University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 43

Asymptotic Results for Simultaneous Group Sequential Analysis of Rank-Based and Weighted Kaplan-Meier Tests with Paired Survival Data in the Presence of Censoring. Technical report

Adin-Cristian Andrei* Susan Murray†

*University of Michigan - Biostatistics, andreia@umich.edu

†University of Michigan Biostatistics, skmurray@umich.edu

Asymptotic Results for Simultaneous Group Sequential Analysis of Rank-Based and Weighted Kaplan-Meier Tests with Paired Survival Data in the Presence of Censoring. Technical report

Adin-Cristian Andrei and Susan Murray

Abstract

This research sequentially monitors paired survival differences using a new class of non-parametric 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 and non rank-based 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.

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

Adin-Cristian Andrei
Department of Biostatistics,
University of Michigan,
1420 Washington Heights, Ann Arbor, Michigan, 48109, U.S.A.
Phone: (734)-936-4035 and/or (734)-764-6872
Fax: (734)-763-2215
Email: andreia@umich.edu

Susan Murray
Department of Biostatistics,
University of Michigan,
1420 Washington Heights, Ann Arbor, Michigan, 48109, U.S.A.

.

SUMMARY. This research sequentially monitors paired survival differences using a new class of non-parametric 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 and non rank-based 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), Mantel (1966) or 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 rank-based tests. Fundamental independent groups sequential methods for families of weighted LR (WLR) tests have been developed and studied by Tsiatis (1981, 1982), Sellke and Siegmund (1983), Slud (1984), 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 weighted LR ($PWLR$) 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), Oakes (1989) 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 on an individual. For example, 3711 patients with diabetic retinopathy in both eyes were enrolled in the Early Treatment Diabetic Retinopathy Study (ETDRS 1991a, 1991b) 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 and it is often difficult to choose between available methods. Relatedly, Oakes and Feng (2003) propose tests combining different versions of $PWLR$ tests in cases when PH assumptions hold either within pairs or marginally between groups. 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 paper is organized as follows. In Section 2, the sequential joint limiting distribution of $PWLR$ and $PWKM$ is derived, from which the joint distribution of $PEMAX$ over time is estimated. 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 $PWKM$

Assume that during an accrual period $[0, A]$, 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, where $A < \tau < \infty$. By examining the data repeatedly and systematically, one might detect significant survival differences early on, should they occur. Although in practice pair members usually enter the study simultaneously, this research allows for differential pair member entry times.

Suppose that pair $l = 1, ..., n$, 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}). Although correlation between same pair members' entry times is likely, different pair members' entry times pairs are independent. The same assumptions apply separately to survival and to censoring times. For each pair member, the entry, the survival and the censoring times are assumed to be independent. For technical reasons, we require that the correlation between paired survival times is strictly less than 1.

At analysis (calendar) time t, one observes $\{E_{gl}, X_{gl}(t), \Delta_{gl}(t)\}$, where $X_{gl}(t) = min\{T_{gl}, L_{gl}, max(t - E_{gl}, 0)\}$ and $\Delta_{gl}(t) = I\{T_{gl} \leq min(L_{gl}, t - E_{gl})\}$, representing the follow-up time and the censoring indicator, respectively. If pair l, member g survives past calendar time t without being censored by L_{gl}, then it is considered censored at analysis time t, although this status might change in the future. As usual, pair l, member g is observed as a censored value at analysis time t if L_{gl} occurs prior to both T_{gl} and the time since entry $t - E_{gl}$. Further assume

that within group g, $(E_{gl},\ T_{gl},\ L_{gl}, l = 1, ..., n)$ are *i.i.d.* continuously distributed with survival functions $1 - G_g(e) = P(E_{gl} > e)$, $S_g(s) = P(T_{gl} > s)$ and $C_g(c) = P(L_{gl} > c)$, respectively. Borrowing notation from Murray (2000), the number of pair members $g = 1, 2$ entered by analysis (calendar) time t is equal to $n_g(t) = \sum_{l=1}^{n} I(E_{gl} \leq t)$. The number of 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 is $n_{g_1 g_2}(t_1, t_2) = \sum_{l=1}^{n} I(E_{g_1 l} \leq t_1, E_{g_2 l} \leq t_2)$.

