STATISTICS IN MEDICINE
*Statist. Med.* 2005; **24**:301–318
Published online 23 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1888

# Non-parametric paired two-sample tests for censored survival data incorporating longitudinal covariate information

Shari Messinger[1,§,¶] and Susan Murray[2,*,†,‡]

[1]*Department of Epidemiology and Public Health, University of Miami School of Medicine, 1801 NW 9th Avenue 3rd floor, Miami, FL 33136, U.S.A.*
[2]*Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.*

SUMMARY

In this manuscript, we present non-parametric two-sample tests for paired censored survival data incorporating longitudinal covariate information. These tests take advantage of information collected at baseline and post-baseline to provide efficiency gains when censoring is uninformative. Additionally, these methods adjust for potential bias from informative censoring that is captured by the baseline and longitudinal covariates. Finite sample properties are investigated with simulation, and we illustrate methodology with an example from the Early Treatment Diabetic Retinopathy Study. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: correlated survival; longitudinal covariates; Kaplan–Meier estimate; Pepe–Fleming test; selection bias

## 1. INTRODUCTION

Most clinical trials collect prognostic baseline information in order to investigate whether treatment imbalances have occurred despite randomization and with the added hope of gaining efficiency from this prognostic information in estimation and testing of treatment effects. In addition, it is becoming more common for clinical trials to collect longitudinal covariates as supplemental outcome information that is relevant to the primary clinical endpoint of interest. For example, the Early Treatment Diabetic Retinopathy Study [1, 2] collected eye-specific baseline retinopathy status classified according to two levels of prognostic disease severity and later, visual acuity scores in a study assessing early versus delayed photocoagulation for

*Correspondence to: Susan Murray, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A.
†E-mail: skmurray@umich.edu
‡ScD.
§E-mail: smessinger@med.miami.edu
¶PhD.


*Received May 2003*
*Accepted March 2004*

the prevention of severe vision loss in patients with diabetic retinopathy in both eyes. In this paired design alternate eyes were randomized to the two competing treatments, minimizing variability in comparing time to severe vision loss.

Since severe vision loss is in part defined in terms of deteriorating visual acuity score, visual acuity information obtained after baseline gives partial information towards treatment-mediated events occurring post-baseline. Hence, if used in a way that bolsters information regarding event-times, baseline information and follow up visual acuity scores should allow for more efficient estimation of survival, especially under relatively heavy censoring as in the ETDRS. At the same time, in the two-sample testing framework, one must be cautious to avoid methods that inappropriately adjust for internal covariates observed post-baseline as described by Kalbfleisch and Prentice [3]. Adjusting a treatment effect for internal covariates observed post-baseline may diminish legitimate treatment differences, since these covariates carry information about the treatment effect.

The particular setting of the ETDRS could potentially benefit from use of estimation and testing procedures for paired, censored survival data structures that incorporate this prognostic longitudinal covariate information. However, methods that accommodate this data structure are not currently available for two-sample testing.

Model-based approaches relevant to two sample testing have been developed for paired, censored survival data that condition upon baseline covariate information and allow adjustment for external time-dependent covariates that do not mediate treatment effects. Frailty models, discussed by many authors including Clayton [4], Vaupel *et al.* [5], Hougaard *et al.* [6] and Oakes [7] handle dependence in survival endpoints by incorporating random effects shared by dependent pair members into a baseline survival model. Spiekerman and Lin [8] use working independence models in the derivation of parameter estimates and account for the correlation structure in determining standard errors. These model-based methods for paired censored survival data are able to adjust for selection bias reflected in baseline covariate imbalances and are able to give correct inference about a treatment effect when modelling assumptions are valid. However, these methods are not designed to make effective use of internal longitudinal covariate information since inference about the treatment effect conditional on covariates included in these parametric models partially adjusts the treatment effect away.

Instead of conditioning upon the covariates as in the analyses described above, one may incorporate covariates in estimating overall marginal survival for each treatment group. In the setting of independent treatment groups, Murray and Tsiatis [9] presented non-parametric two-sample tests comparing marginal survival distributions supplemented by stratified longitudinal covariate information, and demonstrated efficiency gains that grew as more longitudinal information was incorporated over time in the case censoring remained uninformative. Under informative censoring, these tests remain consistent when the relationship between the censoring and failure times is captured by the covariate information. The survival estimates used in their tests have been shown by Murray and Tsiatis [9] to correspond to inverse probability weighted complete case survival estimates discussed in Robins and Rotnitzky [10] and in Robins and Finkelstein [11] when inverse weights are non-parametrically estimated. Similar survival estimates have been developed by Malani [12] using a redistribute to the right algorithm and incorporated into rank-based tests. Although these tests can take advantage of longitudinal covariate information, they do not accommodate paired designs.

Non-parametric methods have been presented for analysing paired censored survival data using a variety of sign- or rank-based tests by authors such as Woolson and Lachenbruch [13],

 

Wei [14], O'Brien and Fleming [15], Dabrowska [16, 17], and Jung [18]. Murray [19, 20] developed paired extensions of both sequentially monitored weighted logrank tests discussed by Gill [21], and tests of integrated survival differences discussed by Pepe and Fleming [22]. In her thesis, Messinger [23] developed non-parametric paired two-sample tests comparing marginal survival adjusting for baseline covariate imbalances in a causal inference manner. Although these tests accommodate paired designs, they are not able to take advantage of prognostic covariate information collected post baseline.

In this manuscript, we develop non-parametric two-sample tests for censored survival data that accommodate dependent treatment groups while incorporating longitudinal covariate information with three goals in mind. These tests extend to the paired setting (1) the ability to adjust for potential bias from informative censoring that is captured by the longitudinal covariates, (2) non-parametric baseline adjustment for treatment imbalances and, under uninformative censoring, (3) efficiency gains as additional prognostic time-dependent covariate information is incorporated without inappropriate alteration of the treatment effect from using these potentially internal covariates. To our knowledge, no single method accomplishing these goals is available in making marginal survival comparisons in the paired censored survival setting. In Section 2, we describe baseline covariate standardized estimation of marginal survival incorporating longitudinal data. In Section 3, we present corresponding two-sample tests for paired, censored survival data. These tests use longitudinal covariate information in a manner that does not diminish treatment effects mediated through the covariates, but instead uses the information to improve inference. We investigate finite sample power and size properties of the proposed tests in Section 4. In Section 5, we revisit the ETDRS study. Discussion follows in Section 6.

## 2. SURVIVAL ESTIMATION INCORPORATING LONGITUDINAL COVARIATES AND BASELINE STANDARDIZATION

Let $T_g$ and $C_g$ denote failure and censoring random variables and $X_g = \min(T_g, C_g)$ the possibly censored event time with corresponding censoring indicator $\Delta_g = I(T_g < C_g), g = 1, 2$. Suppose that for each of $n_g$ subjects in group $g$, a time-dependent categorical covariate is potentially measured at each of times $T_0^*, \ldots, T_s^*$ during the study period. When time-dependent covariates mediate treatment effects they may be viewed as intermediate marker outcomes. For subjects in group $g$, let $Z_{gm}$ denote a time-dependent covariate observed at time $T_{m-1}^*, m = 1, 2, 3, \ldots, s+1$, with realization $i_m, i_m = 0, \ldots, k$. Also let $\theta_{gi_1}$ be the probability that $Z_{g1} = i_1$ at $T_0^*$ for a subject in group $g$, and let $\theta_{gi_1 \ldots i_m}$ be the probability that a subject in group $g$ has $Z_{gm} = i_m$ at $T_{m-1}^*$, conditional on the subject surviving at least to time $T_{m-1}^*$ and previously having $Z_{g1} = i_1$ at $T_0^*, Z_{g2} = i_2$ at $T_1^*, \ldots$, and $Z_{gm-1} = i_{m-1}$ at $T_{m-2}^*$. Also, for a subject in group $g$, let $S_{gi_1 \ldots i_m}(t)$ be the probability that a subject survives past time $t$, conditional on the subject surviving past time $T_{m-1}^*$ and having $Z_1 = i_1$ at $T_0^*, Z_2 = i_2$ at $T_1^*, \ldots$, and $Z_m = i_m$ at $T_{m-1}^*$, and let $H_{gi_1 \ldots i_m}(t)$ and $\lambda_{gi_1 \ldots i_m}(t)$, respectively, denote the corresponding censoring survival and hazard functions. Using conditional probability, we may then express marginal survival for this heterogeneous distribution at times $t, T_{m-1}^* < t \leqslant T_m^*$ as

$$S_g(t) = \sum_{i_1=0}^{k} \sum_{i_2=0}^{k} \cdots \sum_{i_m=0}^{k} S_{gi_1 i_2 \ldots i_m}(t) \theta_{gi_1 i_2 \ldots i_m} \times \cdots \times S_{gi_1 i_2}(T_2^*) \theta_{gi_1 i_2} S_{gi_1}(T_1^*) \theta_{gi_1}$$

