AD_____

Award Number: DAMD17-99-1-9356

TITLE: Innovative Statistical Approaches to Modeling Multiple Outcome Data from the NSABP BCPT

PRINCIPAL INVESTIGATOR: Lisa Weissfeld, Ph.D.
Kiros Berhane, Ph.D.

CONTRACTING ORGANIZATION: University of Pittsburgh
Pittsburgh, Pennsylvania 15260

REPORT DATE: September 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September, 2000	Annual, (1 Sep 99 – 31 Aug 00)

4. TITLE AND SUBTITLE
Innovative Statistical Approaches to Modeling Multiple Outcome Data from the NSABP BCPT

5. FUNDING NUMBERS
DAMD17-99-1-9356

6. AUTHOR(S)
Lisa Weissfeld, Ph.D.
Kiros Berhane, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

E-Mail: lweis@imap.pitt.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The goals of this IDEA award are two fold, namely, to develop and refine statistical methodology for the analysis of a data set where multiple outcomes or disease incidence endpoints are of interest and to apply these methods to the analysis of a data set in prevention such as the National Surgical Adjuvant Breast and Bowel Project's Breast Cancer Prevention Trial. The software for the fitting of the B-spline models, the first type of model that was proposed in this study, is currently being tested. Software development for the pseudospline approach is well underway. The needed software to conduct simulation studies is being tested for generating data from a bivariate exponential distribution and a second routine is under development.

14. SUBJECT TERMS
Breast Cancer, Statistical Methodology, Survival Analysis

15. NUMBER OF PAGES
21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

INTRODUCTION:

This work involves the development of statistical methodology for the analysis of multiple outcome data. The goal of this work is to extend the current statistical methodology, in particular, the method proposed by Wei, Lin and Weissfeld (1989). In the proposed methodology, the standard Cox model used in this multiple outcome procedure is replaced with a spline based version of the Cox model that was proposed by Gray (1992). The advantage of this approach is that researchers obtain a detailed description of the relationship between survival time and a covariate that is not available using the standard Cox regression model.

BODY:

The work included in the statement of work involves several components, the development of flexible marginal models for multiple time to event data using penalized B-spline based models, to extend these models using psuedosplines, and the development of regression diagnostics and goodness-of-fit tests for these models. Progress has been made on each of these aims.

The investigators, Dr. Kiros Berhane and Dr. Lisa Weissfeld, are at the University of Southern California and the University of Pittsburgh, respectively. There is a graduate student researcher at each site who works closely with the faculty member. These individuals are Mr. Zekarias Berhane at the University of Pittsburgh and Ms. Maria Faccuseh at the University of Southern California. There have been two meetings between the investigators over the past year. The first meeting took place in March when Dr. Berhane visited the University of Pittsburgh and the second meeting took place in August at the Joint Statistical Meetings in Indianapolis. The Pittsburgh meeting was used to work on software development, to discuss inferential procedures for the proposed methodology, and to meet with Dr. Costantino, the NSABP investigator who is affiliated with the project. The meeting in August was used to set priorities and goals for the upcoming 6 months. The graduate student researcher from the University of Pittsburgh was also at both of these meetings.

Throughout much of the academic year a research group examining the use of spline based survival models was formed. This group of University of Pittsburgh researchers consists of Dr. Weissfeld, Dr. Joyce Chang, Dr. Jeong and three Ph.D. students who are working with Dr. Weissfeld. Dr. Chang did much of the work on residual analysis that will be extended to the spline based model setting. Dr. Jeong is an NSABP researcher and will also help with the analysis of NSABP BCPT data. This group met weekly and discussed literature in the area of spline based models.

