Confidence Band for the Differences between Two Direct Adjusted Survival Curves

Hong Wang
University of Pittsburgh

Xu Zhang
University of Mississippi Medical Center

Abstract

This paper focuses on comparing the survival outcomes of two treatments while adjusting for multiple demographic and clinical factors. We wish to answer the question: "At what times do the survival outcomes of two treatments differ?". Previous work by Zhang and Klein (2001) addressed this issue by considering a stratified Cox model and by constructing the confidence band for the differences between survival functions of treatments, for a given subject. In this study, we utilize direct adjusted survivals of treatments, so that treatment comparison can be performed on a general basis. The confidence band for the differences between direct adjusted survivals was constructed using the simulation method. We developed two SAS macros that compute the band. The first macro was developed for data sets consisting of fixed covariates only. The second macro was developed based on the piecewise proportional hazards Cox model and allows for time-dependent variables. We illustrate both macros by analyzing the survival data of a cohort of lymphoma patients.

Keywords: Cox model, direct adjustment, time-varying effect, confidence band.

1. Introduction

One fundamental question in many clinical studies is to evaluate survival outcomes of different treatments. In controlled clinical trials where treatment groups are balanced in covariates (such as demographics and clinical characteristics), the standard analytical method is to present Kaplan-Meier survival curves by treatment group and perform the log-rank test. In observational studies where the covariates are not balanced over treatment groups, a regression model suitable for censored data would be needed to control for effects of these covariates. In this situation the multi-variable Cox model is a common choice. This model assumes a constant hazard ratio between two levels of a covariate.

When the Cox model is used to analyze treatment-related survival data, the estimation of treatment specific survival is problematic. Several methods have been discussed. One possible method is to set covariate values equal to the means across the sample and then predict survivals of treatments in the fitted Cox model. The underlying rationale behind this method was that the survival of a treatment based on mean covariate values tended to be close to the average survival of the treatment. However, this rationale is actually a flawed one (Nieto and Coresh 1996) and interpretation of this type of survival is awkward when discrete covariates are present. Makuch (1982) and Gail and Byar (1986) advocate the direct adjustment method based on a Cox model. In this method the survival estimates of a treatment for all subjects in the reference population are computed and averaged. Zhang, Loberiza, Klein, and Zhang (2007) studied the direct adjusted survival curves based on a Cox model stratified by treatment groups. Such a model allows treatment effects to vary over time. Zhang *et al.* (2007) provided a SAS macro that implemented the pointwise comparison between treatments. The propensity-score-based approach has been studied by some statisticians and a review of this method can be found in MacKenzie, Brown, Likosky, Wu, and Grunkemeier (2012).

When treatment effects change over time, it may be crucial to evaluate at what times the outcomes of two treatments differ. Such a question can be answered by constructing a confidence band in an interval $[t_1, t_2]$ for the differences between two treatment-specific survivals. Lin, Fleming, and Wei (1994) proposed the simulation method to construct a confidence band for a survival function based on a Cox model. This method has recently been recognized to be an example of the general wild bootstrap which was first applied to regression analysis by Wu (1986). In the wild bootstrap method the residuals are multiplied by independent random variates with mean 0 and variance 1 and the subsequent variance estimator protects against error variance heteroscedasticity. We refer the reader to the recent paper by Beyersmann, Termini, and Pauly (2013) for a detailed description of the wild bootstrap method. Zhang and Klein (2001) applied the wild bootstrap method to the problem of treatment comparison based on the stratified Cox model. They predicted the survival functions of treatments for a given set of covariate values, and developed the simulation-based confidence band for differences between two survival functions. The band helps researchers to assess when the outcomes of two treatments differ for a given subject.

This paper focuses on the comparison of direct adjusted survivals of treatments, based on a stratified Cox model. We aim at obtaining a confidence band for the differences between two direct adjusted survivals, using the simulation method. This study extends the previous work by Zhang and Klein (2001). Usage of direct adjusted survivals for the band allows for treatment comparison on a general basis. It provides a decision making tool for physicians and patients who want to know whether a short term survival benefit exists for one treatment or at which time point one treatment starts to have a survival advantage. We developed two SAS macros, SURVBAND and TIMEBAND, that compute the bands. The first macro was developed for data sets consisting of fixed covariates only. The second macro was developed based on the piecewise proportional hazards Cox model allowing for time-dependent variables. We illustrate both macros by analyzing the survival data of a cohort of lymphoma patients. The macros were written using the SAS/IML language (SAS Institute Inc. 2008) and the PHREG procedure in SAS/STAT (SAS Institute Inc. 2009). Both products, SAS/STAT and SAS/IML, need to be installed in order to run the macros.

The remainder of the paper is organized as follows. In Section 2 we briefly describe the theoretical foundations and present the formulas of the method. Section 3 presents two SAS

macros and explains the macro parameters. A simulation study was conducted to assess the performance of the simulation-based band. The simulation results are provided in Section 4. In Section 5 we apply both macros to a real data set on survival of lymphoma patients. Final discussions are given in Section 6.

2. Comparison of direct adjusted survivals

2.1. Notations

The sample can be summarized as $\{X_{ij}, \Delta_{ij}, \boldsymbol{Z}_{ij}\}$ for $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$, where the first subscript refers to the treatment group. For the jth subject in the ith treatment group, X_{ij} is the observed time, Δ_{ij} is the event indicator, and $\boldsymbol{Z}_{ij}$ is the vector of covariates. We consider a sample with two treatments ($K = 2$), and the sample size is given by $n = n_1 + n_2$.

Assume that the Cox model, when stratifying on treatment group, is valid to describe the hazard functions. Given $\boldsymbol{z}$, the conditional hazard function of ith treatment at time t is

$$\lambda_i(t|\boldsymbol{z}) = \lambda_{0i}(t) \exp(\boldsymbol{\beta}^\top \boldsymbol{z}), \qquad i = 1, 2, \tag{1}$$

where $\lambda_{0i}(t)$ is an arbitrary nonnegative function, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{z}$ is a vector of covariates. Let $\Lambda_{0i}(t)$ be the cumulative baseline hazard function, $\Lambda_{0i}(t) = \int_0^t \lambda_{0i}(s)ds$.

