Biometrics 61, 715–720
September 2005

DOI: 10.1111/j.1541-0420.2005.00337.x

Simultaneous Group Sequential Analysis of Rank-Based and Weighted Kaplan–Meier Tests for Paired Censored Survival Data

Adin-Cristian Andrei* **and Susan Murray**

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109, U.S.A.
**email:* andreia@umich.edu

Summary. This research sequentially monitors paired survival differences using a new class of nonparametric tests based on functionals of standardized paired weighted log-rank (PWLR) and standardized paired weighted Kaplan–Meier (PWKM) tests. During a trial, these tests may alternately assume the role of the more extreme statistic. By monitoring PEMAX, the maximum between the absolute values of the standardized PWLR and PWKM, one combines advantages of rank-based (RB) and non-RB paired testing paradigms. Simulations show that monitoring treatment differences using PEMAX maintains type I error and is nearly as powerful as using the more advantageous of the two tests in proportional hazards (PH) as well as non-PH situations. Hence, PEMAX preserves power more robustly than individually monitored PWLR and PWKM, while maintaining a reasonably simple approach to design and analysis of results. An example from the Early Treatment Diabetic Retinopathy Study (ETDRS) is given.

Key words: Clinical trials; Group sequential monitoring; Nonparametric; Paired weighted Kaplan–Meier; Paired weighted log-rank.

1. Introduction

At the design stages of clinical trials comparing survival outcomes in independent groups, a common plan is to base the design upon a log-rank (LR) statistic of some form (see, for example, Gehan, 1965; Gill, 1980). Another approach for stochastically ordered alternatives is to compare areas under survival curves (see, for example, Pepe and Fleming, 1989). Versatile tests combining rank-based (RB) and non-RB statistics for independent groups are studied by Chi and Tsai (2001), while Kosorok and Lin (1999) develop sophisticated methods for combining various RB tests. Fundamental independent group sequential methods for families of weighted LR (WLR) tests have been developed and studied by Tsiatis (1981, 1982), Sellke and Siegmund (1983), Slud (1984), and Gu and Lai (1991), among others, and sequential methods for comparing areas under survival curves were developed by Murray and Tsiatis (1999).

For paired censored survival data, where optimality properties for the paired WLR (PWLR) have not been studied, competing methodologies exist to a lesser extent. Some RB and frailty methods are presented by O'Brien and Fleming (1987), Dabrowska (1986, 1990), Murray (2000), and Oakes and Jeong (1998), among others, and paired Pepe–Fleming tests are developed by Murray (2001, 2002). Paired survival data arise in various situations including time to death, disease occurrence or other morbidity in twins, time to vision loss in paired eyes, or failure of matched allografts. For example, 3711 patients with diabetic retinopathy in both eyes were enrolled in the Early Treatment Diabetic Retinopathy Study (ETDRS, 1991a,b) from April 1980 to July 1985, with one eye per patient randomly assigned to early photocoagulation and the other to deferral of photocoagulation until detection of high-risk proliferative retinopathy.

In paired settings such as ETDRS, little research involving multiple test statistics is available. Further complicating the design choice in the group sequential setting, the preferred test may change from one interim analysis to the next.

This research is motivated by a desire to formalize inference in the following scenario. Assume that in a paired censored survival analysis with group sequential monitoring, an investigator first uses a PWLR and fails to reject the null hypothesis by a narrow margin. Then, a paired weighted Kaplan–Meier (PWKM) test is recalled as an attractive alternative and it leads to statistical significance. Or perhaps at different analysis times, statistical advantages are attributed alternately to PWLR or PWKM. In this setting, we provide a middle ground that allows monitoring of both tests, while adjusting for their joint use over time. The proposed test, PEMAX, which is the maximum of the absolute values of the standardized PWLR and PWKM, will be seen to preserve type I error and to have power comparable to the better of these competing tests.

The rest of this article is organized as follows. In Section 2, the sequential joint limiting distribution of PWLR and PWKM is outlined, from which the joint distribution of PEMAX over time is estimated. Although this section is

useful in reviewing some general notation for group sequential analysis of clinical trials and for understanding how to program the methodology, some practitioners may skip it if only interested in the operating characteristics of the method. Technical details regarding closed-form asymptotic covariances and corresponding estimates are relegated to the Appendix. Section 3 presents simulations assessing the moderate-sized sample performance of PEMAX as compared to PWLR and PWKM. Sequential monitoring of the ETDRS using PEMAX is shown in Section 4, and Section 5 is dedicated to comments and conclusions.

