of whether the benefits of tamoxifen outweighs the potential risk was controversial enough that the National Cancer Institute (NCI) sponsored a workshop on the subject in July, 1998, leading to a risk-benefit analysis as reported in Gail et al. (1999).

The results indicate that age and baseline-predicted risks for breast cancer play a significant role in determining whether the benefits of tamoxifen outweigh the associated risks. In this paper, we use the newly developed techniques to simultaneously analyze several outcomes in a way that allows for risks that may not be constant across such factors as age. We focus on invasive breast cancer (IBC), ischemic heart disease (IHD), and endometrial cancer (ENDO) as our outcomes of interest. The primary covariates of interest were treatment (TRT, placebo vs. tamoxifen), age at time of entry (AGE, in years), 5-year breast cancer risk at time of entry (based on a multivariate logistic model of Gail et al. 1989) (PR5YR), lobular carcinoma in situ (LCIS), and atypical hyperplasia of the breast (ATYPH, history at entry). The two continuous covariates that could be modeled using the spline approach, in order to examine non-linearity in their effects, were age and the 5-year breast cancer probability from the Gail model.

The results from the marginal models on each of the three outcomes are given in Table 3 and the corresponding smooth function estimates for AGE and PR5YR are given in Figure 1(a)–1(f). Note that the panels of Figure 1 depict penalized B-spline based estimates of the functions on AGE and PR5YR, along with 95 % pointwise confidence bands. The results from the marginal models indicate that use of tamoxifen is associated with reduced risk of invasive breast cancer ($p < 0.01$), but that it was also associated with significantly increased risk of endometrial cancer. The increased risk in ischemic heart disease appeared to be marginal and not statistically significant. Age of the subjects appeared to be positively associated only with ischemic heart disease, but this association appeared to be linear (Figure 1[c]). On the other hand, the 5-year probability of breast cancer (as estimated form Gail model) was nonlinearly associated with onset of invasive breast cancer (Figure 1[b]). Here, the estimated curve (Figure 1[b]) indicates an initial rise in risk up to 6–7, with a decline in risk starting at about 10. The test for nonlinearity was marginally significant, indicating that a simple linear term may not suffice to control for this variable.

The question still remains as to whether there was evidence of an overall beneficial or detrimental effect of tamoxifen and other prognostic factors when inferences are drawn simultaneously on more than one outcome. This is the question that the new modeling techniques are best suited to answer. Here, we considered bivariate models that simultaneously model invasive breast cancer with ischemic heart disease and endometrial cancer; the results from these two bivariate models are given in Table 4. These results indicate that the