The prediction of survival functions is a complex issue when time-dependent covariates are present. A detailed discussion of time-dependent covariates is given in Section 3.2. For simplicity, we only present the formulas for fixed covariates in Sections 2.1 and 2.2. For a given set of covariate values $\boldsymbol{z}$, the survival function for the ith treatment is given by

$$S_i(t|\boldsymbol{z}) = \exp\left(-\Lambda_{0i}(t)e^{\boldsymbol{\beta}^\top \boldsymbol{z}}\right).$$

The direct adjusted survival of the ith treatment averages the survival estimates of the ith treatment for all subjects in the sample,

$$\bar{S}_i(t) = \frac{1}{n}\sum_{k=1}^{2}\sum_{j=1}^{n_k} S_i(t|\boldsymbol{Z}_{kj}) = \frac{1}{n}\sum_{k=1}^{2}\sum_{j=1}^{n_k} \exp\left(-\Lambda_{0i}(t)e^{\boldsymbol{\beta}^\top \boldsymbol{Z}_{kj}}\right). \tag{2}$$

In the following we introduce additional relevant notations. We define the counting process that counts occurrence of the event by time t by $N_{ij}(t) = I(X_{ij} \leq t, \Delta_{ij} = 1)$ and the at-risk process by $Y_{ij}(t) = I(X_{ij} \geq t)$. The martingale process is given by

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(u)\lambda_{0i}(u)\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_{ij})du.$$

We also define the following notations,

$$\boldsymbol{S}_i^{(p)}(\boldsymbol{\beta}, t) = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}(t)\boldsymbol{Z}_{ij}^{\otimes p}\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_{ij}), \qquad p = 0, 1, 2,$$

$$\boldsymbol{E}_i(\boldsymbol{\beta}, t) = \boldsymbol{S}_i^{(1)}(\boldsymbol{\beta}, t)/S_i^{(0)}(\boldsymbol{\beta}, t),$$

$$\boldsymbol{e}_i(\boldsymbol{\beta}, t) = \mathsf{E}\{\boldsymbol{S}_i^{(1)}(\boldsymbol{\beta}, t)\}/\mathsf{E}\{\boldsymbol{S}_i^{(0)}(\boldsymbol{\beta}, t)\},$$

where for a column vector $\boldsymbol{a}$, $\boldsymbol{a}^{\otimes 0} = 1, \boldsymbol{a}^{\otimes 1} = \boldsymbol{a}, \boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}^\top$. Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimate of the partial likelihood. $\boldsymbol{\Sigma}^{-1}$ is the covariance matrix for the limiting distribution of $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. $\widehat{\boldsymbol{\Sigma}}$ denotes the observed information matrix. $\widehat{\Lambda}_{0i}$ denotes the Breslow estimator of the cumulative hazard function for the ith stratum. The explicit formulas of these estimators can be found in Zhang and Klein (2001, pp. 245–246). The survival function for the ith treatment, given $\boldsymbol{z}$, can be estimated by

$$\widehat{S}_i(t|\boldsymbol{z}) = \exp\left(-\widehat{\Lambda}_{0i}(t)e^{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{z}}\right).$$

2.2. Comparison of direct adjusted survivals

Treatment effect assessment is performed in terms of finding the confidence band for $\bar{S}_1(t) - \bar{S}_2(t)$ over the interval $[t_1, t_2]$. The hypothesis to be tested is

$$H_0 : \bar{S}_1(t) - \bar{S}_2(t) = 0, \quad \forall t : t_1 \le t \le t_2,$$

versus the two-sided alternative. According to Zhang *et al.* (2007), $\bar{S}_1(t)$ and $\bar{S}_2(t)$ can be estimated by

$$\widehat{\bar{S}}_1(t) = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \exp\left[-\widehat{\Lambda}_{01}(t)e^{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_{ij}}\right], \tag{3}$$

$$\widehat{\bar{S}}_2(t) = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \exp\left[-\widehat{\Lambda}_{02}(t)e^{\widehat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_{ij}}\right]. \tag{4}$$

Let

$$W(t) = \sqrt{n} \left\{ \widehat{\bar{S}}_1(t) - \widehat{\bar{S}}_2(t) - (\bar{S}_1(t) - \bar{S}_2(t)) \right\}.$$

A general test statistic for testing H_0 is given by

$$\widetilde{Q} = \sup_{t \in [t_1, t_2]} K(t)|W(t)|, \tag{5}$$

where $K(t)$ is a positive weight function. Possible choices of $K(t)$ include the size of the risk set and $\widehat{S}(t)^p[1 - \widehat{S}(t)]^q$, $p \ge 0, q \ge 0$, where $\widehat{S}(t)$ is the Kaplan-Meier estimate of the survival function of the whole cohort. When the size of the risk set or $\widehat{S}(t)^p$ is used as the weight function, the test statistic is adjusted to emphasize the survival differences during the early time period. In the remainder of this section as well as in the SAS macro developed, we let $K(t) \equiv 1$ and focus on the test statistic

$$Q = \sup_{t \in [t_1, t_2]} |W(t)|. \tag{6}$$

Following the standard asymptotic result for the Cox model (Andersen and Gill 1982) and the previous work on survival comparison (Zhang and Klein 2001), $W(t)$ is asymptotically

equivalent to

$$
\begin{aligned}
\widetilde{W}(t) \;=\; & -\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}S_1(t|\boldsymbol{Z}_{ij})e^{\boldsymbol{\beta}^{\top}\boldsymbol{Z}_{ij}}\right]\int_0^t \frac{d\bar{M}_1(u)}{n_1 S_1^{(0)}(\boldsymbol{\beta},u)} \\
& +\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}S_2(t|\boldsymbol{Z}_{ij})e^{\boldsymbol{\beta}^{\top}\boldsymbol{Z}_{ij}}\right]\int_0^t \frac{d\bar{M}_2(u)}{n_2 S_2^{(0)}(\boldsymbol{\beta},u)} \\
& -\left\{\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}[S_1(t|\boldsymbol{Z}_{ij})\boldsymbol{h}_1(t|\boldsymbol{Z}_{ij})-S_2(t|\boldsymbol{Z}_{ij})\boldsymbol{h}_2(t|\boldsymbol{Z}_{ij})]^{\top}\right\}\left(\boldsymbol{\Sigma}^{-1}\right) \\
& \times \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\int_0^{\infty}[\boldsymbol{Z}_{ij}-\boldsymbol{e}_i(\boldsymbol{\beta},u)]dM_{ij}(u)\right\},
\end{aligned}
$$

where $\bar{M}_i(t)=\sum_{j=1}^{n_i}M_{ij}(t)$ and $\boldsymbol{h}_i(t|\boldsymbol{z})=\int_0^t e^{\boldsymbol{\beta}^{\top}\boldsymbol{z}}[\boldsymbol{z}-\boldsymbol{e}_i(\boldsymbol{\beta},u)]\lambda_{0i}(u)du$. Let $\bar{N}_i(t)=\sum_{j=1}^{n_i}N_{ij}(t)$. $\boldsymbol{h}_i(t|\boldsymbol{z})$ can be estimated by

$$
\widehat{\boldsymbol{h}}_i(t|\boldsymbol{z})=\int_0^t e^{\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{z}}\left[\boldsymbol{z}-\boldsymbol{E}_i(\widehat{\boldsymbol{\beta}},u)\right]\frac{d\bar{N}_i(u)}{n_i S_i^{(0)}(\widehat{\boldsymbol{\beta}},u)}.
$$