2. Joint Sequential Distribution of PWLR and PWKM

2.1 *Background and Notation*

During accrual, n *i.i.d.* data pairs (e.g., n pairs of twins) are enrolled (at least one pair member) in a prospective study ending at time τ. Although in practice pair members usually enter the study simultaneously, this research allows for differential pair member entry times. Suppose that pair $l = 1, \ldots, n$, member $g = 1, 2$, enters the study at time E_{gl} (a calendar time during accrual), has underlying survival time T_{gl}, and potential censoring or loss-to-follow-up time L_{gl} (both regarded as study times measured since E_{gl}). Further assume that within group g, $(E_{gl},\ T_{gl},\ L_{gl},\ l = 1, \ldots, n)$ are *i.i.d.* continuously distributed with survival functions $P(E_{gl} > e)$, $S_g(s) = P(T_{gl} > s)$, and $C_g(c) = P(L_{gl} > c)$, respectively. For technical reasons, the correlation between paired survival times is assumed to be strictly less than 1.

At analysis time t, $n_g(t) = \sum_{l=1}^{n} I(E_{gl} \le t)$ pair members g have entered the study, while $X_{gl}(t) = \min\{T_{gl},\ L_{gl},\ \max(t - E_{gl}, 0)\}$ and $\Delta_{gl}(t) = I\{T_{gl} \le \min(L_{gl}, t - E_{gl})\}$ are the follow-up time and the censoring indicator, respectively, for pair l, member g. There are $n_{g_1 g_2}(t_1, t_2) = \sum_{l=1}^{n} I(E_{g_1 l} \le t_1, E_{g_2 l} \le t_2)$ pairs whose member g_1 has entered the study by analysis time t_1 and member g_2 has entered by analysis time t_2. Let $N_g(t, u) = \sum_{l=1}^{n} I\{X_{gl}(t) \le u, \Delta_{gl}(t) = 1\}$ and $Y_g(t, u) = \sum_{l=1}^{n} I\{X_{gl}(t) \ge u\}$ be the number of failed and, respectively, the number of at-risk pair members g at analysis time t and study time u.

2.2 *PWLR(t) and PWKM(t) Test Statistics*

Let

$$J(t,u) = I(0 \le u \le \tau) I\{Y_1(t,u) Y_2(t,u) > 0\},$$

$$p(t,u) = I(0 \le u \le \tau) I[P\{X_{1l}(t) \ge u\} P\{X_{2l}(t) \ge u\} > 0]$$

and assume that $J(t,u) \xrightarrow{\mathcal{P}} p(t,u)$, for all fixed t.

At time t,

$$\mathrm{PWLR}(t) = \sqrt{n^*(t)} \int_0^\infty J(t,u) K(t,u) \times \sum_{g=1}^{2} (-1)^{g+1} \{Y_g(t,u)\}^{-1} N_g(t, du),$$

with

$$n^*(t) = \left\{ \prod_{g=1}^{2} n_g(t) \right\} \left\{ \sum_{g=1}^{2} n_g(t) \right\}^{-1},$$

$$K(t,u) = W_{pwlr}(t,u) \left\{ \prod_{g=1}^{2} Y_g(t,u) \right\} \left\{ n^*(t) \sum_{g=1}^{2} Y_g(t,u) \right\}^{-1},$$

and the weighting function $W_{pwlr}(t,\ u)$ converging in probability to a deterministic function $w_{pwlr}(t,\ u)$ on $[0,\ t]$. Weights such as $W_{pwlr}(t,u) = 1$ or $W_{pwlr}(t,u) = \prod_{g=1}^{2} Y_g(t,u) \{n_g(t)\}^{-1}$ yield the paired LR and the paired Gehan test, respectively.

Also at time t, $\mathrm{PWKM}(t) = (n^*(t))^{1/2} \int_0^\infty J(t,u) \hat{W}_{pwkm}(t,u) \times \sum_{g=1}^{2} (-1)^{g+1} \hat{S}_g(t,u)\, du$, where $\hat{S}_g(t,u)$ is the Kaplan–Meier (KM) estimator of $S_g(u)$ based on time t data and the weighting process $\hat{W}_{pwkm}(t,u)$ converges in probability to a deterministic function $w_{pwkm}(t,\ u)$ on $[0,\ t]$. With $\hat{\pi}_g(t) = n_g(t) \{\sum_{h=1}^{2} n_h(t)\}^{-1}$ and $\hat{H}_g(t,u)$ being the KM estimator of $H_g(t,u) = P(L_g \ge u, t - E_g \ge u \,|\, E_g \le t)$, possible $\hat{W}_{pwkm}(t,u)$ choices are $\{\prod_{g=1}^{2} \hat{H}_g(t, u-)\} \{\sum_{g=1}^{2} \hat{\pi}_g(t) \hat{H}_g(t, u-)\}^{-1}$, which is in the spirit of the weighting recommended by Pepe and Fleming (1989), or alternatively $\hat{W}_{pwkm}(t,u) = 1$, interpreted as paired years-of-life saved (PYLS) over τ years of study.