Specific Aim 1:

The goal of this aim is to develop flexible marginal models for multiple time to event data using penalized B-spline based models. We have completed the theoretical development of the model and justified the methods for inference. This work is presented in the attached paper. We now have preliminary software to implement these models. The development of the software has taken considerable time. At this point in time we have a PC-based program that we are using. The development of this PC-based software is key since Dr. Weissfeld and Dr. Berhane are at two different locations with Dr. Weissfeld being at the University of Pittsburgh and Dr. Berhane being at the University of Southern California. The initial work in this area involved the use of a UNIX based program that required access to either a Sun Work Station of a UNIX based mainframe. Dr. Robert Gray, who wrote the original program kindly provided us with a Windows version of the software in early 2000. We now have a preliminary version of the program for multiple time to event data, which we are in the process of testing. We are also in the process of testing the

software for the simulation study. Programs are e-mailed between Drs. Weissfeld and Berhane and the graduate student researchers who are also working on the project.

We are currently able to simulate data from a bivariate exponential distribution and are in the process of finishing a second routine for the generation of data from a bivariate exponential distribution that was proposed by Sarkar. We expect that this aspect of the work will be completed shortly so that the simulation portion can be added to the attached draft of the paper. We are also in the process of requesting a data set from the NSABP BCPT and should have a data set shortly. We are approximately 2 to 3 months behind schedule on the work on this aim.

Specific Aim 2:

The goal of this aim is to develop flexible marginal models for multiple time to event data using pseudospline based models for time to event data. We have completed the theoretical development of this model and the justification of the proposed inferential procedures. We are in the process of developing software to implement the model. The software development is nearing completion for this part of the project. Dr. Berhane and his graduate student researcher at USC have worked on this intensively over the past month because of the graduate student researcher's decision to leave USC. The work done on the software will be linked with work that has been done at the University of Pittsburgh. The graduate student researcher at the University of Pittsburgh is very familiar with the program and the work that needs to be done to see the project through to completion. The simulation programs written for Aim 1 will apply directly to simulation from this model as well so that new software development is not necessary for this phase of the simulation study.

The proposed work on this aim is well ahead of schedule and should be completed within the next several months.

Specific Aim 3:

We have begun work on the development of regression diagnostics for this model. Zekarias Berhane, the graduate student researcher based at the University of Pittsburgh, will work on the development of regression diagnostics for these models as part of his dissertation work. He has begun to work on the review of the literature in this area. He is currently spending time reviewing the dissertation work of Dr. Joyce Chang, which was used as the springboard for this specific aim. Work on this aim is a bit ahead of schedule.

Specific Aim 4:

The work on this aim is related to that of aim 3. We have not begun the literature review for this work and have instead focused on pushing the work on aim 2 forward.

KEY RESEARCH ACCOMPLISHMENTS:

The key research accomplishments to date from this work are:

- a preliminary version of a program for multiple outcomes using a spline based model.
- a preliminary version of a program for multiple outcomes using a pseudo-spline based model
- software to simulate data from bivariate exponential distributions
- several new lines of research that will be pursued as a result of this work: an extension of the model to handle recurrent event data, an analysis of the NSABP BCPT data using the method of Wei, Lin and Weissfeld (1989).

REPORTABLE OUTCOMES:

- a draft manuscript for the spline based model is attached
- a manuscript for the pseudospline model is currently under development

CONCLUSIONS:

This work will provide researchers with another tool to analysis multiple outcome survival data. The real advantage of this method is that it will allow researchers to examine the effect of a covariate over the course of the study rather than relying on the "average" measure that is provided by the Cox proportional hazards model. Two changes occurred in the plan of the project over the first year: software was moved from the mainframe to a windows-based PC program and greater emphasis was placed on Aim 2 due to an anticipated change in the graduate student researcher at the University of Southern California. Because of this part of the work on Aim 1 was not finished. The work plan for the coming year will essentially follow that proposed in the grant, with the work on Aim 1 being completed shortly.

REFERENCES:

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* **87,** 942-950.

Sarkar, S. K. (1987). A Continuous Bivariate Exponential Distribution. *J. Amer. Statist. Assoc.* **82,** 667-675.

Wei, L.J., Lin, D.Y. and Weissfeld, L.A. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.* **84,** 1065-1073

Appendix

Draft of "Modeling Multiple Time-to-Event Data Using Penalized B-splines"

Modeling Multiple Time-to-Event Data Using Penalized B-splines

Kiros Berhane and Lisa A. Weissfeld *

(August 8, 2000)

---

* Kiros Berhane is Assistant Professor, Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street CHP-220, Los Angeles, CA. Lisa A. Weissfeld is Professor, Department of Biostatistics, University of Pittsburgh, 303 Parran Hall, Pittsburgh, PA 15261.