Based on the martingale central limit theorem, $\widetilde{W}(t)$ converges weakly to a zero mean Gaussian process. The pointwise variance of $\widetilde{W}(t)$ can be consistently estimated by

$$
\begin{aligned}
\widehat{\sigma}^2(t) \;=\; & n\left[\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\widehat{S}_1(t|\boldsymbol{Z}_{ij})e^{\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{Z}_{ij}}\right]^2\int_0^t \frac{d\bar{N}_1(u)}{\left(n_1 S_1^{(0)}(\widehat{\boldsymbol{\beta}},u)\right)^2} \\
& +n\left[\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\widehat{S}_2(t|\boldsymbol{Z}_{ij})e^{\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{Z}_{ij}}\right]^2\int_0^t \frac{d\bar{N}_2(u)}{\left(n_2 S_2^{(0)}(\widehat{\boldsymbol{\beta}},u)\right)^2} \\
& +\widehat{\boldsymbol{\theta}}(t)^{\top}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)\widehat{\boldsymbol{\theta}}(t),
\end{aligned}
$$

where

$$
\widehat{\boldsymbol{\theta}}(t)=\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left[\widehat{S}_1(t|\boldsymbol{Z}_{ij})\widehat{\boldsymbol{h}}_1(t|\boldsymbol{Z}_{ij})-\widehat{S}_2(t|\boldsymbol{Z}_{ij})\widehat{\boldsymbol{h}}_2(t|\boldsymbol{Z}_{ij})\right].
$$

The process $\widetilde{W}(t)$ does not have independent increments, so the limiting distribution is difficult to describe analytically. We use the wild bootstrap resampling method of Lin *et al.* (1994) to approximate the limiting distribution. Note that the martingale $M_{ij}(t)$ in $\widetilde{W}(t)$ has mean 0 and variance $\mathsf{E}[N_{ij}(t)]$. Let G_{ij}'s $(i=1,2;j=1,\ldots,n_i)$ be independent standard normal variates. We replace the martingale $M_{ij}(t)$ with $G_{ij}N_{ij}(t)$ and plug in the estimated

regression coefficients and covariance matrix, leading to the process

$$
\begin{aligned}
\widehat{W}(t) \;=\; & -\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\widehat{S}_1(t|\boldsymbol{Z}_{ij})e^{\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{Z}_{ij}}\right]\sum_{j=1}^{n_1}\int_{0}^{t}\frac{G_{1j}dN_{1j}(u)}{n_1 S_1^{(0)}(\widehat{\boldsymbol{\beta}},u)} \\
& +\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\widehat{S}_2(t|\boldsymbol{Z}_{ij})e^{\widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{Z}_{ij}}\right]\sum_{j=1}^{n_2}\int_{0}^{t}\frac{G_{2j}dN_{2j}(u)}{n_2 S_2^{(0)}(\widehat{\boldsymbol{\beta}},u)} \\
& -\widehat{\boldsymbol{\theta}}(t)^{\top}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)\frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\int_{0}^{\infty}[\boldsymbol{Z}_{ij}-\boldsymbol{E}_i(\widehat{\boldsymbol{\beta}},u)]G_{ij}dN_{ij}(u)\right\}.
\end{aligned}
$$

We treat the G_{ij}'s as random and $\{X_{ij},\Delta_{ij},\boldsymbol{Z}_{ij}\}$ $(i=1,2;j=1,\ldots,n_i)$ as fixed. Following the proof by Lin *et al.* (1994), the limiting distribution of $\widetilde{W}(t)$ is the same as that of $\widehat{W}(t)$. To approximate the limiting distribution of $\widetilde{W}(t)$, a large number of realizations of $\widehat{W}(t)$ can be generated. To construct the $100(1-\alpha)\%$ confidence band for $\bar{S}_1(t)-\bar{S}_2(t)$ in $[t_1,t_2]$, one can approximate the critical value by finding the $(1-\alpha)\times 100$th percentile of the sample $Q^b=\sup_{t_1\le t\le t_2}|\widehat{W}^{(b)}(t)/\widehat{\sigma}(t)|$ for $b=1,\cdots,B$. Let q_α be the $(1-\alpha)\times 100$th percentile of the sample. The band for $\bar{S}_1(t)-\bar{S}_2(t)$ can be constructed by evaluating $\widehat{\bar{S}}_1(t)-\widehat{\bar{S}}_2(t)\pm n^{-1/2}q_\alpha\widehat{\sigma}(t)$.

It is useful to report the p value for testing H_0. Bajorunaite and Klein (2007) proposed the simulation-based approach for the two-sample test of the cumulative incidence function. They provided a procedure for finding the p value for the test. We consider a similar procedure to derive the p value for testing H_0:

1. Evaluate the test statistic on the sample. Let $Q^*=\sup_{t\in[t_1,t_2]}\sqrt{n}\left|\left(\widehat{\bar{S}}_1(t)-\widehat{\bar{S}}_2(t)\right)/\widehat{\sigma}(t)\right|$.

2. Generate standard normal variates and evaluate $\sup_{t\in[t_1,t_2]}|\widehat{W}(t)/\widehat{\sigma}(t)|$.

3. Repeat Step 2 B times. Let Q^b be the evaluated value of $\sup_{t\in[t_1,t_2]}|\widehat{W}^{(b)}(t)/\widehat{\sigma}(t)|$ in the bth iteration. The estimated p value is calculated by the formula $p=\sum_{b=1}^{B}I(Q^b>Q^*)/B$, where $I(\cdot)$ is the identity function.