For stochastically ordered survival curves, the null hypothesis is $\mathcal{H}_0 : S_1(\cdot) = S_2(\cdot) = S(\cdot)$ on $[0,\ \tau]$. If $t_1 < t_2 < \cdots < t_D$ are successive analysis times such that the statistical information expected between them is sufficient to warrant additional analyses, then it follows that $\{\mathrm{PWLR}(t_1), \mathrm{PWKM}(t_1), \ldots, \mathrm{PWLR}(t_D), \mathrm{PWKM}(t_D)\}^{\mathrm{T}} \xrightarrow{\mathcal{D}} N_{2D}(\mathbf{0_{2D}}, \mathbf{\Sigma})$. The covariance matrix $\mathbf{\Sigma}$ is described in the Appendix, together with consistent estimators of its entries.

One may now obtain quantiles for any functional of PWLR and PWKM by means of Monte Carlo simulations of the estimated joint sequential distribution. Instructive examples of Monte Carlo simulations used in the group sequential monitoring context are given, for example, in Murray (2002) and the companion technical report by Andrei and Murray (2004).

3. Simulation Studies

Simulations are conducted under four different scenarios with stochastic ordering to assess the finite-sample behavior of PEMAX as compared to PWLR and PWKM, when pairing in data is accounted for, and separately when pairing is ignored. In the latter case, we denote the statistics of interest by EMAX rather than PEMAX. We chose PYLS to represent the PWKM family because of its simple interpretation as the number of years-of-life saved while on study. Under each scenario, both under the null and the alternative hypotheses, 1000 Monte Carlo simulation runs consisting of 100 pairs of correlated survival times are generated with correlation of approximately 0.25 and we employ an O'Brien–Fleming error-spending function, using calendar time as a surrogate for the total information accrued, to spend an overall 5% type I error. Also of major interest is to understand how correctly accounting for the paired nature of the data improves the operating characteristics of PEMAX.

Under scenario piecewise exponential (PE), paired PE survival times are generated. We assume a common pair entry time to be uniform (0, 0.25) and conduct interim analyses at times 0.5, 1, and 2. In both groups, the hazard rates are piecewise constant equal to 0.7, 1.3, and 1 in group 1 and 1.3, 0.7, and 1, respectively, in group 2, with changes occurring at times 1 and 1.5. This formulation gives approximately proportional hazards (PH) at the first two interim analyses and crossing hazards at the last one. For the second scenario paired Weibull (PW), PW survival times are generated from Weibull (2.8, 0.5) and Weibull (1.5, 0.8), the common entry times are $U(0, 1)$, and interim analyses are conducted at times 1, 2.5, and 4. Under the third scenario accelerated failure time (AFT), the common entry times are $U(0, 2)$, and the survival times (T_1, T_2) are based upon AFT models $\log(T_g) = \mu_g + 0.5Z_g + 0.25\,W_g$, $g = 1, 2$, where $\mu_1 = 1$, $\mu_2 = 1.2$, Z_1, Z_2 are correlated $U(0, 1)$-distributed, and the error terms W_1, W_2 are *i.i.d.* $N(0, 1)$. Interim analyses are performed at times 2, 4, and 6. Finally, under the PH scenario, the common entry times are $U(0, 12)$, and the correlated survival times are based upon PH models with hazards $h_g(t \mid X_g, Y_g) = h_{0,g} \exp\{\alpha_g X_g + \beta_g Y_g\}$, $g = 1, 2$, where $h_{0,1} = 0.05$, $h_{0,2} = 0.02857$, $\alpha_1 = 0.2$, $\beta_1 = 0.4$, $\alpha_2 = 0.3$, $\beta_2 = 0.6$, X_1, X_2 are independently *Gamma*(0.2, 1) and *Gamma*(0.4, 1)-distributed, respectively, and Y_1, Y_2 are independently generated from $N(0, 0.03)$. Then, *i.i.d.* pairs (U_1, U_2) of correlated $U(0, 1)$ random variables are generated. If $T_g = -\{h_g(t \mid X_g, Y_g)\}^{-1} \times \log(1 - U_g)$, then (T_1, T_2) are correlated survival times generated based on PH structures. Interim analyses are conducted at times 12, 18, and 24.