 

Let $n_{gi_1\ldots i_m}$ be the number of people in group $g$ having $Z_{g1}=i_1$ at $T_0^*, Z_{g2}=i_2$ at $T_1^*,\ldots,$ and $Z_{gm}=i_m$ at time $T_{m-1}^*$, let $n_{gi_1\ldots i_{m-1}}=n_{gi_1\ldots i_{m-1}0}+\cdots+n_{i_1\ldots i_{m-1}k}$ represent the number at risk in group $g$ at $T_{m-1}^*$ having previous covariate history $i_1,\ldots,i_{m-1}$, and let $\hat{S}_{gi_1\ldots i_m}(t)$ be the Kaplan–Meier survival estimate at time $t$ among those in group $g$ who were at risk at time $T_{m-1}^*$ with past covariate values corresponding to $i_1,\ldots,i_m, m=1,2,\ldots,s+1$. An estimate for survival in group $g$ at times $t, T_{m-1}^*<t\leqslant T_m^*$ is

$$WS_g(t)=\sum_{i_1=0}^{k}\sum_{i_2=0}^{k}\cdots\sum_{i_m=0}^{k}\hat{S}_{gi_1i_2\ldots i_m}(t)\hat{\theta}_{gi_1i_2\ldots i_m}\times\cdots\times\hat{S}_{gi_1i_2}(T_2^*)\hat{\theta}_{gi_1i_2}\hat{S}_{gi_1}(T_1^*)\hat{\theta}_{gi_1}$$

where $\hat{\theta}_{gi_1}=n_{gi_1}/n_g$ represents the group specific proportion of patients in each covariate stratum as seen in each treatment group population at baseline and, for $m\neq 1, \hat{\theta}_{gi_1i_2\ldots i_m}=n_{gi_1i_2\ldots i_m}/n_{gi_1i_2\ldots i_{m-1}}$ are group specific proportions. This estimate for survival and corresponding variance results have been previously described by Murray and Tsiatis [9, 24].

An alternative estimate for survival at times, $T_{m-1}^*<t\leqslant T_m^*$, that we propose is

$$\widetilde{WS}_g(t)=\sum_{i_1=0}^{k}\sum_{i_2=0}^{k}\cdots\sum_{i_m=0}^{k}\hat{S}_{gi_1i_2\ldots i_m}(t)\hat{\theta}_{gi_1i_2\ldots i_m}\times\cdots\times\hat{S}_{gi_1i_2}(T_2^*)\hat{\theta}_{gi_1i_2}\hat{S}_{gi_1}(T_1^*)\hat{\theta}_{i_1}$$

where $n_{.i_1}=n_{1i_1}+n_{2i_1}, n=n_1+n_2$, and $\hat{\theta}_{i_1}=n_{.i_1}/n$ represents the overall proportion of patients in covariate stratum $i$ at baseline so it is common across treatment groups $g=1,2$. This estimate of survival adopts an approach commonly used in lifetable analysis to adjust for selection bias, where a common standardized distribution of a confounding factor is used to estimate mortality rates in each of two groups under comparison. Under successful randomization schemes, using a standardized baseline covariate distribution common to both groups in calculating $\widetilde{WS}_g, g=1,2$, provides efficiency gains over the unstandardized estimate, $WS_g(t)$ since more data is used to estimate the common $\theta_{gi}=\theta_i$. When baseline covariate imbalances exist, use of a standardized baseline covariate distribution to estimate marginal survival in a two-sample testing framework evens out baseline covariate imbalances between the groups so that differences in marginal survival observed across groups are due only to differences in the strata-specific survival distributions themselves, and not attributable to bias caused by baseline covariate disparities. Hence, use of the unstandardized survival estimate, $WS_g(t)$ $g=1,2$, in a paired testing framework would accomplish goals (1) and (3) as stated in the Introduction, whereas use of the standardized estimate, $\widetilde{WS}_g(t)$ $g=1,2$ would accomplish goals (1), (2) and (3).

In describing $\widetilde{WS}_g(t)$ we have assumed, without loss of generality, that $T_0^*$ is zero so that the first set of covariates are measured at baseline. However, this estimation strategy may also accommodate settings described by Murray and Tsiatis [9, 24] in which the first covariate incorporated is measured post-baseline, and hence potentially altered by treatment, by using an artificially created baseline covariate that is identical for all patients in each treatment group. By incorporating artificial baseline covariates in this way, the first covariates observed at $T_1^*$ remain unstandardized in $\widetilde{WS}_g(t)$ and the resulting estimate corresponds to work of Murray and Tsiatis [9] where baseline imbalances could be an issue in making inferences.

 

## 3. DEVELOPMENT OF CORRESPONDING TWO-SAMPLE TESTS FOR PAIRED DATA STRUCTURES

Using standardized survival estimates $\widetilde{WS}_g(t)$ that incorporate covariate information collected at $T_0^*, T_1^*, \ldots, T_k^*$, we can now consider a test statistic for paired censored survival data of the form $T_{PSM_k} = (n_1 n_2/n)^{1/2} \int_0^\infty \hat{w}(t)\{\widetilde{WS}_1(t) - \widetilde{WS}_2(t)\}\,\mathrm{d}t$, where the subscript $k$ on the test statistic matches the $k$ subscript on the last time covariate information was collected and incorporated into the survival estimation procedure, $T_k^*$. This test statistic is powered to detect alternatives of the form $\int_0^\infty w(t)S_1(t)\,\mathrm{d}t \neq \int_0^\infty w(t)S_2(t)\,\mathrm{d}t$, which is the same set of alternatives considered by the original Pepe–Fleming statistic. The random weight function, $\hat{w}(t)$, must be chosen so that it vanishes whenever the number at risk within any group and stratum is zero and $\sup_{u\in[0,t)} |\hat{w}(u) - w(u)|$ approaches zero in probability for deterministic $w(u)$. For instance, if we define $Y_{gi_1,\ldots,i_{k+1}}(t)$ as the number of individuals at risk at time $t$ for covariate history $i_1,\ldots,i_{k+1}$ in group $g, \hat{w}(t) = \prod_{g,i_1,\ldots,i_{k+1}} I(Y_{gi_1,\ldots,i_{k+1}}(t) > 0)$ will result in an interpretation corresponding to the average years of life saved (YLS) between treatment groups during the period from $(0, \max(t): \hat{w}(t) > 0)$. The limiting function $w(u)$ associated with this choice of weight is an indicator function with positive value over the time frame where patients corresponding to each group and strata profile have positive probability of remaining at risk.

In calculating the asymptotic variance of this two-sample test in the paired setting, we must fully appreciate and account for correlation between survival and marker endpoints longitudinally across treatment groups. The required methodological derivation of the asymptotic test variance is included in Appendix A along with estimates for the each of its components.

Under successful randomization, it can be shown that the asymptotic variance of $T_{PSM_K}$ is smaller than that of $T_{PM_K} = (n_1 n_2/n)^{1/2} \int_0^\infty w(t)\{WS_1(t) - WS_2(t)\}\,\mathrm{d}t$, the corresponding test we also develop in this manuscript using unstandardized $WS$ survival estimators in place of $\widetilde{WS}$, under the null hypothesis as well as under many interesting alternatives where we have uninformative censoring. For instance such alternatives includes all cases with $\pi_1 = \pi_2 = 1/2$, where $\pi_g$ is the probability of being assigned to treatment group $g$. Having complete pairs is a special case where $\pi_1 = \pi_2$. Another example is the set of alternatives where $A_{2i_1}(0) = A_{1i_1}(0) + C$, where $A_{gi_1}(0) = \int_0^\infty w(t)S_{gi_1}(t)\,\mathrm{d}t$, and $C$ can be interpreted as a constant, perhaps weighted, YLS due to treatment for each stratum. The null hypothesis of $C = 0$ is a special case of this set of alternatives. Additionally, it can be shown that efficiency gains using the standardized survival estimators are achieved under all alternatives when treatment groups are independent. Each test statistic developed for paired censored survival data in this manuscript exceeds by far the efficiency of methods currently available for independent treatment groups although we recommend use of $T_{PSM_K}$. A study of gains made from use of longitudinal data follows in the next section.