3. Two **SAS** macros

3.1. Fixed covariates only

The majority of clinical studies collect only fixed covariates such as demographics and baseline clinical characteristics. The formulas described in the previous section are only applicable to data of fixed covariates and were implemented in the **SAS** macro SURVBAND. The macro should be loaded to the current **SAS** session using the %include statement. Suppose that the file containing the macro is named band.sas, and is saved in the local directory. The following statement should appear in a **SAS** program to load the macro.

```
%include "band.sas";
```

The name of the macro is `SURVBAND` and has the following parameters:

```
%macro SURVBAND(inputdata, time, event, group, covlist, starttime,
  stoptime, alpha, nsimu, seed, outdata);
```

There are nine macro parameters. A full explanation of the macro parameters, as well as the data preparation required, is given in the following.

(1) `inputdata`: This parameter specifies the input SAS data set name. This SAS data set needs to contain the observed time variable, the event indicator, the group variable with values $1, \ldots, K$ indicating K treatment groups, as well as the covariates as regressors in the Cox model.

(2) `time`: This indicates the name of the variable of observed time. Either the time-to-event or the follow-up time is provided in this variable.

(3) `event`: The event indicator takes values 0 or 1. The value is 1 if time-to-event is observed, and it is 0 if the follow-up time is observed.

(4) `group`: The group variable must take values in $\{1, \ldots, K\}$ to distinguish the K treatment groups in the sample. A K-strata Cox model will be constructed which does not require constant hazards ratio between any two treatments.

(5) `covlist`: This parameter specifies the variables that directly enter as regressors in the stratified Cox model. Continuous variables can be directly included. Categorical factors need to be recoded as binary variables. For example, the race factor, containing three levels, White, Black, and Other, has to be recoded as two binary variables. If White is used as reference, one should create the binary variables `black`, which takes value 1 for Black race and 0 otherwise, and `other`, which takes value 1 for other race, 0 otherwise. The variables in the list should be separated by white space. Please note that proportionality tests should be performed beforehand to guarantee all variables pass the tests.

(6) `starttime`: This parameter together with the next one specifies the time interval $[t_1, t_2]$ for which the band will be generated. Determination of t_1 and t_2 should guarantee large enough risk sets as well as a stable comparison between treatments. Subramanian and Zhang (2013) recommended setting t_1 to be slightly greater than the maximum of all first event times from the treatment groups. The comparison of treatments after such a time point has a valid precision estimation. We followed this guideline in implementing the macro under the two-sample framework. The larger value of the first event times from two treatment groups is the default value for t_1. The macro sets t_1 to this value if a smaller value is specified for t_1 by the user.

(7) `stoptime`: This parameter specifies t_2. Subramanian and Zhang (2013) suggested t_2 to be slightly smaller than the minimum of all last event times from the treatment groups. However, such a time may not guarantee a sufficiently large risk set if events are observed for the majority of the study cohort. When analyzing the real data set, Zhang and Klein (2001) chose t_2 so that both treatment groups had at least ten subjects remaining in the study. We believe this is a more practical guideline for determining t_2.

(8) `alpha`: This parameter controls the confidence level for the band of the differences between two direct adjusted survival curves. The significance level α needs to be supplied for a $100(1 - \alpha)\%$ confidence band. For example, one should set α to be 0.05 to produce a 95% confidence band. It should be noted that one may want to set a Bonferroni adjusted α value in case of multiple comparisons. Suppose that there are three treatment groups and one wishes to produce bands for all three pairwise comparisons, then the Bonferroni adjusted α value is 0.0167. It is well-known that Bonferroni adjustment is conservative and easily produces unnecessarily wide confidence intervals or bands. Due to this reason, one should carefully decide on the set of pairwise comparisons. For example, assume that a study involves one control group and three treatment groups. If study interest centers on the three pairwise comparisons between treatments and the control, one can set Bonferroni adjusted α to be 0.0167. Although confidence bands for all pairwise comparisons will be included in the output data set, only the bands for comparisons between treatments and the control shall be fully examined and interpreted.

(9) `nsimu`: This parameter sets the number of simulation realizations that will be generated for determining the confidence band. Bajorunaite and Klein (2007) evaluated the number of realizations in a simulation study and recommended 1000 simulated samples. Their finding was that a larger number does not improve accuracy but prolongs the computational time. A couple of thousands should be sufficient to acquire an accurate band. A size of this level will not invoke a substantial computational time even with a large sample.

(10) `seed`: This parameter has to be a large positive integer and is used as the seed number of the random number generation algorithm. Random number generation routinely utilizes the congruential generator to generate pseudo-random numbers. A seed number has to be supplied as the initial value of the iterative algorithm. When the value zero is supplied, SAS uses the current computer time to create a large positive integer as the seed. As a result, running the program at some other time will lead to a different set of pseudo-random numbers, and a slightly different critical value.

(11) `outdata`: The output data set name is specified by this parameter. The output data set contains the time variable (`time`), the direct adjusted survival estimates (`surv1`, ..., `survK`), the differences between any two adjusted survivals (`diff1_2, diff1_3,`...), the standard error estimates for the adjusted survivals and differences between adjusted survivals(`se1`, ..., `seK`, `se1_2`, `se1_3`, ...), as well as the lower and upper limits of the bands for each pairwise comparison (`diff1_2_low, diff1_2_up,` ...).

We also wrote an R (R Core Team 2016) function that implements the same method as the SAS macro `SURVBAND`. We hope that this R function will offer flexibility to users who prefer to work in the R environment. We posted this function, together with instructions and the example data file, at the following link http://www.umc.edu/faculty/xu_zhang/.

3.2. One time-dependent variable with piecewise proportional hazards

Unlike demographics, patients' clinical characteristics may change over time. In addition, treatment regimen may have to be adjusted due to, for example, exhibition of toxicity symptoms. The covariates which have changing values over time are known as time-dependent

covariates. The distinction between internal and external time-dependent covariates was introduced for convenience when discussing survival function prediction. The aforementioned changing clinical characteristics or treatment regimen belong to the internal category. Survival function prediction with internal time-dependent covariate is subtle because observing the covariate value indicates that the patient is alive. Survival function prediction is normally discussed with external time-dependent covariates, which are characterized by the independence between the covariate processes and the survival status of the subject. For example, longitudinally monitored air pollution level is a typical external time-dependent covariate. It would be practically meaningful to predict the probability of asthma onset for a senior given a process of air pollution level.