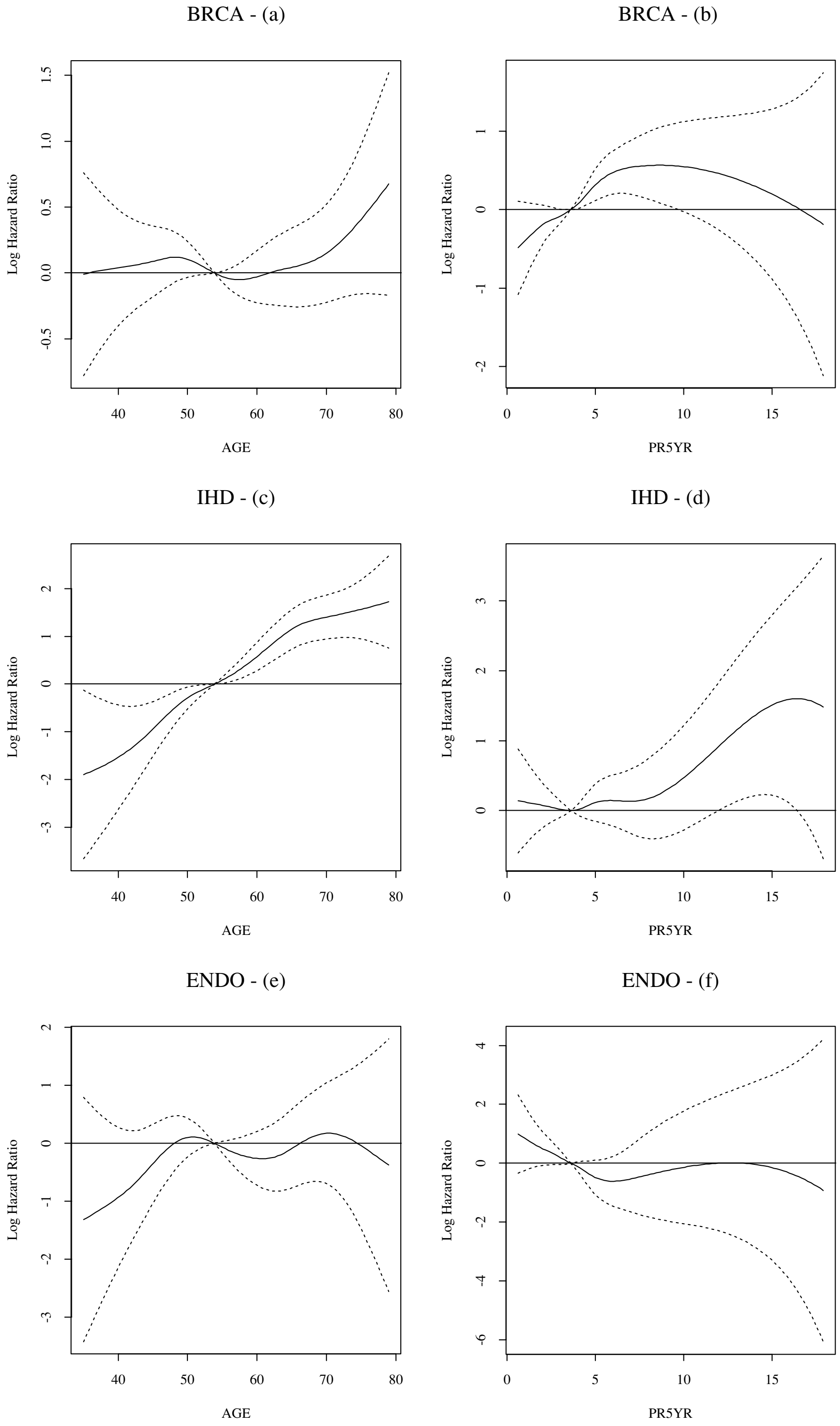

**Figure 1.** Spline-based estimates of the log hazard ratio for breast cancer as functions of age and five-year probability of breast cancer for models on invasive breast cancer (BRCA, panels [**a**] and [**b**]), ischemic heart disease (IHD, panels [**c**] and [**d**]), and endometrial cancer (ENDO, panels [**e**] and [**f**]).

benefits of tamoxifen as a preventive agent significantly outweigh the detrimental effect of an increased risk in ischemic heart disease. The appropriately weighted combined estimate of treatment effect for IBC and IHD was $-0.41$ ($p < 0.01$). It is also interesting that even though there was a significant increased risk in endometrial cancer that was associated with the use of tamoxifen, it did not appear to wash out its benefit of reducing the risk of breast cancer. In fact, there was still

**Table 4**

*Bivariate proportional hazards models on breast cancer, ischemic heart disease, and endometrial cancer*

| Outcome | Covariate | Combined estimate | Test statistic | df | P-value |
|---|---|---|---|---|---|
| IBC | TRT | −0.41 | 28.85 | 2.00 | <0.01 |
| and | LCIS | −0.56 | 2.24 | 1.97 | 0.32 |
| IHD | AGE (overall) | | 419.84 | 8.00 | <0.01 |
| | AGE (linearity) | | 5.62 | 6.00 | 0.48 |
| | PR5YR (overall) | | 24.78 | 8.00 | <0.01 |
| | PR5YR (linearity) | | 10.92 | 6.00 | 0.07 |
| IBC | TRT | −0.55 | 36.54 | 2.00 | <0.01 |
| and | LCIS | −0.58 | 0.91 | 2.00 | 0.62 |
| ENDO | AGE (overall) | | 7.96 | 8.00 | 0.44 |
| | AGE (linearity) | | 7.30 | 6.00 | 0.27 |
| | PR5YR (overall) | | 27.27 | 8.00 | <0.01 |
| | PR5YR (linearity) | | 13.76 | 6.00 | 0.02 |