Abstract

Penalized B-splines have been applied to time-to-event data, providing an extension of the proportional hazards model for a single outcome (Gray, 1994). We use this technique to extend the marginal models of Wei, Lin and Weissfeld (1989). This allows for greater flexibility in modeling the margins and makes formal development of inferential procedures possible. This method is illustrated with an example using data from the NSABP Breast Cancer Prevention Trial.

KEY WORDS: Survival analysis; Smoothing; Ridge regression; Additive models; Splines.

1 Introduction

The advent of promising drugs like tamoxifen in the treatment and/or prevention of breast cancer has ignited both hope and controversy in the scientific world and the general public. The controversy revolves around the adverse side effects of tamoxifen (ref.). *some details about the NSABP-BCPT* In order to demonstrate the positive or negative effectiveness of tamoxifen, one needs to compare the advantages of the drug to its disadvantages in a simultaneous and comprehensive manner. To do this, one needs to be able to make simultaneous inferecne on several time-to-event outcomes and also be able to flexibly model the effect of risk or prognostic factors that have non-linear effects. Considerable progress has been made over the years in the development of models that handle multiple time-to-event outcome data and models that allow for flexible modeling of effects of prognostic factors for sigle time-to-event outcome. But, to date, flexible methods do not exist that allow for! simultaneous inferenc e of multiple, or recurrent, time-to-event outcomes.

The proportional hazards model (Cox 1972) has received considerable attention as a popular way of modeling, possibly censored, time-to-event data. In addition to the proportionality of the hazards, the model assumes that the effects of the predictors (risk factors) on the response follow a parametric (mostly linear) form. Recently, this assumption has been relaxed to allow for data-dependent, and possibly non-linear, covariate effects by exploiting the flexibility of nonparametric regression techniques (Hastie and Tibshirani 1990). Fully non-parametric proportional hazards models (O'Sullivan (1988) and Hastie and Tibshirani (1990)), while attractively flexible, usually suffer from heavy computational load and lack of formal inferential procedures. Gray (1994) used the concept of pseudo-smoothers, with emphasis to penalized B-splines, to develop formal inference for proportional hazards models. Penalized B-splines provide an elegant compromise between regression splines and smo! othi

ng splines.

Another issue in the analysis of time-to-event data is the modeling of multiple, or recurrent, outcomes. The problem of modeling multiple, or recurrent, time to event data has received considerable attention in the statistical literature. For multiple outcome data, Wei, Lin and Weissfeld (1989) propose the use of marginal modeling. For the analysis of recurrent event data, Prentice, Williams and Peterson (1981) propose the use of conditional models, Andersen and Gill (1982) propose a modification of the proportional hazards model and Wei, Lin and Weissfeld (1989) apply the marginal approach for modeling such data. However, these methods have not been extended to include flexible and possibly nonlinear effects of prognostic factors. On the other hand, many researchers have demonstrated that important prognostic factors (e.g. BMI) have a markedly non-linear effect on breast cancer survival and/or prognosis (Gray, 1994). These methods, however, are limited to single outcomes and do not lend themselves to simultaneous inference of several time-to-event outcomes.

In this article, we extend the marginal models of Wei, Lin and Weissfeld (1989) to allow modeling flexibility via the use of penalized B-splines in the style of Gray (1994). See also Hastie (1996) for a detailed discussion on a more general class of pseudo-smoothers. The remainder of the paper is organized as follows. In §2, we give background material on penalized B-splines and details on the proposed flexible marginal models. In §3, we perform extensive simulation studies to study the small sample properties of the proposed inferential procedures. §4 summarizes the results from applications of the proposed methodology to data from the NSABP Breast Cancer Prevention Trial (BCPT), known as Protocol B-14, comparing tamoxifen to placebo for the prevention of recurrence in subjects with breast cancer. In §5, we summarize the main results and give details on future directions for research. The details on the theoretical development and asymptotic properties of the inferential procedures are given in the Appendix (?). *Are we still planning to do this?*