One important application of external time-dependent covariates is to evaluate and solve the problem of nonproportionality in the Cox analysis. A commonly used method for testing proportionality of a fixed covariate Z is to test the significance of $Z(t) = Z \times f(t)$. Here, $Z(t)$ is an intentionally created external time-dependent covariate and common choices for $f(t)$ include t and logarithm of t. Significance of $Z(t)$ would indicate violation of proportionality. One method to deal with nonproportionality on Z is to search for a suitable time cut point c so that the hazard ratios in the time interval before and after c are two constants, leading to the piecewise proportional hazards model. In general, more than one cut point on time can be employed but this increases complexity in modeling and interpretation. For the single cut point c, the coding technique for the piecewise proportional hazards is to create two external time-dependent covariates based on Z,

$$Z_1(t) = \left\{ \begin{array}{ll} Z, & t \le c \\ 0, & t > c \end{array} \right. , \qquad\qquad Z_2(t) = \left\{ \begin{array}{ll} 0, & t \le c \\ Z, & t > c \end{array} \right.$$

The next step is to include both $Z_1(t)$ and $Z_2(t)$ in the Cox regressor. This is a useful method of addressing the issue of nonproportionality. Therefore, we developed the second macro of band generation based on the piecewise proportional hazards model in which the effect of one covariate is modeled as two constant hazard ratios. The macro is saved in the file `timeband.sas` with the macro name `TIMEBAND`.

```
%macro TIMEBAND(inputdata, time, event, group, covlist, timecov,
   cutpoint, starttime, stoptime, alpha, nsimu, seed, outdata);
```

This macro contains two additional parameters, `timecov` and `cutpoint`.

(1) `timecov`: The covariate that is modeled as two time-dependent variables and has two constant hazard ratios. It should be noted that this covariate should not be included in `covlist`.

(2) `cutpoint`: The time cut point c used to create the time-dependent variables.

In Section 5.2 the macro is illustrated with a real example of survival data of lymphoma patients.

4. Simulation study

A simple simulation study was conducted to examine the actual performance of the simulation-based confidence band given in Section 2.2. We considered a sample size of 100 and two

Setting	Baseline hazard functions	% of replications where H_0 was rejected
(1)	$\lambda_1(t) = \lambda_2(t) = 0.1$	0.062
(2)	$\lambda_1(t) = 0.1, \lambda_2(t) = 0.15$	0.332
(3)	$\lambda_1(t) = 0.1, \lambda_2(t) = 0.2$	0.665

Table 1: Proportion of replications where the bands did not cover the zero reference line.

treatment groups. The treatment assignment was determined by a Bernoulli random variable with $p = 0.5$ so that the two groups were roughly of equal size. The baseline hazard functions associated with two treatments were constant. A total of three settings were simulated:

(1) $\lambda_1(t) = \lambda_2(t) = 0.1$,

(2) $\lambda_1(t) = 0.1, \lambda_2(t) = 0.15$,

(3) $\lambda_1(t) = 0.1, \lambda_2(t) = 0.2$.

We considered the same covariate setting that contained a standard normal covariate and a discrete covariate with value 0/1. Distribution of the standard normal covariate was independent from treatment assignment while for the discrete covariate value 1 appeared more frequently in treatment 2 group (65%) than treatment 1 group (40%). Censoring variables were generated from exponential distributions with varying hazard rates for different settings to control the censoring rate to be about 30%.

There were 1000 replications in each setting. Since the range of event times varied dramatically across samples, we chose to use a sample-based interval $[t_1, t_2]$ for simulation. In each sample, t_1 was set to be the smallest time when events have occurred in both treatment groups. The largest event time was used as t_2. The 95% confidence band was constructed based on 1000 simulated processes. We examined if the band fully covered the zero reference line in $[t_1, t_2]$. Rejection of the null hypothesis $\bar{S}_1(t) - \bar{S}_2(t) = 0$ was concluded if zero was outside of the band for some subintervals contained in $[t_1, t_2]$. Table 1 depicts the proportion of replications where the null hypothesis was rejected in the three simulation settings.

The results demonstrated good performance of the simulation-based band. The proportion of rejection should be interpreted as the realized power of the test using a 5% significance level. In Setting (1), the proportion of rejection is 0.062, close to the nominal type I error rate 0.05. Reasonable proportions of rejection are observed in Settings (2) and (3) where the two treatment arms are associated with different baseline hazard functions. In addition, as the treatment-specific hazard functions become more different, the observed proportion of rejection increases.

5. Real data example

Fifty-eight patients with primary central nervous system (CNS) lymphoma received chemotherapy for blood-brain barrier disruption (BBBD) at Oregon Health Sciences University from 1982 to 1992. 39 patients received initial chemotherapy for BBBD without prior radiation, while 19 patients were administered cranial radiation prior to the referral for chemotherapy. The survival data of these 58 patients were originally analyzed by Dahlborg, Henner, Crossen, Tableman, Petrillo, Braziel, and Neuwelt (1996). They presented the Kaplan-Meier curves of two treatment groups and evaluated the log-rank statistic. The curves and test result

supported the conclusion that the initial chemotherapy without prior radiation is associated with a significantly lower chance of mortality. Although Dahlborg *et al.* (1996) reported low Karnofsky score and being male as worse prognostic factors, the investigators did not present the adjusted survival curves of the two treatments.

5.1. Data analysis using the first macro

Tableman and Kim (2003) reported the data set and conducted a Cox analysis to assess the treatment effect as well as the effects of a few clinical covariates. In Tableman and Kim (2003), based on a 5% significance level, the final Cox model included the treatment variable, the Karnofsky performance score, sex, dichotomized age split on 60, as well as the variable for interaction between sex and age. Although there was no evidence for violation of proportionality between the two treatments, we still considered the Cox model stratified by treatment group, so that we could capture dynamic changes in the treatment effect over time. Other covariates, Karnofsky performance score, sex, age, interaction between sex and age, consisted of the parametric part of the Cox model.