a statistically significant protective effect of tamoxifen when the risks for breast cancer and endometrial cancer were considered simultaneously. The inverse-variance weighted combined estimate of treatment effect for IBC and ENDO was −0.55 ($p < 0.01$). The results indicate a strong linear effect of age in the bivariate model for invasive breast cancer and ischemic heart disease. On the other hand, PR5YR appears to have a strong nonlinear effect in both bivariate models, indicating that it should be modeled as a nonlinear term. Note that the common estimates on bivariate outcomes for each of the prognostic factors are obtained by using a linear combination of the $\boldsymbol{\eta}_g$'s in a way that takes the appropriate variance-covariance matrix into account. So, they allow for a truly combined inference across outcomes, as opposed to the relatively ad-hoc methods of visually comparing the marginal estimates.

## 6. Discussion

The analyses in this article demonstrate that when examining the effectiveness of chemopreventive agents on diseases such as breast cancer, appropriate modeling techniques are needed to (i) to allow for simultaneous examination of the beneficial and potentially adverse effects of the agent, and (ii) enable the proper modeling of prognostic and/or risk factors that may have nonlinear exposure response relationship.

The methods proposed here have the advantage of being able to estimate a relatively realistic functional form for the covariate effects of interest, while enabling formal inference on the overall significance or adequacy of a certain parametric form (e.g., linearity) across several time-to-event outcomes. This is made possible by using penalized B-splines that are known to offer an attractive compromise between fully nonparametric regression smoothers, such as smoothing splines, and flexible, but inherently parametric, techniques such as regression splines (Hastie and Tibshirani, 1990b; Gray, 1994).

In this article, we have introduced a way of conducting simultaneous inference across several outcomes by extending the methods of Gray (1994) and Wei et al. (1989). The results from the analysis of the BCPT data demonstrate its immediate usefulness in health-related research. The simulations demonstrate that the asymptotic inferential procedures are reliable when adequately large sample sizes are used, and also provide rough guidelines on how to select realistic values for the degrees of freedom (hence, smoothing parameters), and number and location of knots. The small sample properties of the proposed tests may be improved by extending a covariance estimator as in, say, Fay and Graubard (2001). Note that parametric regression splines are much simpler to apply and still play an important role in practical applications, especially when the number of knots are appropriate and the positions of such knots reasonably placed.

There are many open areas of research that would extend the methods in this article. Some of the most important areas of research include development of diagnostic measures in the multivariate setting, testing for trends in some parametric but monotonic subclass of the general spline approach (linearity has been explored here), and a more in-depth examination of the issue of proportionality of hazards. A more general class of models that is based on the notion of pseudosplines, as in Hastie (1996), is currently being developed by our group; results will be reported elsewhere. In this class of models, examination of the adequacy of increasingly complex forms of polynomials would be natural, due to the general structure of orthogonal-polynomial based pseudosplines, as opposed to the penalized B-splines discussed in this article.

The issue of dependent censoring, and hence competing risks, was not particularly germane to the analysis of the BCPT data. This was because subjects were not censored after observation of any one of the outcomes. Rather, treatment was stopped, but subjects were followed until the end of the study, possible death, and other noninformative censoring processes. So, for the analysis of the BCPT data, allowing for time-dependent treatment was adequate. But, one could easily envision a scenario wherein subjects are censored for most outcomes as soon as one of the outcomes is observed. In such cases, the development of methods that allows for dependent censoring becomes important. Generally speaking, the marginal modeling paradigm that we have followed in this article is not amenable to such dependent censoring problems.

## Acknowledgements

This research was partially supported by an idea award from the U.S. Department of Defense (DAMD17-99-9356). We would like to thank Dr Joe Costantino of the NSABP-BCPT for providing the illustrative data set and for helpful discussions about the study. We would also like to thank the editor, associate editor, and two anonymous reviewers for providing us with very insightful comments.