## 4. SIMULATIONS

Simulations relating to an 11 year study period with baseline and 4 year dichotomous marker information were constructed to investigate finite sample power and size properties with independent and dependent treatment groups. Each treatment group consists of 4 possible survival profiles constructed from piecewise bivariate lognormal failure times with a 4-year changepoint

 

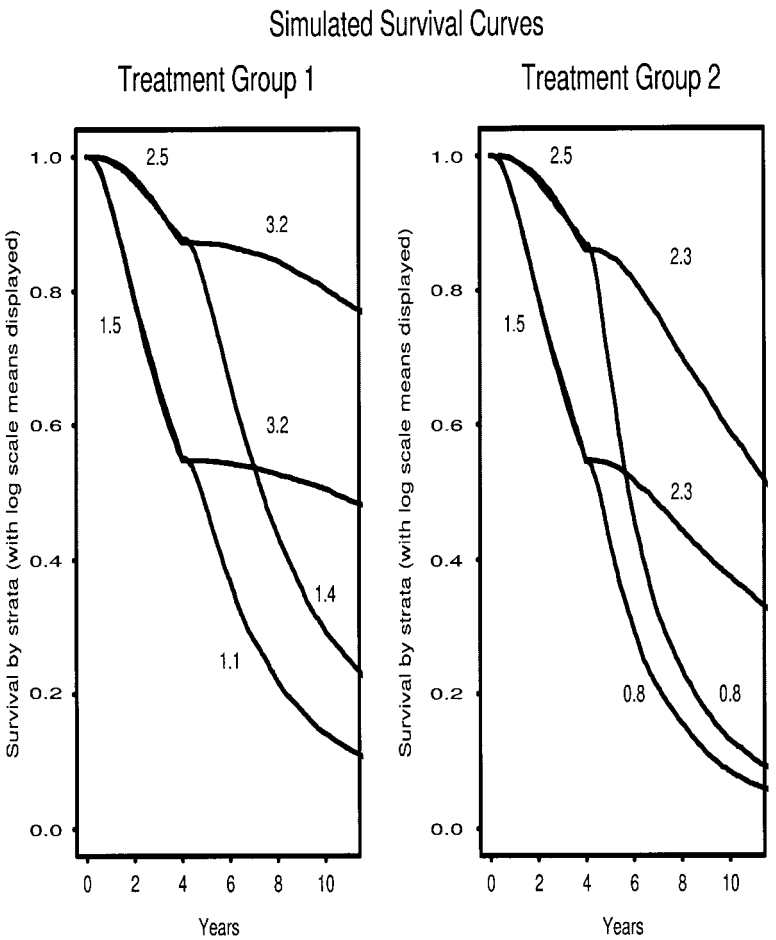


Figure 1. The piecewise bivariate lognormal survival curves displayed above correspond to the scenario underlying results in Tables I and II, where the average difference in years lived over the first 4 years due to baseline stratum is approximately 6 months in each treatment group. Average differences in years lived after 4 years due to covariate stratum determined at $T_1^*$ for each treatment group depending on baseline strata 1 and 2 were approximately 2.5 and 1.5 years, respectively. Log-scale means are displayed in proximity to the curves with a 4-year change point.

as illustrated in Figure 1. Correlation, or the lack thereof, comes from two different sources, the log-scale correlation parameter used in generating the piecewise bivariate lognormal curves and the degree of correlation imposed upon the longitudinal covariates. In simulations conducted with independent treatment groups, the log-scale correlation parameter, $\rho$, was taken to be zero and the covariates were generated independently across treatment groups. In the correlated setting, $\rho = 0.3$ and pair members shared the same covariate profiles, unless there was induced selection bias in which case covariates were generated independently.

Parameter values corresponding to the null and alternative hypotheses, with and without selection bias, are outlined in the following paragraphs along with the results of each simulation.

 

Prognostic ability of the covariates in the various simulations are described using the average difference in years lived between covariate stratum survival profiles over a specific time period. To target the degree of censoring during the 11 year study period to be 20 per cent and encourage identical study periods across all simulations, we modelled the censoring mechanism acting on the endpoints using the random variable $C = L \times I(B = 1) + 11 \times I(B = 0)$. In this expression, $L$ is a bivariate lognormal random variable with mean and variance on the log scale equal to 2.5 and correlation equal to that of the corresponding failure times, and $B$ is a Bernoulli random variable with success probability related to the desired level of censoring. For each investigation, 1000 Monte Carlo simulations were run with 300 either uncorrelated or correlated failure time pairs.

To verify size under the null hypothesis, the log-scale means and variances of the baseline failure time distributions for each treatment group were $(2.5, 1)$ and $(1.5, 1)$ for covariate strata 1 and 2, respectively, corresponding to an average difference in years lived over the first 4 years of approximately 6 months. The log-scale means and variances of the conditional failure time distributions applied to those at risk at $T_1^*$ for each baseline covariate stratum were $(2.3, 1), (0.8, 1)$ for covariate strata 1 and 2, respectively. These parameters affecting survival beyond $T_1^*$ correspond to an average difference in years lived over the remaining study period between covariate stratum determined at $T_1^*$ of approximately 2.5 and 1.5 years according to baseline stratum 1 and 2, respectively.

Selection bias scenarios displayed for the null hypothesis assume $\theta_{11} = \theta_{22} = 0.55$ and $\theta_{12} = \theta_{21} = 0.45$. Otherwise, in the absence of selection bias, equal proportions were designed to fall into the different baseline strata. Covariates generated at $T_1^*$ were designed to fall into the different strata with equal proportions.

Size results are located in Table I. From top to bottom, the rows in Table I correspond to results from the paired test statistics incorporating covariate information from 2 looks ($T_{PSM_1}$), baseline only ($T_{PSM_0}$), and ignoring all covariate information ($T_P$). From left to right, the columns of the table represent results with independent treatment groups and then correlated treatment groups without and with selection bias. Appropriate size results were observed for paired, standardized tests using either baseline information alone or both baseline and longitudinal information in each scenario. The test statistic that ignores covariate information, $T_P$, had inflated type I errors for the selection bias setting.

To study power, the log-scale means and variances of the four failure time distribution profiles in each treatment group were generated to study alternatives over a range of prognostic ability of the baseline covariate. Baseline covariate prognostic ability corresponded to an average difference in years lived between strata over the first 4 years of approximately $0, 6,$

Table I. Size under null hypothesis.

| Test | Independent | Paired | Paired with selection bias |
|---|---|---|---|
| $T_{PSM_1}$ | 0.048 | 0.058 | 0.052 |
| $T_{PSM_0}$ | 0.052 | 0.043 | 0.049 |
| $T_P$ | 0.048 | 0.042 | 0.109 |

Average differences in years lived after 4 years between covariate stratum determined at $T_1^*$ for each treatment group depending on baseline strata 1 and 2 were approximately 2.5 and 1.5 years, respectively.

 

Table II. Power under alternative hypothesis*.

| Test | Independent | Paired | Paired with mild selection bias |
|---|---|---|---|
| $T_{PSM_1}$ | 0.608 | 0.717 | 0.666 |
| $T_{PSM_0}$ | 0.589 | 0.664 | 0.643 |
| $T_P$ | 0.569 | 0.619 | 0.264 |

*Average difference in years lived over the first 4 years due to baseline stratum is approximately 6 months in each treatment group. Average differences in years lived after 4 years between covariate stratum determined at $T_1^*$ for each treatment group depending on baseline strata 1 and 2 were approximately 2.5 and 1.5 years, respectively.

or 9 months. When simulating no average difference in years lived over the first 4 years according to baseline stratum, log-scale means and variances of the baseline failure time distributions were $(2,1)$ for both covariate strata 1 and 2 and treatment groups 1 and 2. For an average difference in years lived of 6 months over the first 4 years due to baseline covariate stratum, these parameters become $(2.5,1)$ and $(1.5,1)$ for covariate strata 1 and 2, respectively. For an average difference in years lived of 9 months over the first 4 years due to the baseline covariate stratum, these parameters become $(2.8,1)$ and $(1.2,1)$ for covariate strata 1 and 2, respectively. The log-scale means and variances of the conditional failure time distributions applied to those at risk at $T_1^*$ depends on treatment group $g$, and covariate history $i_1 i_2$. For those in treatment group 1, these parameters were $(3.2,1),(1.4,1),(3.2,1)$, and $(1.1,1)$, respectively, corresponding to covariate histories $11,12,21,22$. For treatment group 2, these parameters were $(2.3,1)(0.8,1),(2.3,1)(0.8,1)$, respectively, corresponding to covariate histories $11,12,21,22$. These parameters affecting survival beyond $T_1^*$ correspond to an average difference in years lived over the remaining study period between covariate stratum determined at $T_1^*$ of approximately 2.5 and 1.5 years according to baseline stratum 1 and 2, respectively, in the case shown in Figure 1. Further information on prognostic ability for the various simulation results are footnoted in the appropriate power tables.