2 Proposed Model

2.1 Background

To fix ideas, we first consider a non-parametric regression model in the univariate framework. Let $(x_1, y_1), ..., (x_n, y_n)$ denote a set of n independent observations and consider a regression model of the form

$$y_i = f(x_i) + \epsilon_i \ , \tag{1}$$

where $i = 1, ..., n$, $f(x)$ is an unspecified smooth function and $\epsilon_i \in N(0, \sigma_\epsilon^2)$. In the non-parametric regression setup, one estimates $f(x)$ via a scatterplot smoother. A scatterplot smoother is said to be linear if, concentrating on the computations of the function only at the design points in $\mathbf{x} = (x_1, ..., x_n)$, it can be written as a linear map $S : R^n \to R^n$ defined by $\hat{\mathbf{y}} = S\mathbf{y}$, where $\mathbf{y} = (y_1, ..., y_n)$ is the response vector. Here S is referred to as a smoother matrix and is analogous to the hat matrix in linear regression. From this point onwards, our discussion focuses on penalized regression splines, even though the idea of pseudo-smoothers applies, in principle, to any linear smoother.

For a given number of knots and fixed positions of the knots, a regression spline representation that uses the B-spline basis functions $B_1(x), ..., B_{m+4}(x)$ is given as

$$f(x) = \gamma_0 + \gamma_1 x + \sum_{l=2}^{m+3} \gamma_l B_l(x) \ .$$

Note that the constant and linear functions are stated explicitly and only (m+2) of the B-spline basis functions are used for identifiability (De Boor, 1974). A penalized form of this B-spline representation is given by subtracting the following roughness penalty from the resulting residual sum of squares:

$$\lambda \int [f''(u)]^2 du \ .$$

Here, λ is a smoothing parameter that determines the amount of smoothness. Recognizing that the penalty function given above is quadratic in the parameter vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_{m+3})$, one could rewrite it as

$$\lambda \boldsymbol{\gamma}^T \mathbf{K} \boldsymbol{\gamma} \ ,$$

where $\mathbf{K}$ is a positive defiite matrix that is a function of the covariate, and more specifically, of the knot locations. Note that $\mathbf{K}$ is an $(m+4) \times (m+4)$ matrix with the the first two rows and two columns as zeros, since the constant and linear functions pass unpenalized.

This idea was first introduced in Hastie and Tibshirani (1990) and its use in univariate proportional hazards models was detailed in Gray (1994). Gray (1994) also develops (and validates) appropriate testing procedures for main effects, interactions and non-linear time depenency of covariate effects for the proportional hazards model *(any more details here?)*. In this paper, we extend this technology to the multivariate proprtional hazards models of Wei, Lin and Weissfeld (1989).

2.2 The model

To model marginal distributions of multivariate time-to-event data, let us consider a flexible proportional hazards model for each of the G failure types. For the g^{th} type of failure of the i^{th}, $i = 1, ..., n$, subject, the model can be written as

$$\lambda_{gi}(t) = \lambda_{g0}(t) exp\{\sum_j f_{jg}(Z_{jgi})\} \ , \quad t \geq 0 \ , \tag{2}$$

where $\lambda_{g0}(t)$ is an unspecified baseline hazard function and f_{jg}, $j = 1, ..., p$, denotes the unspecified smooth functions. In the usual setup (Cox, 1972), one observes data of the form $(X_{gi}, Z_{gi}, \Delta_{gi})$, where $X_{gi} = min(\tilde{X}_{gi}, C_{gi})$, C_{gi} is the censoring time, $Z_{gi}(t) = (Z_{1gi}(t), ..., Z_{pgi}(t))^T$ and $\Delta_{gi} = 1$ if $X_{gi} = \tilde{X}_{gi}$ and 0 otherwise.