## Résumé

Dans le cadre du "National Surgical Adjuvant Breast and Bowel Project" (projet national de traitement post-opératoire dans les cancers du sein et de l'intestin), un essai thérapeutique randomisé (le "Breast Cancer Prevention Trial," BCPT, essai de prévention du cancer du sein) a été mis en place pour évaluer l'efficacité du tamoxifène dans la prévention du cancer du sein. Outre l'incidence du cancer du sein, des données ont été recueillies sur plusieurs autres critères, éventuellement délétères, comme le cancer invasif de l'endomètre, la cardiopathie ischémique, l'accident ischémique transitoire, la thrombose veineuse profonde et/ou l'embolie pulmonaire. Dans cet article, nous présentons les résultats d'une analyse des données du BCPT qui illustre une nouvelle technique de modélisation pour évaluer l'efficacité du tamoxifène dans la prévention du cancer du sein. Nous généralisons le modèle flexible de Gray (1994: Biometrics; 50, 640-652) afin d'autoriser des inférences sur des critères multiples dépendant du temps dans le cadre de la modélisation marginale de Wei, Lin et Weissfeld (1989: JASA;84, 1065-1073). Le modèle proposé permet de faire des inférences à partir de délais de survenue d'événements multiples tout en permettant une plus grande souplesse dans la modélisation des effets des facteurs pronostiques avec des relations exposition-réponse non linéaires. Les résultats d'études de simulation sur les propriétés des tests asymptotiques dans de petits échantillons sont aussi présentés.

## References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics* **10,** 1100–1120.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

DeBoor, C. (1974). *A Practical Guide to Splines.* New York: Springer-Verlag.

Fay, P. F. and Graubard, B. I. (2001). Small-sample adjustments for Wald-Type tests using sandwich estimators. *Biometrics* **57,** 1198–1206.

Fisher, B., Dignam, J., Bryant, J., et al. (1996). Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. *Journal of the National Cancer Institute* **88,** 1529–1542.

Fisher, B., Costantino, J. P., and Wickerham, D. L., et al. (1998). Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *Journal of the National Cancer Institute* **90,** 1371–1388.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., and Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* **81,** 1879–1886.

Gail, M. H., Costantino, J. P., Bryant, J., Croyle, R., Freedman, L., Helzlsouer, K., and Vogel, V. (1999). Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute* **91,** 1829–1845.

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

Gray, R. J. (1994). Spline-based test in survival analysis. *Biometrics* **50,** 640–652.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models.* London: Chapman and Hall.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association* **55,** 698–707.

Hastie, T. J. (1996). Pseudosplines. *Journal of the Royal Statistical Society, Series B* **58,** 379–396.

Hastie, T. J. and Tibshirani, R. J. (1990a). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46,** 1005–1016.

Hastie, T. J. and Tibshirani, R. J. (1990b). *Generalized Additive Models.* London: Chapman and Hall.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross validation. *SIAM Journal of Science and Statistical Computation* **9,** 531–542.

Wang, Y. and Taylor, J. M. G. (1995). Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine* **14,** 1205–1218.

Wei, L. J. and Stram, D. O. (1988). Analysing repeated measurements with possibly missing observations by modeling marginal distributions. *Statistics in Medicine* **7,** 139–148.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

*Received December* 2001. *Revised March* 2003.
*Accepted March* 2003.

## Appendix

### *Calculation of $W_g$*

The robust variance estimator introduced by Wei et al. (1989) for inference across margins uses a plug-in estimator for covariances between the scores of the $g$th and $v$th margins.