In the absence of selection bias, equal proportions were designed to fall into the different baseline strata. A range of selection bias scenarios were investigated under the alternative hypotheses. Parameter values corresponding to mild selection bias were $\theta_{11}=\theta_{22}=0.55$ and $\theta_{12}=\theta_{21}=0.45$. These parameters become $\theta_{11}=\theta_{22}=0.60$ and $\theta_{12}=\theta_{21}=0.40$ for alternatives with moderate selection bias, and become $\theta_{11}=\theta_{22}=0.65$ and $\theta_{12}=\theta_{21}=0.35$ for an even greater degree of selection bias. Covariates generated at $T_1^*$ were designed to fall into the different strata with equal proportions.

Power results located in Table II show power increasing under independent and paired settings both with and without mild selection bias with each additional incorporated covariate look. In this selection bias setting, the test statistic that ignores covariate information, $T_P$, is too conservative while the paired standardized tests using either baseline covariate information alone or baseline and longitudinal information adjust for potential bias.

Power results located in Table III further illustrate the performance of the test statistics over a range of selection bias and prognostic ability of the baseline covariate. The left column of the table indicates the average difference in years lived between covariate strata over

 

Table III. Power with varying degrees of baseline covariate prognostic ability and selection bias.

| Baseline prognostic ability (YLS over 4 years) | Test | $\theta_{11} = \theta_{22} = 0.55$ $\theta_{12} = \theta_{21} = 0.45$ | $\theta_{11} = \theta_{22} = 0.60$ $\theta_{12} = \theta_{21} = 0.40$ | $\theta_{11} = \theta_{22} = 0.65$ $\theta_{12} = \theta_{21} = 0.35$ |
|---|---|---|---|---|
| 0* | $T_{PSM_1}$ | 0.693 | 0.702 | 0.662 |
| | $T_{PSM_0}$ | 0.626 | 0.642 | 0.621 |
| | $T_P$ | 0.633 | 0.636 | 0.613 |
| 6 months† | $T_{PSM_1}$ | 0.666 | 0.723 | 0.707 |
| | $T_{PSM_0}$ | 0.643 | 0.680 | 0.653 |
| | $T_P$ | 0.264 | 0.092 | 0.047 |
| 9 months† | $T_{PSM_1}$ | 0.695 | 0.693 | 0.672 |
| | $T_{PSM_0}$ | 0.678 | 0.673 | 0.661 |
| | $T_P$ | 0.162 | 0.081 | 0.306‡ |

*Average difference in years lived after 4 years in each treatment group and baseline covariate stratum between covariate stratum determined at $T_1^*$ is approximately 2.5 years.

†Average differences in years lived after 4 years between to covariate stratum determined at $T_1^*$ for each treatment group depending on baseline strata 1 and 2 were approximately 2.5 and 1.5 years, respectively.

‡All but one of these were rejections of the null hypothesis in favour of the inferior treatment. This is worse than having zero power since the wrong treatment is recommended in most cases.

the first 4 years of the study period. The next column indicates the test statistic for which the results are displayed in the following three columns with increasing degrees of baseline imbalance. In cases where the baseline covariate is prognostic over the first 4 years, results show power increasing with each additional incorporated covariate look as in Table II. Comparable power results for $T_P$ and $T_{PSM_0}$ are displayed for all levels of selection bias when the baseline covariate is not prognostic over the first 4 years with power gains only apparent for $T_{PSM_1}$ in these cases. Increasing selection bias only modestly affects the power of $T_{PSM_0}$ and $T_{PSM_1}$ as sample size increases for estimating survival curves and $\theta_{ggi_2}$ parameters associated with $Z_{11}$ and $Z_{22}$, and sample size decreases for estimating survival curves and $\theta_{g(3-g)i_2}$ parameters associated with $Z_{12}$ and $Z_{21}$. However increasing selection bias has a more profound affect upon the power of $T_P$, which offers no adjustment for bias due to an imbalanced and prognostic baseline covariate. This impact of increased selection bias on $T_P$ is greater when baseline covariates have greater prognostic ability. In the worst case, illustrated in the bottom right corner of the table, the power is even worse in consequence than 0 per cent since rejections of the null hypothesis are most often in favour of the inferior treatment.

We additionally investigated power of the test statistics when the effect of treatment manifested itself through changes in the covariate marker after baseline. Parameters to generate data were identical to those for paired data under the null hypothesis, with the exception of the parameters corresponding to the post baseline covariate distribution. In this scenario, covariates generated at $T_1^*$ were designed to fall into covariate strata 1 and 2 with probability 0.65 and 0.35, respectively, for treatment group 1 and with probability 0.35 and 0.65, respectively, for treatment group 2 regardless of baseline status. Results are displayed in Table IV for mild selection bias or no selection bias. Again we see increased power with each additional incorporated covariate look. Although treatment differences appear only after $T_1^*$, $T_{PSM_0}$

 

Table IV. Power when treatment effect is captured by post baseline covariate.

| Test | No selection bias | Mild selection bias |
|---|---|---|
| $T_{PSM_1}$ | 0.672 | 0.684 |
| $T_{PSM_0}$ | 0.647 | 0.656 |
| $T_P$ | 0.608 | 0.311 |

- Covariates generated at $T_1^*$ were designed to fall into covariate strata 1 and 2 with probability 0.65 and 0.35, respectively, for treatment group 1 and with probability 0.35 and 0.65, respectively, for treatment group 2 regardless of baseline status.
- Average differences in years lived after 4 years between to covariate stratum determined at $T_1^*$ for each treatment group depending on baseline strata 1 and 2 were approximately 2.5 and 1.5 years, respectively.

is still able to make gains in power through a coarsened capturing of the treatment effect. In the selection bias setting presented, the test statistic that ignores covariate information, $T_P$, is even more conservative than it was with no selection bias present.

## 5. EXAMPLE

Recall the Early Treatment Diabetic Retinopathy Study (ETDRS) described in the Introduction enrolling 3711 patients each having some degree of diabetic retinopathy in both eyes. For each patient, one eye was randomized to a treatment group receiving early photocoagulation therapy and the other to a treatment group deferring photocoagulation therapy until a time when high-risk proliferative retinopathy was detected. The clinical endpoint of interest was time to severe vision loss, where this loss was defined as visual acuity less than $\frac{5}{200}$ at two consecutive visits. The statistics discussed in this paper test the difference in average days of sight between treatment groups observed over a study period of about 8 years.

The ETDRS collected eye-specific baseline retinopathy status classified according to two levels of prognostic disease severity. Level of observed baseline retinopathy severity in this group is moderately predictive survival and, to a lesser degree, censoring times. For each level of baseline retinopathy, follow-up visual acuity at 3 years, defined as low or high visual acuity with respect to the median, is significantly predictive of survival conditional on having survived to that point as can be seen in Figure 2. After inspection, the strata specific conditional censoring survival distributions are similar amongst those at risk at 3 years. In this setting, we expect to make gains by incorporating both the baseline and 3-year prognostic covariate information.

Estimates for the average extended days of sight over the 8-year period in the early photocoagulation group along with the associated test variances are displayed in Table V. All test methods considered give similar estimates of average extended days of sight in the early photocoagulation group of about 41 days. Both tests incorporating prognostic covariate information in the paired setting perform better than the $T_P$, the paired Pepe–Fleming test ignoring

 

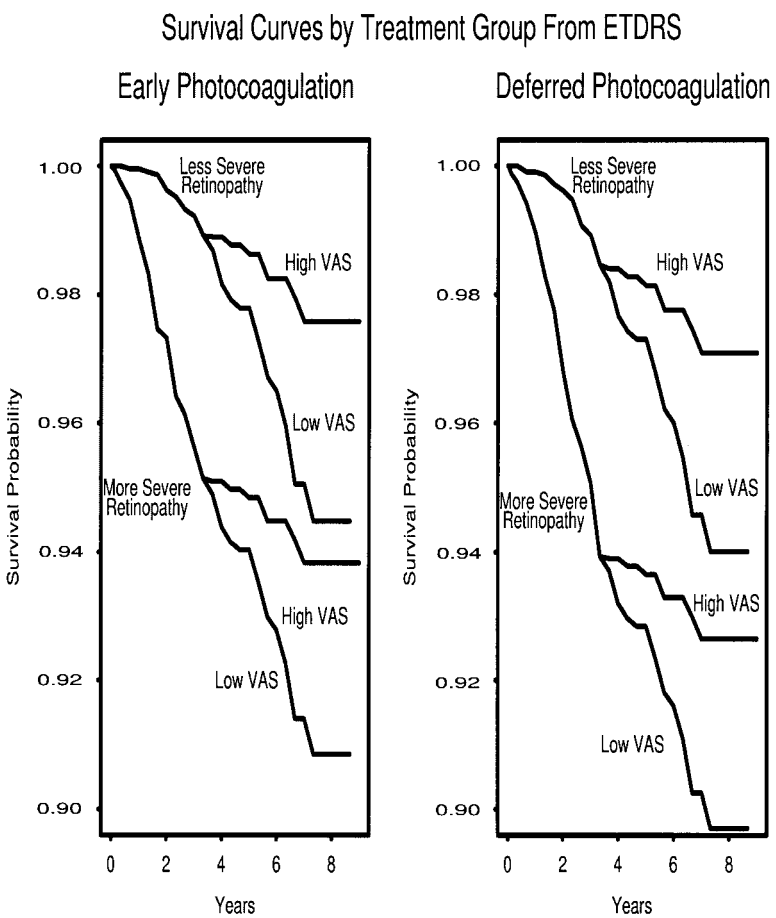


Figure 2. Survival probabilities by baseline retinopathy status and 3-year visual acuity score (VAS) in early and deferred photocoagulation groups of the ETDRS.