Model (2) is fully non-parametric and quite general. Note also that the fully linear model of Wei, Lin and Weissfeld (1989) forms a special case of (2) where $f_{jg}(Z_{jgi}) = \beta_{jg} Z_{jgi}$. For this fully linear model, the partial likelihood is given as

$$PL_g(\beta) = \Pi_{i=1}^{n} \Big(\frac{exp\{\boldsymbol{\beta}_{g(T)} Z_{gi}(X_{gi})\}}{\sum_{l \in \mathcal{R}_g(X_{gi})} exp\{\boldsymbol{\beta}_{g(T)} Z_{gl}(X_{gl})\}} \Big)^{\Delta_{gi}} , \tag{3}$$

where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{1g}, ..., \boldsymbol{\beta}_{pg})^T$ 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 $\partial log PL_g(\boldsymbol{\beta}_g)/\partial \boldsymbol{\beta}_g = 0$, $\hat{\boldsymbol{\beta}}_g$, can be shown to be a consistent estimator of $\boldsymbol{\beta}_g$ provided that the fully linear model is correctly specified (Anderson and Gill, 1982).

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. For simplicity of discussion, we first discuss a model with p parametric and an additional non-parametric term, *i.e.*,

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

where $j = 1, ..., p$. We propose to estimate $f_g(h_{gi})$ using the penalized regression spline approach discussed in §2.1, *i.e.*,

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

Note that, we have now dropped the constant term since it is accounted for by the baseline hazard. Following the notations of Gray (1994), let $\boldsymbol{\gamma}_g = (\gamma_{g2}, ..., \gamma_{g(m+3)})$ 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

$$PL_g^p(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) - 1/2\lambda_g \boldsymbol{\eta}_{g(T)} \mathbf{K}_g \boldsymbol{\eta}_g \ . \tag{6}$$

where $\mathbf{K}$ is a positive defiite matrix that is a function of the covariate h_g as in §2.1. Note that $\mathbf{K}$ is an $(m+3) \times (m+3)$ matrix with the the first row and column as zeros, since the linear function passes unpenalized.

The hypotheses of interest with respect to the smooth function are then $\boldsymbol{\gamma}_g = \mathbf{0}$ and $\boldsymbol{\eta}_g = \mathbf{0}$, representing the hypotheses of "no effect" and "linear effect" respectively. *more details here on summarized version of Gray's tests for univariate outcome*

It is straightforward to extend this model to allow for multiple, say q, non-parametric terms. In this case, $\boldsymbol{\eta}_g$ would be a bigger vector that augments contributions from the basis functions of the q terms. Here, $\boldsymbol{\eta}_g = (\boldsymbol{\eta}_{g1} : ... : \boldsymbol{\eta}_{gq})$ would be of dimension $\sum_{l=1}^{q}(m_l + 3) \times 1$ and the penalty term would be the sum of the q penalty functions leading to

$$PL_g^p(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) - 1/2 \sum_{j=1}^{q} \lambda_{gj} \boldsymbol{\eta}_{gj}^T \mathbf{K}_{gj} \boldsymbol{\eta}_{gj} \ . \tag{7}$$

where each non-parametric term has its own smoothing parameter, λ_{gj}, and penalty function $\mathbf{K}_{gj}$. Here, one could test for the "overall" effect or "linearity" of the individual non-parametric terms or for a combination of them. *more details here*

2.3 Inference

While making inference on each of the margins is important, this could be done easily by using developments in Gray (1994). Our interest is mainly in being able to conduct simultaneaous inference on several time-to-event outcomes in models that have non-parametric smooth

terms. Once the marginal distributions are modeled, then the methods described in Wei, Lin and Weissfeld (1989) can be extended to test for trends across parameter estimates and to combine estimates across margins to test for covariate effects of interest. In our extensions to the multivariate survival data framework, we use slightly diferent but equivalent testing procedures (compared to those of Gray (1994)) for both the univariate (marginal) and the simultaneous inferences. Let us consider the case where we have p parametric terms and one additional non-parametric term as given by (4). Then, for outcome g, the unpenalized part of equation (6) can be written as