Size and power simulation results are presented in Table 1. The PLR, PYLS, and PEMAX tests (those accounting for pairing) maintain size close to the nominal 0.05 level. Ignoring pairing results in size levels diminished by almost 50%, implying overconservativeness for all three tests. Under each scenario, the power loss percentages exhibited by the other tests are presented with respect to the most powered test under the respective scenario. As expected, the paired versions of the tests observed are all more powerful than any of those that ignore pairing. Under the four scenarios, using PEMAX, power gains of 3.79%, 2.20%, 4.36%, and, respectively, 0.38% over the disadvantaged test are observed. Power losses using PEMAX as opposed to the more powered test are minimal under each scenario. The competing PYLS and PLR tests occasionally take turns in being more powered as follow-up increases under the various scenarios presented, but monitoring based on PEMAX produces power comparable to the more powered of PLR and PYLS, when the alternative hypothesis is in doubt. Ignoring pairing induces serious power losses of at least 10% in most cases, some even as large as 18.04%, associated with the unpaired YLS test under PW.

4. Example

Recall the ETDRS example described in the introduction. The 3711 patients enrolled between April 1980 and July 1985 were followed in order to detect vision loss defined as visual acuity less than 5/200 at two consecutive visits, but due to either loss-to-follow-up or administrative censoring, this primary endpoint was not observed for everybody.

In order to make the analysis more interesting, we restrict to about 25% of the data consisting of 999 patients enrolled prior to February 15, 1983 who were taking a placebo pill in a separate randomization process. Because the causes that may ultimately lead to vision loss are common, there tends to exist a mild-to-moderate positive correlation between the loss of visual acuity in the left and right eye of an individual. The staggered entry feature, the presence of censoring, and the ethical reasons requiring a periodic examination of

Table 1

Size and power simulation results for paired and unpaired YLS, LR, and EMAX based on 1000 *replications with an overall type I error* $\alpha = 0.05$. *Power loss results under each scenario (in percentages) are relative to the more powered test (indicated by* **) under the respective scenario.*

		Paired			Unpaired		
Scenario	Test	PYLS	PLR	PEMAX	YLS	LR	EMAX
PE	Size	0.047	0.047	0.046	0.026	0.025	0.019
	Power	0.897*	0.848	0.882	0.840	0.755	0.819
	Power loss	0%	5.46%	1.67%	6.35%	15.83%	8.69%
PW	Size	0.049	0.045	0.045	0.022	0.019	0.019
	Power	0.687	0.726*	0.703	0.595	0.630	0.603
	Power loss	5.37%	0%	3.17%	18.04%	13.22%	16.94%
AFT	Size	0.056	0.049	0.048	0.035	0.028	0.027
	Power	0.780*	0.725	0.759	0.703	0.648	0.690
	Power loss	0%	7.05%	2.69%	9.87%	16.92%	11.54%
PH[a]	Size	0.058	0.045	0.047	0.028	0.029	0.027
	Power	0.782*	0.769	0.772	0.699	0.691	0.689
	Power loss	0%	1.66%	1.28%	10.61%	11.64%	11.89%

[a]Under the PH scenario, PYLS is negligibly better powered than PLR, the latter being most powerful under PH with independent groups. However, when correlation is introduced to the PH setting, there is no theory asserting similar PLR properties, and it might not be surprising that PLR does not seem to take advantage as well as PYLS of the fact that the hazard rates, and implicitly the survival curves, vary in tandem.

Table 2
Paired and unpaired versions of (P)YLS and (P)LR tests together with the O'Brien–Fleming (P)EMAXb stopping boundaries of the corresponding (P)EMAX test for the 999 *patients enrolled prior to February* 15, 1983 *who are taking a placebo pill*

Analysis time	Error spent	Paired			Unpaired		
		PYLS	PLR	PEMAXb	YLS	LR	EMAXb
1	2.85×10^{-5}	−2.119	1.900	4.453	−1.169	1.441	4.051
2	1.42×10^{-4}	−2.530	2.340	4.031	−1.970	1.810	3.663
3	5.74×10^{-4}	−3.060	3.006	3.517	−2.453	2.437	3.287
4	1.18×10^{-3}	−2.846	3.006	3.320	−2.272	2.472	3.106
5	1.31×10^{-3}	−2.482	2.674	3.197	−1.928	2.165	2.940
6	2.34×10^{-3}	−2.700	2.653	2.988	−2.095	2.140	2.722
7	1.33×10^{-3}	−2.716	2.423	2.996	−2.074	1.909	2.715
8 ←	2.27×10^{-3}	**−3.106**	2.828	**2.892**	−2.412	2.284	2.590
9	8.29×10^{-4}	−3.179	2.886	2.918	−2.490	2.348	2.585

Bold font indicates that the PEMAXb boundary has been exceeded.

the data make this example suitable for analysis using group sequential methods and PEMAX will be employed. A number of nine interim analyses are planned, proceeding after the first 50 events have occurred and continuing every 6 months thereafter and the overall 1% type I error is spent using an O'Brien–Fleming error-spending function. The proportion of deaths observed at each analysis time is used as a surrogate for the proportion of total information in the spending function. Strategies for error spending are discussed in O'Brien and Fleming (1979), Lan and DeMets (1983), and summarized in Jennison and Turnbull (2000).