For the $g$th type of failure, let

$$N_{gi}(t) = I(X_{gi} \le t, \Delta_{gi} = 1),$$
$$Y_{gi}(t) = I(X_{gi} \ge t)$$

and

$$M_{gi}(t) = N_{gi}(t) - \int_0^t Y_{gi}(u)\lambda_{gi}(u)\,du,$$

where $I(\cdot)$ denotes the indicator function. Then, it is straightforward to show that the penalized score function has the form

$$U_g^{(p)}(\psi_g) = U_g(\psi_g) - \lambda_g \tilde{K}_g \psi_g$$

where

$$U_g(\psi_g) = \sum_{i=1}^{n} \int_0^t P_{gi}(u)\,dM_{gi}(u) - \int_0^t \frac{\displaystyle\sum_{i=1}^{n} Y_{gi}(u)P_{gi}(u)exp\big\{\psi_g^T P_{gi}(u)\big\}}{\displaystyle\sum_{i=1}^{n} Y_{gi}(u)exp\big\{\psi_g^T P_{gi}(u)\big\}}\,d\bar{M}_g(u) \tag{A.1}$$

and $\bar{M}_g(u) = \sum_{i=1}^{n} M_{gi}(u)$.

Based on arguments that are parallel to those in Wei et al. (1989), the asymptotic covariance matrix between $\sqrt{n}(\hat{\psi}_g - \psi_g)$ and $\sqrt{n}(\hat{\psi}_v - \psi_v)$ is given by

$$\hat{D}_{gv}(\hat{\psi}_g, \hat{\psi}_v) = \hat{V}_g(\hat{\psi}_g) E\big\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\big\}\hat{V}_v(\hat{\psi}_v),$$

where

$$w_{gj} = \int_0^\infty \big\{P_{gj}(t) - s_g^{(1)}(\psi_g;t)\big/s_g^{(0)}(\psi_g;t)\big\}\,dM_{gj}(t),$$

$$s_g^{(1)}(\psi_g;t) = E\big[Y_{gi}(t)P_{gi}(t)\exp\big\{\psi_g^T P_{gi}(t)\big\}\big],$$

and

$$s_g^{(0)}(\psi_g;t) = E\big[Y_{gi}(t)\exp\big\{\psi_g^T P_{gi}(t)\big\}\big].$$

We then use a plug in estimate for $E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\}$, which takes the form of $\hat{C}$ as in (10). This estimator turns out to be asymptotically the same as the estimator proposed in Wei et al. (1989), since the penalty converges to zero under the null hypothesis. For this reason, the penalty term is dropped in the plug-in estimate for $E\{w_{g1}(\hat{\psi}_g) w_{v1}(\hat{\psi}_v)^T\}$. We define

$$W_{gi}(\psi_g) = \Delta_{gi}\left\{P_{gi}(X_{gi}) - \frac{S_g^{(1)}(\psi_g;X_{gi})}{S_g^{(0)}(\psi_g;X_{gi})}\right\} - \sum_{l=1}^{n} \frac{\Delta_{gl} Y_{gi}(X_{gl}) exp\big\{\psi_g^T P_{gi}(X_{gl})\big\}}{n S_g^{(0)}(\psi_g;X_{gl})} \times \left\{P_{gi}(X_{gl}) - \frac{S_g^{(1)}(\psi_g;X_{gl})}{S_g^{(0)}(\psi_g;X_{gl})}\right\}, \tag{A.2}$$

$$S_g^{(1)}(\psi;t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t)P_{gi}(t)exp\big\{\psi_g^T P_{gi}(t)\big\},$$

and

$$S_g^{(0)}(\psi;t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t)exp\big\{\psi_g^T P_{gi}(t)\big\}.$$

The above asymptotic results are based on the approach used in Wei et al. (1989). Note that $\hat{Q}$ is constructed as a function of the information matrix, the penalty matrix, the smoothing parameter and the individual elements of the unpenalized score vector, that is, a separate term is computed for each of the $n$ observations. Note that, for the above approximation, the penalized versions of the likelihood and the score functions are used to compute the information matrix, while the unpenalized score vector is used in the plug-in estimator for the computation of $W$ as given in (A.2). Note also that the penalty matrix $\tilde{K}_g$ contributes to the penalized score and information matrix only for the last $(m + 2)$ components of $\psi_g$. Inferential procedures for the first $p$ parametric terms are directly analogous to those outlined in Wei et al. (1989).