A text file `cns.txt`, provided along with the paper, contains the data including survival time, vital status, treatment variable and three covariates. Suppose that this text file is saved in the local directory. The following SAS syntax performs the following two steps: First, the data step reads in the external text file and creates a SAS data set. Second, the PHREG procedure fits the aforementioned stratified Cox model.

```
data cns;
infile "cns.txt";
input time death treatment KFS male age60 sexage;
run;

proc PHREG;
model time*death(0)=KFS male age60 sexage /rl;
strata treatment;
run;
```

The SAS data set `cns` has the following variables: (1) `time:` the survival time in years from the first BBBD; (2) `death:` the death event indicator; (3) `treatment:` the treatment group variable, 1 for chemotherapy without prior radiation, 2 for chemotherapy with prior radiation; (4) `KPS:` the Karnofsky performance score ranging from 40 to 100; (5) `male:` the male sex indicator, 1 for male, 0 for female; (6) `age60:` the age variable, 1 if age is at least 60, 0 if age is less than 60; (6) `sexage:` the variable for interaction between sex and age, 1 for male aged 60 or older, 0 otherwise. The PHREG procedure implements the Cox regression, producing the partial-likelihood-based estimates of regression coefficients. The `rl` option in the `model` statement adds the hazard ratio estimates to the output. The hazard ratio estimates, as well as the 95% confidence intervals, can be found in Table 2.

Compared to patients in Group 2, patients in Group 1 have higher mean Karnofsky score (84 vs. 75, two-sample t test $p = 0.058$). The Kaplan-Meier curves by treatment group may not precisely describe the treatment effect on survival because of the difference in the Karnofsky score in the two groups. The direct adjusted survival curves of the two treatments, based on the properly specified regression model, can remove the bias caused by imbalanced

	HR	95% CI	p value
Karnofsky performance score	0.97	0.95–0.99	0.003
Sex (male vs. female)	1.12	0.40–3.15	0.827
Age (≥ 60 vs. < 60)	3.47	1.36–8.88	0.009
Sex & age interaction (male and age ≥ 60 vs. other)	0.11	0.02–0.57	0.008

Table 2: Hazard ratios in the Cox analysis. The baseline hazard rate functions varied by the two treatment groups.

distribution of risk factors. The main issue discussed in this paper is to construct a band for the differences between direct adjusted survivals, so that one can identify when the survival outcomes differ. We decided to generate the band up to 2 years. This upper limit was chosen because only a few events occurred after 2 years and in the two treatment groups 17 and 4 patients remained under study at this time point. The sizes of the treatment groups are reasonable so that the treatment comparison is reliable. Below is the syntax to load the macro and invoke it:

```
%include "band.sas";
%SURVBAND(cns, time, death, treatment, KFS male age60 sexage, 0, 2, 0.05,
    2000, 785243, survout);
```

The variables `time`, `surv1` and `surv2` in the output data set allows one to draw the direct adjusted survivals of the two treatments. One has to load the values into an Excel sheet, the R environment for statistical computing, or an other software package, to plot the curves. Some additional values have to be added to set proper x-axis ranges for the curves. The variable `time` in the output data set starts from the first event time and ends with the largest event time. One should let a curve start from time zero and extend to the largest censoring time of the given treatment group. We depicted the direct adjusted curves of two treatments in Figure 1.

The time-varying treatment effect can be examined from the band of the differences between direct adjusted survivals. The above syntax required simulation of 2000 processes. The critical value for constructing the 95% band was 2.729, which was printed in the output. The output data set contains the variables for the differences (`diff1_2`) and the lower, upper limits of the 95% confidence band (`diff1_2_low`, `diff1_2_up`). Figure 2 presents the differences, the 95% confidence interval as well as the 95% confidence band. Note that the the lower limit of the time interval was automatically adjusted to 0.06, though 0 was specified as the lower limit. The macro automatically checks the lower limit of the time interval and sets it to the larger value of two first event times of the two treatment groups if a smaller value is specified. Therefore, the band for the CNS lymphoma example was produced for the interval $[0.06, 2.00]$. In Figure 2, one can observe a monotone increase of the difference in survival probability, between treatment without radiation and treatment with radiation. The lower limit of the band rises beyond the zero reference line about at the time point 1.4 years, which confirms a long-term survival benefit for chemotherapy without prior radiation.

5.2. Data analysis using the second macro

We further used this data example to illustrate the second macro developed for the Cox model containing time-dependent variables. We modeled `KPS` as time-dependent variables with 1.5
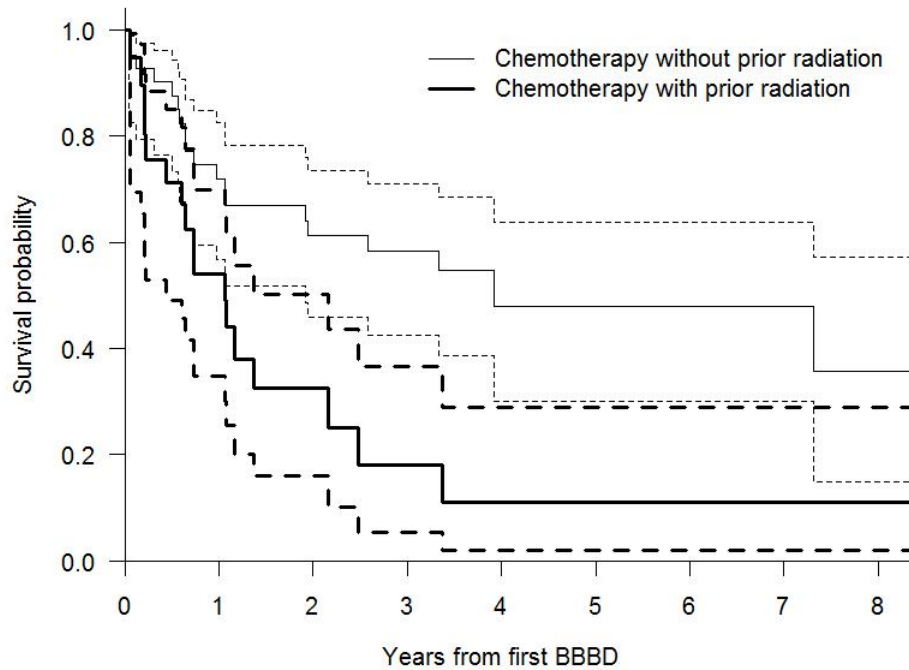
Figure 1: Direct adjusted survival curves and 95% log-log transformed confidence intervals for chemotherapy without prior radiation and chemotherapy with prior radiation.
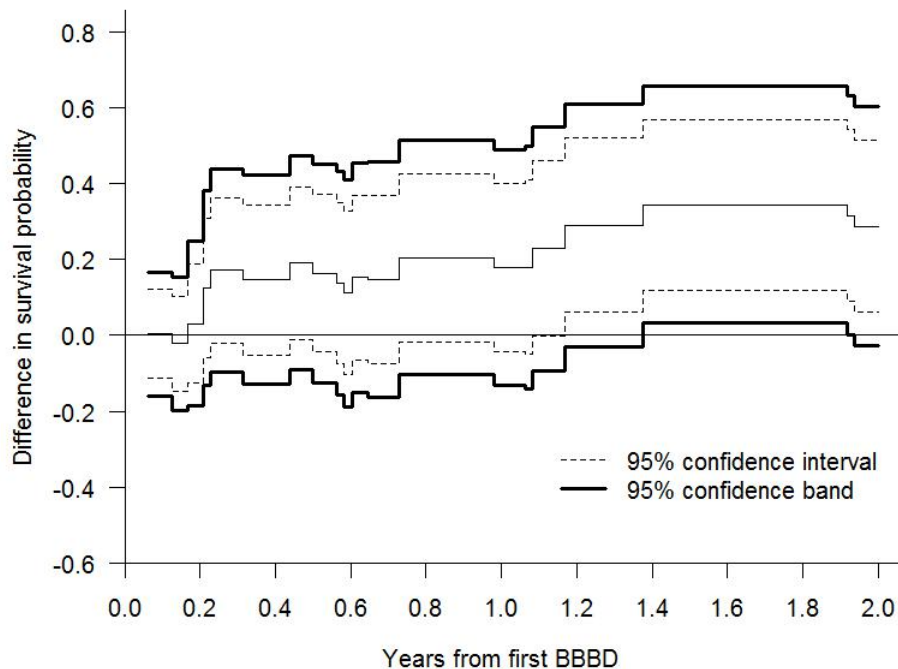