For the 999 placebo patients, the results in Table 2 show that the PEMAX rejects at the eighth interim analysis, where the standardized PYLS exceeds the PEMAX sequential boundary. Interestingly, PLR and PYLS take turns in getting closer to statistical significance as the monitoring process unfolds, making it attractive to monitor both throughout the study. When data pairing is ignored, not one of the tests employed detects significant survival differences between the two treatment groups. Using PEMAX to repeat the same testing procedure for the 1010 patients that receive an aspirin pill instead of placebo results in the detection of significant survival differences at the sixth analysis time, when PLR exceeds the corresponding PEMAX boundary, while PYLS does not (see Table 3). Hence, under a very similar study design, PEMAX detects survival differences driven this time by PLR.

This example illustrates how the favored design choice is not always obvious since the only protocol difference between the two patient cohorts was the assignment to placebo or aspirin in addition to the paired design for studying early versus delayed photocoagulation. In each case, PEMAX tracked well with the more favored design, detecting the difference of interest.

5. Discussion

The newly proposed test, PEMAX, has several features that distinguish it from the individual tests. Although RB tests are generally favored when a PH situation is anticipated, Pepe and Fleming (1989) have shown a lack of sensitivity to the magnitude of the difference between the survival curves and have proposed WKM statistics. PEMAX is set to balance the advantages and disadvantages associated with these families of tests in the paired censored survival data setting. Thus, it should not be surprising that it might provide a degree of robustness to detect ordered survival curves, when dealing with PH or crossing hazards situations.

Table 3
Paired and unpaired versions of (P)LR and (P)YLS tests together with the O'Brien–Fleming (P)EMAXb stopping boundaries of the corresponding (P)EMAX test for the 1010 *patients enrolled prior to February* 15, 1983 *who are taking an aspirin pill*

Analysis time	Error spent	Paired			Unpaired		
		PYLS	PLR	PEMAXb	YLS	LR	EMAXb
1	2.85×10^{-5}	−0.518	0.660	3.847	−0.405	0.512	4.318
2	1.42×10^{-4}	−0.742	1.058	3.661	−0.587	0.828	3.769
3	5.74×10^{-4}	−1.271	1.218	3.501	−1.050	0.982	3.795
4	1.18×10^{-3}	−2.458	2.806	3.179	−2.023	2.305	3.177
5	1.31×10^{-3}	−2.097	2.630	3.103	−1.670	2.124	3.106
6 ←	2.34×10^{-3}	−2.528	**3.155**	**2.933**	−2.021	2.531	2.939
7	1.33×10^{-3}	−2.556	3.166	2.896	−2.047	2.539	2.981
8	2.27×10^{-3}	−2.483	2.965	2.834	−1.997	2.394	2.823
9	8.29×10^{-4}	−2.583	3.046	3.112	−2.077	2.453	3.185

Bold font indicates that the PEMAXb boundary has been exceeded.

Associated with PEMAX come the advantages of being able to: (1) account for correlation between paired outcomes, (2) account for correlation between PWLR and PWKM, and (3) control type I error within the group sequential monitoring framework. Testing frameworks that fail to account for the source of correlation in (1) are generally inefficient. Frameworks that ignore repeated testing in (2) and (3) will have inflated size. Although the focus is on PEMAX, other tests built upon functionals of PWLR and PWKM, such as linear combinations of these, could be devised as seen fit. Their sequential limiting behavior is readily available, given the closed-form expressions for the joint limiting distribution of the PWLR and PWKM.

This methodology adds to the literature available for the analysis of clinical trials involving paired survival outcomes. With overconservativeness being an issue when paired structures are overlooked, using PEMAX would account for the true nature of the data and give the benefits of using the correlation in the data. Although statistical literature has been rapidly advancing in broadening the ability to monitor different types of test statistics with different forms of alternatives in the independent setting, this availability is still in its infancy in the paired setting. This procedure reduces the temptation to use methods designed for independent settings when the censored survival data are paired.

Acknowledgements

The authors would like to thank the Early Treatment Diabetic Retinopathy Study Research Group and particularly Marian R. Fisher for the data used in writing this manuscript. The authors thank the editor and the associate editor for their constructive comments and helpful suggestions.

References

Andrei, A.-C. and Murray, S. (2004). *Asymptotic results for simultaneous group sequential analysis of rank-based and weighted Kaplan–Meier tests for paired survival data in the presence of censoring*. Technical Report, Working Paper Series, Working Paper 43, Department of Biostatistics, The University of Michigan.

Chi, Y. and Tsai, M. H. (2001). Some versatile tests based on the simultaneous use of weighted logrank and weighted Kaplan–Meier statistics. *Communications in Statistics, Part B—Simulation and Computation* **30,** 743–759.