Table V. Results from ETDRS.

| Test | Average extended days of sight | Associated variance | 95 per cent CI |
|---|---|---|---|
| $T_{PSM_1}$ | 41.86 | 65.88 | (25.95,57.77) |
| $T_{PSM_0}$ | 41.23 | 85.83 | (23.07,59.39) |
| $T_P$ | 41.45 | 107.68 | (21.11,61.79) |

$T_{PSM_1}$ is the paired test developed in this work using standardized baseline covariate information and other longitudinal covariates. $T_{PSM_0}$ is the paired test using only standardized baseline covariate information. $T_P$ is the paired test ignoring covariate information.

covariate information. $T_{PSM_1}$, incorporating visual acuity at 3 years in addition to baseline retinopathy, leads to a 23 per cent reduction in the estimated variance of the treatment difference over $T_{PSM_0}$, which uses only baseline retinopathy status.

 

## 6. DISCUSSION

This work develops methodology for studying paired censored survival data that makes use of baseline and internal longitudinal covariate information in a manner that improves statistical inference regarding treatment effects. These paired tests extend the ability to adjust for potential bias from informative censoring that is captured by the longitudinal covariates, adjust for differential selection, and can take advantage of covariate information to make efficiency gains in the absence of these sources of bias. We have demonstrated that additional efficiency gains are achieved with each prognostic covariate look under uninformative censoring in the paired censored survival setting. The potential for gaining information in the paired censored survival setting through incorporation of additional covariate information exceeds that seen when treatment groups are independent since, in the paired setting, covariate information collected longitudinally for one treatment group gives partial information about the opposing treatment group endpoint. Simulations observed in Section 4 show a tendency toward larger increases in power in the paired setting that encourage this intuition. This observation alone is a new contribution to the literature that inspires the need for more research with auxiliary covariate information in the paired censored survival setting.

When treatment groups are comparable at baseline, baseline covariate standardization advocated by this research provides efficiency gains due to additional information being used to estimate marginal survival. It should be underscored that standardization methods are only appropriate when applied to baseline covariates and not subsequent covariate information that potentially mediates treatment effects. Standardizing post-baseline covariate distributions across treatment groups when estimating survival would potentially mask the treatment effect of interest.

In her thesis, Messinger [23] developed stratified tests for dependent treatment groups using weighted integrated survival based methods. In addition to adjusting for potential bias from baseline covariate imbalances and allowing the censoring distributions to vary by baseline stratum, incorporating baseline covariate information through stratification also yields efficiency gains when treatment alternatives are more clearly detectable within each covariate stratum. There may also be advantages to stratified tests in cases where there are great disparities in the follow-up periods of the various baseline covariate strata defined groups, since $T_{PSM}$ does not take advantage of differences beyond the end of follow-up in the least followed stratum. However, stratified tests are unable to exploit the additional covariate information collected after baseline for purposes of either efficiency gain or informative censoring adjustment. Hence in cases where prognostic longitudinal information is available, $T_{PSM}$ gives more advantages than the paired stratified tests developed by Messinger. In addition, $T_{PSM}$ maintains the philosophy of comparing overall marginal survival differences in terms of years of life saved that is appealing when working with non-statistically minded investigators.

## APPENDIX A

In calculating the asymptotic variance of $T_{PSM_K}$, the underlying correlation structure in both the survival times and longitudinal covariate information between the treatment groups must be accounted for. In what follows, we outline the strategy for calculating this asymptotic variance, with further details available from the authors upon request.

 

Recall $T_{PSM_k} = (n_1 n_2/n)^{1/2} \int_0^\infty \hat{w}(t)\{\widetilde{WS}_1(t) - \widetilde{WS}_2(t)\}\,\mathrm{d}t$, which is asymptotically equivalent in distribution to $(\pi_1\pi_2)^{1/2}\sqrt{n}\int_0^\infty w(t)\{\widetilde{WS}_1(t) - \widetilde{WS}_2(t)\}\,\mathrm{d}t$, where $\pi_g$ is the overall probability of falling into treatment group $g$. Hence,

$$\begin{aligned}\mathrm{Var}(T_{PSM_k}) &\approx \pi_1\pi_2\,\mathrm{Var}\left[\sqrt{n}\int_0^\infty w(t)\{\widetilde{WS}_1(t) - \widetilde{WS}_2(t)\}\,\mathrm{d}t\right] \\ &= \pi_1\pi_2\left[\sum_{g=1}^{2}\mathrm{Var}\left\{\sqrt{n}\int_0^\infty w(t)\widetilde{WS}_g(t)\,\mathrm{d}t\right\}\right. \\ &\quad \left. -2\,\mathrm{Cov}\left\{\sqrt{n}\int_0^\infty w(t)\widetilde{WS}_1(t)\,\mathrm{d}t, \sqrt{n}\int_0^\infty w(t)\widetilde{WS}_2(t)\,\mathrm{d}t\right\}\right]\end{aligned}$$

We first consider

$$\begin{aligned}\sum_{g=1}^{2}\mathrm{Var}\left\{\sqrt{n}\int_0^\infty w(t)\widetilde{WS}_g(t)\,\mathrm{d}t\right\} = \sum_{g=1}^{2}\mathrm{Var}\left\{\sqrt{n}\int_0^{T_1^*} w(t)\widetilde{WS}_g(t)\,\mathrm{d}t\right. \\ \left. +\sqrt{n}\int_{T_1^*}^\infty w(t)\widetilde{WS}_g(t)\,\mathrm{d}t\right\}\end{aligned}$$

To calculate this term, we use conditioning arguments with respect to $\mathscr{F}_{T_1^*-Z_2}$, representing the survival, censoring and covariate information up until and including $T_1^*$ with the exception of the value of $Z_2$. At this point, care must be taken so that $\widetilde{WS}_g(t)$ is defined appropriately in each time interval. Let $S_{gi_1\cdot}(t) = P(T > t | T > T_1^*, Z_1 = i_1) = \sum_{i_2=0}^{k}\theta_{i_1 i_2}S_{gi_1 i_2}(t)$. Then, for $T_0^* < t \leqslant T_1^*$, $\widetilde{WS}_g(t)$ becomes $\sum_{i_1=0}^{k}\hat{\theta}_{i_1}\hat{S}_{gi_1}(t)$, and for $T_1^* < t$, $\widetilde{WS}_g(t)$ becomes $\sum_{i_1=0}^{k}\hat{\theta}_{i_1}\hat{S}_{gi_1}(T_1^*)WS_{gi_1\cdot}(t)$, where $WS_{gi_1\cdot}(t) = \sum_{i_2=0}^{k}\hat{\theta}_{gi_1 i_2}WS_{gi_1 i_2}(t)$ corresponds to an unstandardized estimate of $S_{gi_1\cdot}(t)$. Applying conditioning arguments, $\sum_{g=1}^{2}\mathrm{Var}\{\sqrt{n}\int_0^\infty w(t)\widetilde{WS}_g(t)\,\mathrm{d}t\}$ becomes,

$$\begin{aligned}&\sum_{g=1}^{2}\mathrm{Var}\left\{\sum_{i_1=0}^{k}\frac{n_{\cdot i_1}}{\sqrt{n}}\int_0^{T_1^*} w(t)\hat{S}_{gi_1}(t)\,\mathrm{d}t + \sum_{i_1=0}^{k}\frac{n_{\cdot i_1}}{\sqrt{n}}\hat{S}_{gi_1}(T_1^*)\int_{T_1^*}^\infty w(t)S_{gi_1\cdot}(t)\,\mathrm{d}t\right\} \\ &+\sum_{g=1}^{2}E\left[\mathrm{Var}\left\{\sqrt{n}\int_{T_1^*}^\infty w(t)\widetilde{WS}_g(t)\,\mathrm{d}t | \mathscr{F}_{T_1^*-Z_2}\right\}\right] \qquad \text{(A1)}\end{aligned}$$