Figure 2: The 95% confidence band for the differences in direct adjusted survival between chemotherapy without prior radiation and chemotherapy with prior radiation. 2000 realizations of the process were generated and yielded the critical value 2.729 for the band.
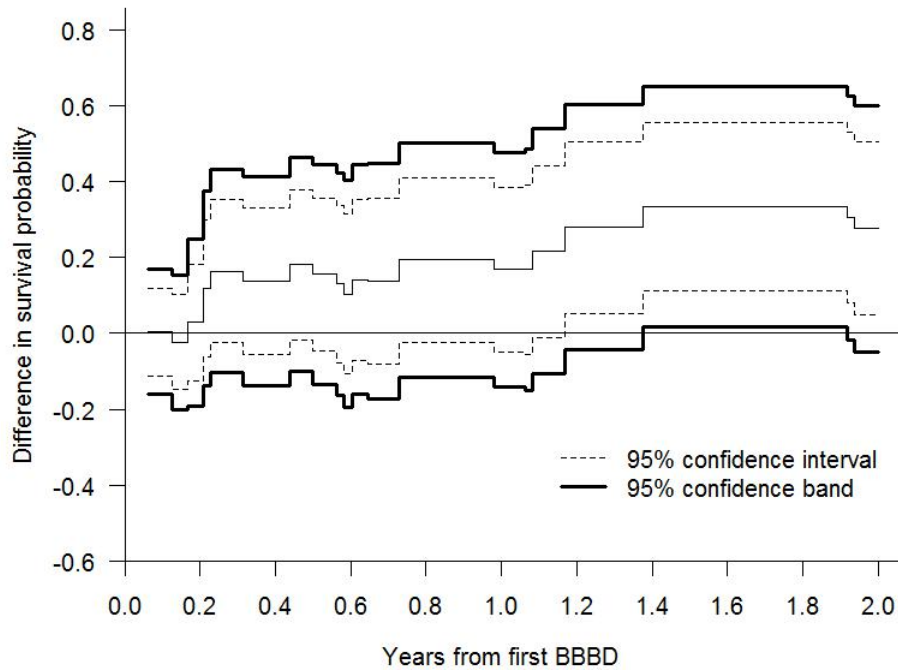
Figure 3: The 95% confidence band for the differences in direct adjusted survival between chemotherapy without prior radiation and chemotherapy with prior radiation. The covariate `KPS` was modeled as a time-dependent variable with a time cut point at 1.5 years. The critical value 2.792 is based on 2000 realizations of the process containing time-dependent variables.

years as the cut point. Note that this variable passed the proportionality test but was modeled to have piecewise proportional hazards for illustrative purpose only. More specifically, two time-dependent variables were created based on `KPS`,

$$\text{KPS1}(t) = \left\{ \begin{array}{ll} \text{KPS,} & t \leq 1.5 \\ 0, & t > 1.5 \end{array} \right. , \qquad \text{KPS2}(t) = \left\{ \begin{array}{ll} 0, & t \leq 1.5 \\ \text{KPS,} & t > 1.5 \end{array} \right. .$$

The following statements should be written in the `PHREG` procedure to create the variables `KPS1` and `KPS2`,

```
proc PHREG;
model time*death(0)=KPS1 KPS2 male age60 sexage /rl;
KPS1=0;
if time<=1.5 then KPS1=KPS;
KPS2=0;
if time>1.5 then KPS2=KPS;
strata treatment;
run;
```

This model produced the hazard ratios 0.96 (95% CI 0.94–0.99) and 0.99 (95% CI 0.94–1.06) for `KPS1` and `KPS2`, respectively. These two hazard ratios are the effect of `KPS` prior to and

after 1.5 years, respectively. The estimated effects of other covariates remained almost the same as the estimation result from the Cox model treating `KPS` as fixed. The macro `TIMEBAND` can produce the direct adjusted survival of treatment based on the above Cox model, as well as the band for differences between two direct adjusted survival functions. For this example, write the following statements to call the macro,

```
%include "timeband.sas";
%TIMEBAND(cns, time, death, treatment, male age60 sexage, KPS, 1.5, 0,
   2, 0.05, 2000, 487012, survout);
```

The critical value yielded from running this macro was 2.792, slightly larger than that in the model with only fixed covariates. This is reasonable because usage of two variables for `KPS` increased the variability in the estimated survival when the proportionality does hold for different values in `KPS` over the whole time range. The band yielded from the macro is depicted in Figure 3. The similar pattern is observed, i.e., that the band starts to be beyond the zero reference line at about 1.4 years.