Dabrowska, D. M. (1986). Rank tests for independence for bivariate censored data. *Annals of Statistics* **14,** 250–264.

Dabrowska, D. M. (1990). Signed-rank tests for censored matched pairs. *Journal of the American Statistical Association* **85,** 478–485.

Early Treatment Diabetic Retinopathy Study Research Group. (1991a). Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics: ETDRS report number 7. *Ophthalmology* **98,** 741–756.

Early Treatment Diabetic Retinopathy Study Research Group. (1991b). Early photocoagulation for diabetic retinopathy: ETDRS report number 9. *Ophthalmology* **98,** 766–785.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **52,** 203–223.

Gill, R. D. (1980). Censoring and stochastic integrals. In *Mathematical Centre Tracts*, Volume 124. Amsterdam: Mathematisch Centrum.

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Annals of Statistics* **19,** 1403–1433.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: CRC Press.

Kosorok, M. R. and Lin, C. Y. (1999). Versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association* **94,** 320–332.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70,** 659–663.

Murray, S. (2000). Nonparametric rank-based methods for group sequential monitoring of paired censored survival data. *Biometrics* **56,** 984–990.

Murray, S. (2001). Using weighted Kaplan–Meier statistics in nonparametric comparisons of paired censored survival outcomes. *Biometrics* **57,** 361–368.

Murray, S. (2002). Group sequential monitoring of years of life saved with paired censored survival data. *Statistics in Medicine* **21,** 177–189.

Murray, S. and Tsiatis, A. A. (1999). Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics* **55,** 1085–1092.

Oakes, D. and Jeong, J. H. (1998). Frailty models and rank tests. *Lifetime Data Analysis* **4,** 209–228.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

O'Brien, P. C. and Fleming, T. R. (1987). A paired Prentice–Wilcoxon test for censored paired data. *Biometrics* **43,** 169–180.

Pepe, M. and Fleming, T. R. (1989). Weighted Kaplan–Meier statistics: A class of distance tests for censored survival data. *Biometrics* **45,** 497–507.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70,** 315–326.

Slud, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics* **12,** 551–571.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* **68,** 311–315.

Tsiatis, A. A. (1982). Group sequential methods for survival analysis with staggered entry. In *Survival Analysis*, Volume 2, J. Crowley and R. A. Johnson (eds), 257–268. Beachwood, Ohio: Institute of Mathematical Statistics.

Received January 2004. *Revised October* 2004.
Accepted November 2004.

Appendix

The asymptotic proportion $\pi_g(t)$ of pair members g entered by analysis time t is estimated by $\hat{\pi}_g(t) = n_g(t)\{\sum_{h=1}^{2} n_h(t)\}^{-1}$.

For group g, the probability $\pi_g(t_1 \mid t_2)$ of study entry by analysis time t_1, given entry by analysis time t_2, where $t_1 \le t_2$, is consistently estimated by $\hat{\pi}_g(t_1 \mid t_2) = n_g(t_1)\{n_g(t_2)\}^{-1}$. The number of dependent pairs in groups g_1 and g_2 at analysis times t_1 and t_2, respectively, is equal to $n_{g_1,g_2}(t_1,t_2)$, so the proportion $\theta_{g_1,g_2}(t_1,t_2)$ of such dependent observations is consistently estimated by $\hat{\theta}_{g_1,g_2}(t_1,t_2) = 2n_{g_1,g_2}(t_1,t_2)\{n_{g_1}(t_1) + n_{g_2}(t_2)\}^{-1}$. The asymptotic proportion $\gamma_{g_1,g_2}(t_1,t_2)$ of pair members g_1 that have entered by analysis time t_1, among the pairs where the other member has entered by analysis time t_2, is estimated by $\hat{\gamma}_{g_1,g_2}(t_1,t_2) = n_{g_1}(t_1)\{n_{g_1}(t_1) + n_{g_2}(t_2)\}^{-1}$. For $0 < p < 1$, let $OR(p) = p(1-p)^{-1}$ and $\psi_{g_1,g_2}(t_1,t_2) = 0.5 \times \theta_{g_1,g_2}(t_1,t_2) \quad (\pi_{3-g_1}(t_1)\pi_{3-g_2}(t_2))^{1/2} \quad [(OR\{\gamma_{g_1,g_2}(t_1,t_2)\})^{1/2} + (OR\{\gamma_{g_2,g_1}(t_2,t_1)\})^{1/2}]$.