After some further calculation following from equation (A1), this asymptotic closed form variance becomes

$$\sum_{g=1}^{2}\left(\sum_{h=1}^{2}\pi_h\sum_{i_1=0}^{k}\theta_{hi_1}\{A_{gi_1}(0) - \bar{A}_{hg}(0)\}^2 \qquad \text{(A2)}$$

$$+\sum_{i_1=0}^{k}\sum_{j_1=0}^{k}\Gamma_{i_1 j_1}A_{gi_1}(0)A_{gj_1}(0) - \Gamma\bar{A}_{1g}(0)\bar{A}_{2g}(0) \qquad \text{(A3)}$$

 

$$+\sum_{i_1=0}^{k}(\theta_{i_1}/\pi_{gi_1})\left[\int_0^{T_1^*}\{S_{gi_1}(t)H_{gi_1}(t)\}^{-1}\{A_{gi_1}(t)\}^2\lambda_{gi_1}(t)\,\mathrm{d}t \tag{A4}$$

$$+S_{gi_1}(T_1^*)H_{gi_1}^{-1}(T_1^*)\mathrm{Var}\left\{\sqrt{n_{gi_1\cdot}}\int_{T_1^*}^{\infty}w(t)WS_{gi_1\cdot}(t)\,\mathrm{d}t|\mathscr{F}_{T_1^*-Z_2}\right\}\Bigg]\Bigg) \tag{A5}$$

where $A_{gi_1}(x)=\int_x^{\infty}w(t)S_{gi_1}(t)\,\mathrm{d}t$, $\bar{A}_{hg}(0)=\sum_{i_1=0}^{k}\theta_{hi_1}A_{gi_1}(0)$ for $g,h=1,2$ with $\bar{A}_{hg}(0)=\bar{A}_g(0)$ for $h=g$ and $\pi_{gi_1}$ is the probability of falling into treatment group $g$ for individuals in baseline strata $i_1$. Also, define $\tilde{n}_{i_1j_1}$ as the number of pair members in baseline stratum $i_1$ of group 1 with corresponding counterparts in baseline stratum $j_1$ of group 2, $\Gamma_{i_1j_1}=E(2\tilde{n}_{i_1j_1}/n)$, and $\Gamma=\sum_{i_1=0}^{k}\sum_{j_1=0}^{k}\Gamma_{i_1j_1}$. A result by Murray and Tsiatis [9] for $\mathrm{Var}\left\{\sqrt{n_{gi_1\cdot}}\int_{T_1^*}^{\infty}w(t)WS_{gi_1\cdot}(t)\,\mathrm{d}t|\mathscr{F}_{T_1^*-Z_2}\right\}$ gives,

$$\begin{aligned}&\mathrm{Var}\left\{\sqrt{n_{gi_1\cdot}}\int_{T_1^*}^{\infty}w(t)WS_{gi_1\cdot}(t)\,\mathrm{d}t|\mathscr{F}_{T_1^*-Z_2}\right\}\\&\approx\sum_{\zeta=2}^{s+1}\sum_{i_\zeta=0}^{k}\frac{S_{gi_1i_2}(T_2^*)}{H_{gi_1i_2}(T_2^*)}\theta_{gi_1i_2}\cdots\sum_{i_\zeta=0}^{k}\theta_{gi_1\ldots i_\zeta}\\&\quad\times\left\{\int_{u=T_{\zeta-1}^*}^{T_\zeta^*}\frac{A_{gi_1\ldots i_\zeta}^2(u)\lambda_{gi_1\ldots i_\zeta}(u)}{H_{gi_1\ldots i_\zeta}(u)S_{gi_1\ldots i_\zeta}(u)}+D_{gi_1\ldots i_\zeta}^2(T_{\zeta-1}^*)\right\}\end{aligned}$$

where $A_{gi_1\ldots i_\zeta}(x)=\int_x^{\infty}w(t)S_{gi_1\ldots i_\zeta}(t)\,\mathrm{d}t$, $\bar{A}_{gi_1\ldots i_{\zeta-1}}(x)=\sum_{i_\zeta=0}^{k}\theta_{gi_1\ldots i_\zeta}A_{gi_1\ldots i_\zeta}(x)$, and $D_{gi_1\ldots i_\zeta}(x)=A_{gi_1\ldots i_\zeta}(x)-\bar{A}_{gi_1\ldots i_{\zeta-1}}(x)$.

To derive the final term in the variance of $T_{PSM_k}$, $2\,\mathrm{Cov}\{\sqrt{n}\int_0^{\infty}w(t)\widetilde{WS}_1(t)\,\mathrm{d}t,\sqrt{n}\int_0^{\infty}w(t)\widetilde{WS}_2(t)\mathrm{d}t\}$, we again use conditioning arguments. Initially we condition only on baseline covariate information, and then calculate subsequent covariance terms by conditioning with respect to $\mathscr{F}_{T_1^*-Z_2}$. Using this strategy, we first address terms relating to covariability in the baseline covariates, and then terms relating to covariability in the corresponding survival estimates.

So,

$$\begin{aligned}&-2\,\mathrm{Cov}\left\{\sqrt{n}\int_0^{\infty}w(t)\widetilde{WS}_1(t)\,\mathrm{d}t,\sqrt{n}\int_0^{\infty}w(t)\widetilde{WS}_2(t)\,\mathrm{d}t\right\}\\&\approx-2\,\mathrm{Cov}\left\{\sum_{i_1=0}^{k}\frac{n_{\cdot i_1}}{\sqrt{n}}\int_0^{\infty}w(t)S_{1i_1}(t),\sum_{j_1=0}^{k}\frac{n_{\cdot j_1}}{\sqrt{n}}\int_0^{\infty}w(t)S_{2j_1}(t)\right\}\\&\quad-2E\left[\mathrm{Cov}\left\{\sqrt{n}\int_0^{\infty}w(t)\widetilde{WS}_1(t),\sqrt{n}\int_0^{\infty}w(t)\widetilde{WS}_2(t)|\mathscr{F}_{T_0}^*\right\}\right]\end{aligned}$$

 

$$= -2\sum_{i_1=0}^{k}\sum_{j_1=0}^{k} n^{-1}A_{1i_1}(0)A_{2j_1}(0)\mathrm{Cov}\{n_{\cdot i_1}, n_{\cdot j_1}\}$$

$$-2E\left[\sum_{i_1=0}^{k}\sum_{j_1=0}^{k}\mathrm{Cov}\left\{\frac{n_{\cdot i_1}}{\sqrt{n}}\int_0^\infty w(t)WS_{1i_1}(t), \frac{n_{\cdot j_1}}{\sqrt{n}}\int_0^\infty w(t)WS_{2j_1}(t)|\mathscr{F}_{T_0}^*\right\}\right] \quad \text{(A6)}$$

where $\mathrm{Cov}\{n_{\cdot i_1}, n_{\cdot j_1}\} = \mathrm{Cov}\{n_{1i_1}, n_{1j_1}\} + \mathrm{Cov}\{n_{2i_1}, n_{2j_1}\} + \mathrm{Cov}\{n_{1i_1}, n_{2j_1}\} + \mathrm{Cov}\{n_{2i_1}, n_{1j_1}\}$. This can be evaluated by recognizing that $\mathrm{Cov}\{n_{gi_1}, n_{gj_1}\}$ is a multinomial covariance with sample size $n_g$, and $\mathrm{Cov}\{n_{1i_1}, n_{2j_1}\}$ becomes $n_p(\tilde{\theta}_{i_1j_1} - \theta_{1i_1}\theta_{2j_1})$, where $n_p$ is the total number of pairs, $\tilde{\theta}_{i_1j_1} = \Gamma_{i_1j_1}/\Gamma$ represents the proportion of pairs where the group 1 member is in baseline stratum $i_1$ and the group 2 member is in baseline stratum $j_1$ relative to the total number of pairs. Therefore, $-2\sum_{i_1=0}^{k}\sum_{j_1=0}^{k} n^{-1}A_{1i_1}(0)A_{2j_1}(0)\mathrm{Cov}\{n_{\cdot i_1}, n_{\cdot j_1}\}$ becomes

$$-\left\{\sum_{i_1=0}^{k}\sum_{j_1=0}^{k} A_{1i_1}(0)A_{2j_1}(0)(\Gamma_{i_1j_1} + \Gamma_{j_1i_1}) - \Gamma(\bar{A}_1(0)\bar{A}_2(0) + \bar{A}_{12}(0)\bar{A}_{21}(0))\right\} \quad \text{(A7)}$$

$$-2\pi_1\left\{\sum_{i_1=0}^{k}\theta_{1i_1}A_{1i_1}(0)A_{2i_1}(0) - \bar{A}_1(0)\bar{A}_{12}(0)\right\} \quad \text{(A8)}$$

$$-2\pi_2\left\{\sum_{i_1=0}^{k}\theta_{2i_1}A_{1i_1}(0)A_{2i_1}(0) - \bar{A}_2(0)\bar{A}_{21}(0)\right\} \quad \text{(A9)}$$

Combining expressions (A3) and (A7) we have

$$\sum_{i_1=0}^{k}\sum_{j_1=0}^{k}\Gamma_{i_1j_1}\{A_{1i_1}(0)A_{1j_1}(0) + A_{2i_1}(0)A_{2j_1}(0) - A_{1i_1}(0)A_{2j_1}(0) - A_{2i_1}(0)A_{1j_1}(0)\}$$

$$-\Gamma\{\bar{A}_1(0)\bar{A}_{21}(0) + \bar{A}_{12}(0)\bar{A}_2(0) - \bar{A}_1(0)\bar{A}_2(0) - \bar{A}_{12}(0)\bar{A}_{21}(0)\}$$

which reduces to

$$\sum_{i_1=0}^{k}\sum_{j_1=0}^{k} PF_{i_1}PF_{j_1}(\Gamma_{i_1j_1} - \Gamma\theta_{1i_1}\theta_{2j_1}) \quad \text{(A10)}$$

where $PF_{i_1} = \int_0^\infty w(t)\{S_{1i_1}(t) - S_{2i_1}(t)\}\,\mathrm{d}t = A_{1i_1}(0) - A_{2i_1}(0)$.