$$PL_g(\boldsymbol{\beta}_g, \boldsymbol{\eta}_g) = \Pi_{i=1}^{n} \Big(\frac{exp\{\sum_{j=1}^{p} Z_{gj}\beta_{gj}(X_{gi}) + h_g\gamma_1(X_{gi}) + \sum_{l=2}^{m+3} B_{lg}(h_g)\gamma_{lg}(X_{gi})\}}{\sum_{s\in\mathcal{R}_g(X_{gi})} exp\{\sum_{j=1}^{p} Z_{gj}\beta_{gj}(X_{gs}) + h_g\gamma_1(X_{gs}) + \sum_{l=2}^{m+3} B_{lg}(h_g)\gamma_{lg}(X_{gs})\}} \Big)^{\Delta_{gi}} , \quad (8)$$

where all components are as defined in §2.2, for the g^{th} type of failure. Let $\boldsymbol{\psi}_g = (\boldsymbol{\beta}_g, \boldsymbol{\eta}_g)$ and $P_g = (Z_{1g} : ... : Z_{pg} : h_g : B_{2g}(h_g) : ... : B_{m+3,g}(h_g))$ with P_{gr} denoting the r^{th} column vector, $r = 1, ..., (m + p + 3)$. Letting $\hat{A}_g$ be the unpenalized information matrix for the g^{th} outcome as a function of $\boldsymbol{\psi}$, it can be shown that

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

where $U_g(\boldsymbol{\psi}_{g(T)})$ is the score vector and $\boldsymbol{\psi}_{g(T)}$ is the vector of true parameter values for the g^{th} outcome (Gray, 1994). 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 where

$$V = (A_g + \lambda_n\mathbf{K})^{-1}A_g(A_g + \lambda_n\mathbf{K})^{-1} , \quad (9)$$

To develop the simultaneous inferential procedures for several outcomes, we first define

$$W_{gir}(\hat{\psi}_{gr}) = \Delta_{gi}\Big(\sum_{j=1}^{m} P_{gr(jr)}(X_{gi}) - \frac{S_g^{(1)}(\hat{\psi}_{gr}; X_{gi})}{S_g^{(0)}(\hat{\psi}_{gr}; X_{gi})}\Big) - \sum_{j=1}^{n} \frac{\Delta_{gj} Y_{gi}(X_{gj}) exp(\sum_{k=1}^{p} P_{gk}\psi_{gk}(X_{gj})}{n S_g^{(0)}(\psi_{gr}; X_{gj})}$$

$$\Big(\sum_{j=1}^{m} P_{gr(jr)}(X_{gj}) - \frac{S_g^{(1)}(\psi_{gr}; X_{gj})}{S_g^{(0)}(\psi_{gr}; X_{gj})}\Big) \ , \tag{10}$$

where

$$S_g^{(1)}(\psi_r; t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t)(\sum_{j=1}^{m} P_{gr(jr)}) exp(\sum_{s=1}^{m} P_{gs}\psi_{gs}(t)) \ ,$$

$$S_g^{(0)}(\psi_r; t) = n^{-1} \sum_{i=1}^{n} Y_{gi}(t) exp(\sum_{s=1}^{m} P_{gs}\psi_{gs}(t)) \ ,$$

and $Y_{gi}(t) = I(X_{gi} \geq t)$. Then, the asymptotic covariance matrix between $\sqrt{n}(\hat{\psi}_u - \psi_u)$ and $\sqrt{n}(\hat{\psi}_v - \psi_v)$ can be consistently estimated by

$$\hat{D}_{uv}(\hat{\psi}_u, \hat{\psi}_v) = \hat{V}_u(\hat{\psi}_u)\hat{\psi}_{uv}(\hat{\psi}_u, \hat{\psi}_v)\hat{V}_v(\hat{\psi}_v) \ , \tag{11}$$

where $\hat{\psi}_{uv}(\hat{\psi}_u, \hat{\psi}_v) = n^{-1} \sum_{j=1}^{n} W_{uj}(\hat{\psi}_u) W_{vj}(\hat{\psi}_v)^T$, where W_{uj} and W_{vj} are defined in (10). Thus, the covariance matrix of $(\hat{\psi}_1, ..., \hat{\psi}_G)$ can be consistently estimated by

$$\hat{Q} = \begin{pmatrix} \hat{D}_{11}(\hat{\psi}_1, \hat{\psi}_1) & \dots & \hat{D}_{1G}(\hat{\psi}_1, \hat{\psi}_G) \\ \vdots & \ddots & \vdots \\ \hat{D}_{G1}(\hat{\psi}_G, \hat{\psi}_1) & \dots & \hat{D}_{GG}(\hat{\psi}_G, \hat{\psi}_G) \end{pmatrix} . \tag{12}$$