6. Final discussions

The `SAS` macros presented in this paper can be used to compute the confidence bands for differences between direct adjusted survival curves based on stratified Cox models. The bands provide information regarding at what times the survival outcomes of two treatments differ. The macros are quite efficient. The run time can be expected to be less than one minute to generate 2000 realizations even for a large sample. The simulation method can be extended to compare the differences of direct adjusted survival curves based on other models, such as Aalen's additive model (Aalen 1989). The direct adjustment method can be utilized to compute a quantity for a general reference population. Our macro averages the survivals of a treatment across all subjects in a sample, which is the routine reference population. Laubender and Bender (2010) discussed usage of a subgroup of a sample as a reference population, so that one can estimate the beneficial effect when patients receiving inferior treatment would switch to the better treatment.

When there are more than two treatment arms, the Bonferroni method can be used to adjust for multiple pairwise comparisons. The Bonferroni method is conservative because it ignores the correlation between tests. The sharper method that accounts for correlation among tests was considered by Logan, Wang, and Zhang (2005). Define $W_{ij}(t)$ as the difference between the direct adjusted survivals of treatment i and treatment j, and form a vector $\boldsymbol{W} = (W_{12}, \ldots, W_{K-1,K})$. The vector $\boldsymbol{W}$ would converge to a multivariate normal process with mean 0 and some covariance matrix. To control the familywise error, one may set the critical value as the $1 - \alpha$ percentile of the distribution of $\max |W_{ij}(t)|$. One can use the wild bootstrap method to obtain realizations of the vector and then find the appropriate percentile. This method is not implemented in the macros because we focus on the two-sample problem in this paper. The reader can revise the source code to implement this method.

References

Aalen O (1989). "A Linear Regression Model for the Analysis of Life Time." *Statistics in Medicine*, **8**, 907–925. [doi:10.1002/sim.4780080803](https://doi.org/10.1002/sim.4780080803).

Andersen P, Gill R (1982). "Cox's Regression Model for Counting Processes: A Large Sample Study." *The Annals of Statistics*, **10**, 1100–1120. `doi:10.1214/aos/1176345976`.

Bajorunaite R, Klein J (2007). "Two-Sample Tests of the Equality of Two Cumulative Incidence Functions." *Computational Statistics & Data Analysis*, **51**, 4269–4281. `doi:10.1016/j.csda.2006.05.011`.

Beyersmann JM, Termini S, Pauly J (2013). "Weak Convergence of the Wild Bootstrap for the Aalen-Johansen Estimator of the Cumulative Incidence Function of a Competing Risk." *The Scandinavian Journal of Statistics*, **78**, 387–402. `doi:10.1111/j.1467-9469.2012.00817.x`.

Dahlborg S, Henner W, Crossen J, Tableman M, Petrillo A, Braziel R, Neuwelt E (1996). "Non-AIDS Primary CNS Lymphoma: The First Example of a Durable Response in a Primary Brain Tumor Using Enhanced Chemotherapy Delivery without Cognitive Losss and without Radiotherapy." *The Cancer Journal from Scientific American*, **2**, 166–174.

Gail M, Byar D (1986). "Variance Calculations for Direct Adjusted Survival Curves, with Applications to Testing for No Treatment Effect." *Biometrical Journal*, **28**, 587–599. `doi:10.1002/bimj.4710280508`.

Laubender R, Bender R (2010). "Estimating Adjusted Risk Difference (RD) and Number Needed to Treat (NNT) Measures in the Cox Regression Model." *Statistics in Medicine*, **29**, 851–859. `doi:10.1002/sim.3793`.

Lin D, Fleming T, Wei J (1994). "Confidence Bands for Survival Curves Under the Proportional Hazards Model." *Biometrika*, **81**, 73–81. `doi:10.2307/2337051`.

Logan B, Wang H, Zhang M (2005). "Pairwise Multiple Comparison Adjustment in Survival Analysis." *Statistics in Medicine*, **24**, 2509–2523. `doi:10.1002/sim.2125`.

MacKenzie T, Brown J, Likosky D, Wu Y, Grunkemeier G (2012). "Review of Case-Mix Corrected Survival Curves." *The Annals of Thoracic Surgery*, **93**, 1416–1425. `doi:10.1016/j.athoracsur.2011.12.094`.

Makuch R (1982). "Adjusted Survival Curve Estimation Using Covariates." *Journal of Chronic Diseases*, **35**, 437–443. `doi:10.1016/0021-9681(82)90058-3`.

Nieto F, Coresh J (1996). "Adjusting Survival Curves for Confounders: A Review and a New Method." *American Journal of Epidemiology*, **143**, 1059–1068. `doi:10.1093/oxfordjournals.aje.a008670`.

R Core Team (2016). *R: A Language and Enviroment for Statistical Computing*. Vienna, Austria. URL `https://www.R-project.org/`.

SAS Institute Inc (2008). *SAS/IML 9.2 User's Guide*. Cary, NC. URL `http://www.sas.com/`.

SAS Institute Inc (2009). *SAS/STAT 9.2 User's Guide, 2nd Edition*. Cary, NC. URL `http://www.sas.com/`.

Subramanian S, Zhang P (2013). "Model-Based Confidence Bands for Survival Functions." *Journal of Statistical Planning and Inference*, **143**, 1166–1185. `doi:10.1016/j.jspi.2013.01.012`.

Tableman M, Kim J (2003). *Survival Analysis Using S: Analysis of Time-to-Event Data.* Chapman & Hall/CRC.

Wu C (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis." *The Annals of Statistics*, **14**, 1261–1295. doi:10.1214/aos/1176350142.

Zhang M, Klein J (2001). "Confidence Bands for the Difference of Two Survival Curves under Proportional Hazards Model." *Lifetime Data Analysis*, **7**, 243–254.

Zhang X, Loberiza F, Klein J, Zhang M (2007). "A SAS Macro for Estimation of Direct Adjusted Survival Curves Based on a Stratified Cox Regression Model." *Computer Methods and Programs in Biomedicine*, **88**, 95–101. doi:10.1016/j.cmpb.2007.07.010.

Affiliation:

Hong Wang
Department of Biostatistics, University of Pittsburgh
Biostatistics Facility, University of Pittsburgh Cancer Institute
201 North Craig Street
Pittsburgh, PA 15213, United States of America

Xu Zhang
Center of Biostatistics & Bioinformatics
Cancer Institute
University of Mississippi Medical Center
2500 North State Street
Jackson, MS 39216, United States of America
E-mail: xzhang2@umc.edu