We can also combine expressions (A2), (A8), (A9) giving us,

$$\pi_1\sum_{i_1=0}^{k}\theta_{1i_1}\{(A_{1i_1}(0) - \bar{A}_1(0))^2 + (A_{2i_1}(0) - \bar{A}_{12}(0))^2\}$$

$$+\pi_2\sum_{i_1=0}^{k}\theta_{2i_1}\{(A_{1i_1}(0) - \bar{A}_{21}(0))^2 + (A_{2i_1}(0) - \bar{A}_2(0))^2\}$$

$$-2\pi_1\left(\sum_{i_1=0}^{k}\theta_{1i_1}A_{1i_1}(0)A_{2i_1}(0) - \bar{A}_1(0)\bar{A}_{12}(0)\right) - 2\pi_2\left(\sum_{i_1=0}^{k}\theta_{2i_1}A_{1i_1}(0)A_{2i_1}(0) - \bar{A}_{21}(0)\bar{A}_2(0)\right)$$

 

which reduces to

$$\sum_{g=1}^{2} \pi_g \sum_{i_1=0}^{k} \theta_{gi_1} (PF_{i_1} - \overline{PF_g})^2 \tag{A11}$$

where $\overline{PF_g} = \sum_{i_1=1}^{k} \theta_{gi_1} PF_{i_1}$.

Define $n^*_{i_1 j_1} = n_{1i_1} n_{2j_1} / \tilde{n}_{i_1 j_1}$. Then combining expressions (A4)–(A6), (A10) and (A11), and noting that $n_{\cdot i_1} n_{\cdot j_1} / n n^*_{i_1 j_1} \xrightarrow{P} \Gamma_{i_1 j_1} / 2(\pi_{1i} \pi_{2j})^{-1/2}$, the asymptotic variance for the standardized test statistic in the paired setting incorporating time-dependent covariates becomes

$$\begin{aligned}
&\mathrm{Var}(T_{PSM_k}) \\
&= \pi_1 \pi_2 \left( \sum_{g=1}^{2} \pi_g \sum_{i_1=0}^{k} \theta_{gi_1} (PF_{i_1} - \overline{PF_g})^2 + \sum_{i_1=0}^{k} \sum_{j_1=0}^{k} PF_{i_1} PF_{j_1} (\Gamma_{i_1 j_1} - \Gamma \theta_{1i_1} \theta_{2j_1}) \right. \\
&\quad + \sum_{g=1}^{2} \sum_{i_1=0}^{k} \frac{\theta_{i_1}}{\pi_{gi_1}} \left[ \int_0^{T_1^*} \frac{\{A_{gi_1}(t)\}^2 \lambda_{gi_1}(t)\, \mathrm{d}t}{S_{gi_1}(t) H_{gi_1}(t)} + \frac{S_{gi_1}(T_1^*)}{H_{gi_1}(T_1^*)} \right. \\
&\times \left. \left. \mathrm{Var} \left\{ n_{gi_1 \cdot}^{1/2} \int_{T_1^*}^{\infty} w(t) W S_{gi_1 \cdot}(t)\, \mathrm{d}t \,\middle|\, \mathscr{F}_{T_1^* - Z_2} \right\} \right] \right. \\
&- \sum_{i_1=0}^{k} \sum_{j_1=0}^{k} \Gamma_{i_1 j_1} (\pi_{1i_2} \pi_{2j_1})^{-1} \\
&\times \left. E \left[ \mathrm{Cov} \left\{ \sqrt{n^*_{i_1 j_1}} \int_0^{\infty} w(t) W S_{1i_1}(t), \sqrt{n^*_{i_1 j_1}} \int_0^{\infty} w(t) W S_{2j_1}(t) | \mathscr{F}^*_{T_0} \right\} \right] \right)
\end{aligned} \tag{A12}$$

We have left to evaluate $\mathrm{Cov}\{\sqrt{n^*_{i_1 j_1}} \int_0^{\infty} w(t) W S_{1i_1}(t)\, \mathrm{d}t, \sqrt{n^*_{i_1 j_1}} \int_0^{\infty} w(t) W S_{2j_1}(t)\, \mathrm{d}t | \mathscr{F}^*_{T_0}\}$, which involves terms from correlated survival estimates after removing covariability related to baseline covariates. Before presenting the asymptotic results for this remaining covariance term, we must first define some additional notation. Let $T_{gi_1,\ldots,i_m}$ and $C_{gi_1,\ldots,i_m}$, respectively, denote random variables corresponding to the conditional survival and censoring survival functions $S_{gi_1 \ldots i_m}(t)$ and $H_{gi_1 \ldots i_m}(t)$ with $X_{gi_1,\ldots,i_m} = \min(T_{gi_1,\ldots,i_m}, C_{gi_1,\ldots,i_m})$. Let $S_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)$, $H_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)$, and $\lambda_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)$ represent the bivariate survival, censoring survival, and hazard functions where the group 1 pair member was at risk at time $T^*_{m-1}$ with covariate history corresponding to $i_1,\ldots,i_m$ and the group 2 pair member was at risk at time $T^*_{m-1}$ with covariate history corresponding to $j_1,\ldots,j_m$. Also let $\lambda_{g_1 i_1,\ldots,i_m | g_2 j_1,\ldots,j_m}(u|v)$ represent the associated conditional hazard where the pair member from group $g_1$ has covariate history $i_1,\ldots,i_m$, and the pair member from group $g_2$ has covariate history $j_1,\ldots,j_m$. Using this notation, define $G_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v) = P(X_{1i_1,\ldots,i_m} \geqslant u, X_{2j_1,\ldots,j_m} \geqslant v)\{P(X_{1i_1,\ldots,i_m} \geqslant u) P(X_{2j_1,\ldots,j_m} \geqslant v)\}^{-1} \{\lambda_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v) - \lambda_{1i_1,\ldots,i_m | 2j_1,\ldots,j_m}(u|v) \lambda_{2j_1,\ldots,j_m}(v) - \lambda_{2j_1,\ldots,j_m | 1i_1,\ldots,i_m}(v|u) \lambda_{1i_1,\ldots,i_m}(u) + \lambda_{1i_1,\ldots,i_m}(u) \lambda_{2j_1,\ldots,j_m}(v)\}$. Let $\tilde{n}_{i_1,\ldots,i_m,j_1,\ldots,j_m}$ represent the number of pairs with covariate history $i_1,\ldots i_m$, $j_1,\ldots,j_m$ at $T^*_{m-1}$ and let $\tilde{n}_{i_1,\ldots,i_{m-1},j_1,\ldots,j_{m-1}}$ represent the number of pairs with covariate history $i_1,\ldots,i_{m-1},j_1,\ldots,j_{m-1}$