The marginal cause-specific hazard function for pair member g at study time $0 \le u \le t$ is defined as $\lambda_g(u) = \lim_{\delta u \to 0} \frac{1}{\delta u} P\{X_{gl}(t) < u + \delta u, \Delta_{gl}(t) = 1 \mid X_{gl}(t) \ge u\}$. Using information on pair member g_1 at analysis time t_1 and pair member g_2 at analysis time t_2, the joint hazard at study time $0 \le u \le t_1$ (for member g_1) and study time $0 \le v \le t_2$ (for member g_2) is $\lambda_{g_1,g_2}\{(t_1,u),(t_2,v)\} = \lim_{\delta u,\delta v \to 0} \frac{1}{\delta u \delta v} \times P\{X_{g_1 l}(t_1) < u + \delta u, X_{g_2 l}(t_2) < v + \delta v, \Delta_{g_1 l}(t_1) = 1, \Delta_{g_2 l}(t_2) = 1 \mid X_{g_1 l}(t_1) \ge u, X_{g_2 l}(t_2) \ge v\}$. The cause-specific conditional hazard for pair member g_1 at study time $0 \le u \le t_1$, given that pair member g_2 is at-risk at study time $0 \le v \le t_2$, is $\lambda_{g_1 \mid g_2}\{(t_1,u) \mid (t_2,v)\} = \lim_{\delta u \to 0} \frac{1}{\delta u} P\{X_{g_1 l}(t_1) < u + \delta u, \Delta_{g_1 l}(t_1) = 1 \mid X_{g_1 l}(t_1) \ge u, X_{g_2 l}(t_2) \ge v\}$. Let

$$
\begin{aligned}
&R_{g_1,g_2}\{(t_1,u),(t_2,v)\} \\
&\quad = \lambda_{g_1,g_2}\{(t_1,u),(t_2,v)\} \\
&\qquad - \lambda_{g_1|g_2}\{(t_1,u) \mid (t_2,v)\}\lambda_{g_2}(v) \\
&\qquad - \lambda_{g_2|g_1}\{(t_2,v) \mid (t_1,u)\}\lambda_{g_1}(u) + \lambda_{g_1}(u)\lambda_{g_2}(v), \\
&B_{g_1,g_2}\{(t_1,u),(t_2,v)\} \\
&\quad = P\{X_{g_1 l}(t_1) \ge u, X_{g_2 l}(t_2) \ge v \mid E_{g_1 l} \le t_1, E_{g_2 l} \le t_2\} \\
&\qquad \times [P\{X_{g_1 l}(t_1) \ge u \mid E_{g_1 l} \le t_1\} P\{X_{g_2 l}(t_2) \ge v \mid E_{g_2 l} \le t_2\}]^{-1},
\end{aligned}
$$

and $G_{g_1,g_2} = R_{g_1,g_2} B_{g_1,g_2}$.

If $Q_g(t,\ u)$ denotes $S_g(u)H_g(t,\ u)$ and, as a reminder, $H_g(t,\ u) = P(L_g \ge u,\ t - E_g \ge u \mid E_g \le t)$ is the censoring survival function among group g members entered by analysis time t, then $k(t,u) = \lim_{n_1(t),n_2(t) \to \infty} K(t,u) = w_{pwlr}(t,u)\{\prod_{g=1}^{2} Q_g(t,u)\}\{\sum_{g=1}^{2} \pi_g(t) Q_g(t,u)\}^{-1}$ on $[0,t]$. For $0 \le u \le t$, define $A_g(t,u) = \int_u^\infty p(t,y) w_{pwkm}(t,y) S_g(y)\, dy$.

A.1 *Description of* $\boldsymbol{\Sigma}$

The entries of $\boldsymbol{\Sigma}$ are of the form $\mathrm{cov}\{\mathrm{PWLR}(t_i),\ \mathrm{PWLR}(t_j)\}$, $\mathrm{cov}\{\mathrm{PWKM}(t_i),\ \mathrm{PWKM}(t_j)\}$, $\mathrm{cov}\{\mathrm{PWLR}(t_i),\ \mathrm{PWKM}(t_j)\}$, or $\mathrm{cov}\{\mathrm{PWKM}(t_i), \mathrm{PWLR}(t_j)\}$, where $1 \le i \le j \le D$. If we let $\eta_{g,3-g}(t_i,\ t_j) = (\pi_{3-g}(t_i)\pi_{3-g}(t_j)\pi_g(t_i \mid t_j))^{1/2}$ and $p(t,\ u)k(t,\ u) = r(t,\ u)$, results detailed in a technical report by Andrei and Murray (2004) lead to $\mathrm{cov}\{\mathrm{PWLR}(t_i), \mathrm{PWLR}(t_j)\} = \sum_{g=1}^{2} [\eta_{g,3-g}(t_i,t_j) \int_0^\infty r(t_i,u) r(t_j,u) \{Q_g(t_j,u)\}^{-1} \lambda_g(u)\, du - \psi_{g,3-g}(t_i,t_j) \int_0^\infty \int_0^\infty r(t_i,u) r(t_j,v) G_{g,3-g}\{(t_i,u),(t_j,v)\}\, dv\, du]$.