The above asymptotic results are based on the approach used in Wei, Lin and Weissfeld (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 score vector, that is,

a separate term is computed for each of the n observations. Note also that, for the above approximation, the penalized version of the likelihood is used to compute the information matrix while the original (unpenalized) version of the likelihood is used for the computation of the individual elements of the score vector. An alternative estimator can be obtained by using the penalized version of the likelihood for the computation of W as given in equation (10) *Is this true?.*

Note that the penalty matrix K_g contributes to the penalized score and information matrix only for the last $(m+2)$ components of ψ_g. The inference for the first p parametric terms is directly analogous to Wei, Lin and Weissfeld (1989). For the non-parametric term, one could conduct simultaneous inference on the "overall" effect and/or "linearity" of h across failure types. Let $\hat{\gamma}_g$ denote the components of $\hat{\psi}_g$ that correspond to the relevant components of the non-parametric term h_g. Let also $\hat{\Gamma}$ denote the relevant sub-matrix of $\hat{Q}$ corresponding to $\hat{\gamma} = (\hat{\gamma}_1, ..., \hat{\gamma}_G)$. Then, one could use the quadratic form $(\hat{\gamma}_1, ..., \hat{\gamma}_G)\hat{\Gamma}(\hat{\gamma}_1, ..., \hat{\gamma}_G)^T$ to conduct a joint test on the null hypotheses given by $H_g = \gamma_g = \mathbf{0}$, $g = 1, ..., G$. 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 ψ_g respectively.

Test for trends? Is it possible in the penalized B-spline framework? This could probably be the advantage of pseudosplines since they have ordered levels of complexity and hence one could test for equality in the comparable components of the smooth functions.

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 that one could develop automaitc procedures for selecting the number and position of the knots (which are usually between 10-15, per outcome) and the value of α_g.

We will disucuss the potential effects of various choices of number of knots in our simulation studies. We follow Gray (1994) in putting the knots at locations that yield approximately equal numbers of observations between knots. The issue of the value of the smoothing paramters could also be addressed as a model selection procedure. But, we do not pursue this issue any further in this manuscript. We, however, intend to report results elsewhere *Is this enough or the right strategy?.*

3 Simulation Study

Initial details as in the outline?

4 Examples: The NSABP Data

Initial details as in the two substantive papers from Joe Costantino?

5 Discussion

- summarize main results and findings
- relevance to breast cancer research
- discuss related research and open areas of research

REFERENCES

Cox, D.R. (1972), "Regression models and life tables (with discussion)", J. R. Statist. Soc. B 34, 187-220.

Golub, G. H. and Van Loan, C. F. (1983), Matrix Computations, Baltimore: Johns Hopkins University Press.

Gray, R. J. (1994), "Spline-Based test in survival analysis", Biometrics 50, 640-652.

Hastie, T. J. (1996). "Pseudosplines", J. R. Statist. Soc. B 58, 379-396.

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

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

Jorgensen, B (1984). "The delta algorithm and GLIM", Int. Stat. Rev. 52, 283-300.

O'Sullivan, F. (1988). "Nonparametric estimation of relative risk using splines and cross validation", SIAM J. Sci. and Stat. Comp. 9, 531-542.

Thisted, R. A. (1988), Elements of Statistical Computing, London: Chapman and Hall.

Wahba, G. (1990), Spline Functions for Observational Data, CBMS-NSF Regional Conf. Series, SIAM. Philadelphia.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989), "Rgeression analysis of multivariate incomplete failure time data by modeling marginal distributions", JASA 84, 1065-1073.