 

where both members were at risk at $T^*_{m-1}$, and let $\tilde{\theta}_{i_1,\ldots,i_m,j_1,\ldots,j_m} = E(\tilde{n}_{i_1,\ldots,i_m,j_1,\ldots,j_m}/\tilde{n}_{i_1,\ldots,i_{m-1},j_1,\ldots,j_{m-1}})$. Then, after some further evaluation, expression A12 becomes,

$$
\begin{aligned}
&\mathrm{Var}(T_{PSM_k}) \\
&= \sum_{i_1=0}^{k}\sum_{j_1=0}^{k}(\pi_{1i_1}\pi_{2j_1})^{-1}\Gamma_{i_1j_1}C(0) \\
&\quad + \sum_{l=1}^{s}\sum_{i_1,\ldots,i_l}\sum_{j_1,\ldots,j_l}(\pi_{1i_1}\pi_{2j_1})^{-1}\Gamma\prod_{\zeta=1}^{l}\left\{\frac{S_{i_1,\ldots,i_\zeta,j_1,\ldots,j_\zeta}(T^*_\zeta,T^*_\zeta)H_{i_1,\ldots,i_\zeta,j_1,\ldots,j_\zeta}(T^*_\zeta,T^*_\zeta)}{H_{1i_1,\ldots,i_\zeta}(T^*_\zeta)H_{2j_1\ldots j_\zeta}(T^*_\zeta)}\,\tilde{\theta}_{i_1,\ldots,i_\zeta,j_1,\ldots,j_\zeta}\right\} \\
&\quad \times\left[\sum_{i_1,\ldots,i_{l+1}}\sum_{j_1,\ldots,j_{l+1}}\tilde{\theta}_{i_1,\ldots,i_{l+1},j_1,\ldots,j_{l+1}}\{A_{1i_1,\ldots,i_{l+1}}(T^*_l)A_{2j_1,\ldots,j_{l+1}}(T^*_l)+C(l)\}\frac{A_{1i_1,\ldots,i_l}(T^*_l)A_{2j_1,\ldots,j_l}(T^*_l)}{S_{1i_1,\ldots,i_l}(T^*_l)S_{2j_1,\ldots,j_l}(T^*_l)}\right]
\end{aligned}
$$

where for $l=1,\ldots,s$,

$$
\begin{aligned}
&C(l) \\
&= \int_{T^*_l}^{T^*_{l+1}}\int_{T^*_l}^{T^*_{l+1}}[\{A_{1i_1,\ldots,i_{l+1}}(T^*_{l+1})+A_{1i_1,\ldots,i_{l+1}}(u)\}\{A_{2j_1,\ldots,j_{l+1}}(T^*_{l+1})+A_{2j_1,\ldots,j_{l+1}}(v)\} \\
&\quad -2A_{1i_1,\ldots,i_{l+1}}(T^*_{l+1})A_{2j_1,\ldots,j_{l+1}}(T^*_{l+1})]G_{i_1,\ldots,i_{l+1},j_1,\ldots,j_{l+1}}(u,v)\,\mathrm{d}u\,\mathrm{d}v
\end{aligned}
$$

In estimating the variance of the test statistic, variance terms may be substituted by their maximum likelihood estimates. Let $\hat{A}_{gi_1,\ldots,i_m}(x)=\int_x^\infty w(u)\hat{S}_{gi_1,\ldots,i_m}(u)\,\mathrm{d}u$. An estimate for

$$
\begin{aligned}
&G_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)\ \text{is}\ \hat{G}_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v) \\
&\quad = n^*_{i_1,\ldots,i_m,j_1,\ldots,j_m}\{\tilde{Y}_{1i_1,\ldots,i_m}(u)\tilde{Y}_{2j_1,\ldots,j_m}(v)\}^{-1} \\
&\qquad \times\tilde{Y}_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)[\{\tilde{Y}_{i_1,\ldots,i_m,j_1\ldots,j_m}(u,v)\}^{-1}\,\mathrm{d}\tilde{N}_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v) \\
&\qquad -\{\tilde{Y}_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)\tilde{Y}_{2j_1,\ldots,j_m}(v)\}^{-1}\,\mathrm{d}\tilde{N}_{1i_1,\ldots,i_m|2_{j_1,\ldots,j_m}}(u|v)\,\mathrm{d}\tilde{N}_{2j_1,\ldots,j_m}(v) \\
&\qquad -\{\tilde{Y}_{i_1,\ldots,i_m,j_1\ldots,j_m}(u,v)\tilde{Y}_{1i_1,\ldots,i_m}(u)\}^{-1}\,\mathrm{d}\tilde{N}_{2j_1,\ldots,j_m|1i_1,\ldots,i_m}(v|u)\,\mathrm{d}\tilde{N}_{1i_1,\ldots,i_m}(u) \\
&\qquad +\{\tilde{Y}_{1i_1,\ldots,i_m}(u)\tilde{Y}_{2j_1,\ldots,j_m}(v)\}^{-1}\,\mathrm{d}\tilde{N}_{1i_1,\ldots,i_m}(u)\,\mathrm{d}\tilde{N}_{2j_1,\ldots,j_m}(v)]
\end{aligned}
$$

Here, $n^*_{i_1,\ldots,i_m,j_1,\ldots,j_m}=(n_{i_1,\ldots,i_m}n_{j_1,\ldots,j_m}/\tilde{n}_{i_1,\ldots,i_m,j_1,\ldots,j_m})$, $\tilde{Y}_{gi_1,\ldots,j_m}(u)$ counts the number still in group $g$ at risk for failure at time $u$ with covariate history $i_1,\ldots,i_m$, and $\mathrm{d}\tilde{N}_{gi_1,\ldots,i_m}(u)$ counts the number in group $g$ with covariate history $i_l,\ldots,i_m$, that failed at time $u$. Also, $\tilde{Y}_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)$ counts the number of complete pairs still at risk for failure at times $u$ and $v$ with covariate histories $i_1,\ldots,i_m$, $j_1,\ldots,j_m$, $\mathrm{d}\tilde{N}_{i_1,\ldots,i_m,j_1,\ldots,j_m}(u,v)$ counts individuals from complete pairs where the group

 

1 member with covariate history $i_1,\ldots,i_m$, failed at time $u$, and the group 2 member with covariate history $j_1,\ldots,j_m$ failed at time $v$, and $\mathrm{d}\tilde{N}_{g_1 i_1,\ldots,i_m|g_2 j_1,\ldots,j_m}(u|v)$ counts the number of pairs where the group 1 member with covariate history $i_1,\ldots,i_m$ failed at time $u$ and the group 2 member with covariate history $j_1,\ldots,j_m$ is still at risk at time $v$.

## REFERENCES


1. Early Treatment Diabetic Retinopathy Study Research Group. Early treatment diabetic retinopathy study design and baseline patient characteristics: ETDRS report number 7. *Ophthalmology* 1991; **98**:741–756.
2. Early Treatment Diabetic Retinopathy Study Research Group. Early photocoagulation for diabetic retinopathy: ETDRS report number 9. *Ophthalmology* 1991; **98**:766–785.
3. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 1980.
4. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**:141–151.
5. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty and the dynamics of mortality. *Demography* 1979; **16**:439–454.
6. Hougaard P, Harvard B, Holm N. Measuring the similarities between the lifetimes of adult danish twins born between 1881–1930. *Journal of the American Statistical Association* 1992; **87**:17–24.
7. Oakes D. Bivariate survival models induced by frailties. *Journal of the American Statistical Association* 1989; **84**:487–493.
8. Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association* 1998; **93**:1164–1175.
9. Murray S, Tsiatis A. Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. *Lifetime Data Analysis* 2001; **7**:125–141.
10. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*: *Methodological Issues*, Jewell N, Dietz K, Farewell V (eds). Birkhäuser: Boston, 1992.
11. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000; **56**:779–788.
12. Malani HM. A modification of the redistribution to the right algorithm using disease markers. *Biometrika* 1995; **82**:515–526.
13. Woolson R, Lachenbruch P. Rank tests for censored matched pairs. *Biometrika* 1980; **67**:597–606.
14. Wei LJ. A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *Journal of the American Statistical Association* 1980; **75**:634–637.
15. O'Brien P, Fleming T. A paired Prentice–Wilcoxon test for censored paired data. *Biometrics* 1987; **43**: 169–180.
16. Dabrowska D. Rank tests for matched pair experiments with censored data censored matched pairs. *Journal of Multivariate Analysis* 1989; **28**:88–114.
17. Dabrowska D. Signed-rank tests for censored matched pairs. *Journal of the American Statistical Association* 1990; **85**:478–485.
18. Jung S. Rank tests for matched survival data. *Lifetime Data Analysis* 1999; **5**:67–69.
19. Murray S. Nonparametric rank-based methods for group sequential monitoring of paired censored survival data. *Biometrics* 2000; **56**:984–990.
20. Murray S. Using weighted Kaplan–Meier statistics in nonparametric comparisons of paired censored survival outcomes. *Biometrics* 2001; **57**:361–368.
21. Gill RD. *Censoring and Stochastic Integrals*. Mathematical Center Tract 124. Mathematische Centrum: Amsterdam, 1980.
22. Pepe MS, Fleming TR. Weighted Kaplan–Meier statistics: large sample theory and optimality considerations. *Journal of the Royal Statistical Society B* 1991; **53**:341–352.
23. Messinger Cayetano S. Nonparametric paired tests for censored survival data incorporating prognostic covariate information. *Ph.D. Dissertation*, Department of Biostatistics, University of Michigan, Ann Arbor, MI.
24. Murray S, Tsiatis A. A nonparametric approach in incorporating prognostic longitudinal covariate information in survival estimation. *Biometrics* 1996; **52**:137–151.


 