Also,

$$
\begin{aligned}
&\mathrm{cov}\{\mathrm{PWKM}(t_i), \mathrm{PWKM}(t_j)\} \\
&\quad = \sum_{g=1}^{2} \Bigg[\eta_{g,3-g}(t_i,t_j) \int_0^\infty A_g(t_i,u) A_g(t_j,u) \\
&\qquad \times \{Q_g(t_j,u)\}^{-1} \lambda_g(u)\, du - \psi_{g,3-g}(t_i,t_j) \\
&\qquad \times \int_0^\infty \int_0^\infty A(t_i,u) A(t_j,v) G_{g,3-g}\{(t_i,u),(t_j,v)\}\, dv\, du \Bigg].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\mathrm{cov}\{\mathrm{PWLR}(t_i), \mathrm{PWKM}(t_j)\} \\
&\quad = \sum_{g=1}^{2} \Bigg[-\eta_{g,3-g}(t_i,t_j) \int_0^\infty r(t_i,u) A_g(t_j,u) \\
&\qquad \times \{Q_g(t_j,u)\}^{-1} \lambda_g(u)\, du + \psi_{g,3-g}(t_i,t_j) \\
&\qquad \times \int_0^\infty \int_0^\infty r(t_i,u) A_{3-g}(t_j,v) G_{g,3-g}\{(t_i,u),(t_j,v)\}\, dv\, du \Bigg].
\end{aligned}
$$

Finally,

$$
\begin{aligned}
&\mathrm{cov}\{\mathrm{PWKM}(t_i), \mathrm{PWLR}(t_j)\} \\
&\quad = \sum_{g=1}^{2} \Bigg[-\eta_{g,3-g}(t_i,t_j) \int_0^\infty A_g(t_i,u) r(t_j,u) \\
&\qquad \times \{Q_g(t_j,u)\}^{-1} \lambda_g(u)\, du + \psi_{g,3-g}(t_i,t_j) \\
&\qquad \times \int_0^\infty \int_0^\infty A_g(t_i,u) r(t_j,v) G_{g,3-g}\{(t_i,u),(t_j,v)\}\, dv\, du \Bigg].
\end{aligned}
$$

A.2 *Estimation of* $\boldsymbol{\Sigma}$

Let $Y_{g_1,g_2}\{(t_i,u),(t_j,v)\}$ be the number of pairs in which, at analysis time t_i pair member g_1 is at-risk at study time u and at analysis time t_j pair member g_2 is at-risk at study time v. Define $N_{g_1,g_2}\{(t_i,du),(t_j,dv)\}$ to be the number of pairs in which group g_1 member, who has entered by analysis time t_i, fails at study time u and group g_2 member, who has entered by analysis time t_j, fails at study time v. Finally, the number of pairs for which the group g_1 member, who has entered by analysis time t_i, is at-risk until and fails at study time u and group g_2 member, who has entered by analysis time t_j, is still at-risk at study time v is denoted by $N_{g_1|g_2}\{(t_i,du) \mid (t_j,v)\}$. An estimator for $P\{X_g(t_i) \ge u \mid E_g \le t_i\}$ is $Y_g(t_i,u) \times \{n_g(t_i)\}^{-1}$, while $P\{X_{g_1}(t_i) \ge u,\ X_{g_2}(t_j) \ge v \mid E_{g_1} \le t_i,\ E_{g_2} \le t_j\}$ is estimable by $Y_{g_1,g_2}\{(t_i,u),\ (t_j,v)\}\{n_{g_1,g_2}(t_i,\ t_j)\}^{-1}$. Nelson–Aalen-type estimators of $\lambda_g(u)du$, $\lambda_{g_1,g_2}\{(t_i,u),(t_j,v)\}\, du\, dv$, and $\lambda_{g_1|g_2}\{(t_i,u) \mid (t_j,v)\}\, du$ are available through $\{Y_g(t_j,\ u)\}^{-1} N_g(t_j,\ du)$, $N_{g_1,g_2}\{(t_i,du),(t_j,dv)\} \times [Y_{g_1,g_2}\{(t_i,u),(t_j,v)\}]^{-1}$, and $N_{g_1|g_2}\{(t_i,du) \mid (t_j,v)\}\, [Y_{g_1,g_2}\{(t_i,u),(t_j,v)\}]^{-1}$, respectively. A consistent estimator of $G_{g,3-g}\{(t_i,\ u),\ (t_j,\ v)\}\, dv\, du$ can be obtained based on previously described estimators of its components. Thus, an estimator of the covariance matrix $\boldsymbol{\Sigma}$ is now available. More details can be found in Andrei and Murray (2004).