At analysis time t, the marginal cause-specific hazard function for pair member g at study time $0 \leq u \leq t$ is defined as $\lambda_g(u) = \lim_{\delta u \to 0} \frac{1}{\delta u} P\{X_{gl}(t) < u + \delta u, \Delta_{gl}(t) = 1 | X_{gl}(t) \geq u\}$. Using information on pair member g_1 at analysis time t_1 and pair member g_2 at analysis time t_2, define the joint hazard at study time $0 \leq u \leq t_1$ (pertaining to pair member g_1) and study time $0 \leq v \leq t_2$ (pertaining to pair member g_2) as $\lambda_{g_1,g_2}\{(t_1, u), (t_2, v)\} = \lim_{\delta u, \delta v \to 0} \frac{1}{\delta u \delta v} P\{X_{g_1 l}(t_1) < u + \delta u, X_{g_2 l}(t_2) < v + \delta v, \Delta_{g_1 l}(t_1) = 1, \Delta_{g_2 l}(t_2) = 1 | X_{g_1 l}(t_1) \geq u, X_{g_2 l}(t_2) \geq v\}$. Using information on pair member g_1 at analysis time t_1 and pair member g_2 at analysis time t_2, the cause-specific conditional hazard for pair member g_1 at study time $0 \leq u \leq t_1$, given that pair member g_2 is at-risk at study time $0 \leq v \leq t_2$, is defined as $\lambda_{g_1|g_2}\{(t_1, u)|(t_2, v)\} = \lim_{\delta u \to 0} \frac{1}{\delta u} P\{X_{g_1 l}(t_1) < u + \delta u, \Delta_{g_1 l}(t_1) = 1 | X_{g_1 l}(t_1) \geq u, X_{g_2 l}(t_2) \geq v\}$. Let $R_{g_1,g_2}\{(t_1, u), (t_2, v)\}$ $= \lambda_{g_1,g_2}\{(t_1, u), (t_2, v)\} - \lambda_{g_1|g_2}\{(t_1, u)|(t_2, v)\}\lambda_{g_2}(v) - \lambda_{g_2|g_1}\{(t_2, v)|(t_1, u)\}\lambda_{g_1}(u) + \lambda_{g_1}(u)\lambda_{g_2}(v)$, $B_{g_1,g_2}\{(t_1, u), (t_2, v)\} = P\{X_{g_1 l}(t_1) \geq u, X_{g_2 l}(t_2) \geq v | E_{g_1 l} \leq t_1, E_{g_2 l} \leq t_2\} \times [P\{X_{g_1 l}(t_1) \geq u | E_{g_1 l} \leq t_1\} P\{X_{g_2 l}(t_2) \geq v | E_{g_2 l} \leq t_2\}]^{-1}$ and $G_{g_1,g_2} = R_{g_1,g_2} B_{g_1,g_2}$.

The asymptotic proportion $\pi_g(t)$ of pair members $g = 1, 2$ available at analysis time t (among the n pairs that will be accrued by A is estimated by $\hat{\pi}_g(t) = n_g(t) n^{-1}$. For the pair member in group g, the probability $\pi_g(t_1|t_2)$ of study entry by analysis time t_1, given entry by analysis time t_2, where $t_1 \leq t_2$, is consistently estimated by $\hat{\pi}_g(t_1|t_2) = n_g(t_1)\{n_g(t_2)\}^{-1}$. The number of dependent pairs in groups g_1 and g_2 at analysis times t_1 and t_2 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 in group g_1 that have entered by analysis time t_1, among the pairs where the other pair 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 $OR(p)=p(1-p)^{-1}$ be the odds ratio. Define $\psi_{g_1,g_2}(t_1,t_2)=0.5\times\theta_{g_1,g_2}(t_1,t_2)\sqrt{\pi_{3-g_1}(t_1)\pi_{3-g_2}(t_2)}\left[\sqrt{OR(\gamma_{g_1,g_2}(t_1,t_2))}+\sqrt{OR(\gamma_{g_2,g_1}(t_2,t_1))}\right]$. Therefore, an estimator $\hat{\psi}_{g_1,g_2}(t_1,t_2)$ of $\psi_{g_1,g_2}(t_1,t_2)$ is readily available.

At analysis time t, one knows $\{E_{gl},Y_{gl}(t,u),N_{gl}(t,u);0\le u\le t\}$, where $N_{gl}(t,u)=I\{X_{gl}(t)\le u,\Delta_{gl}(t)=1\}$ and $Y_{gl}(t,u)=I\{X_{gl}(t)\ge u\}$, are the failure and the at-risk indicators at study time u based on data available at calendar time t, respectively. $N_g(t,u)$ and $Y_g(t,u)$ are the obvious aggregate versions. Let $M_{gl}(t,u)=N_{gl}(t,u)-\int_0^u Y_{gl}(t,s)\lambda_g(s)ds$ and $M_g(t,u)=\sum_{l=1}^n M_{gl}(t,u)$. For each fixed t, $M_g(t,u)$ is a marginal martingale in u with respect to the filtration containing all survival and censoring information available at analysis time t up to study time u for group g. As an unfortunate consequence of the paired nature of the data, $M_g(t,u)$ is no longer a martingale with respect to the filtration simultaneously containing the above mentioned information from both pair members.

2.1 $PWLR(t)$ and $PWKM(t)$ Test Statistics

At analysis time t, these tests are defined flexibly so that one may test for survival differences over any period of time up to τ by incorporating $J(t,u)=I(0\le u\le\tau)I\{Y_1(t,u)Y_2(t,u)>0\}$ in their integrand. Let $p(t,u)=I(0\le u\le\tau)I[P\{X_{1i}(t)\ge u\}P\{X_{2i}(t)\ge u\}>0]$ for $0\le u\le t$, and assume that $J(t,u)\xrightarrow{\mathcal{P}}p(t,u)$, for all fixed t.

Defined as an integrated weighted difference of the estimated hazard functions, $PWLR(t)=\sqrt{n^*(t)}\int_0^\infty J(t,u)K(t,u)\left[\{Y_1(t,u)\}^{-1}N_1(t,du)-\{Y_2(t,u)\}^{-1}N_2(t,du)\right]$, with $n^*(t)=n_1(t)n_2(t)\times\{n_1(t)+n_2(t)\}^{-1}$, $K(t,u)=\{n^*(t)\}^{-1}W_{pwlr}(t,u)Y_1(t,u)Y_2(t,u)\{Y_1(t,u)+Y_2(t,u)\}^{-1}$ and the weighting function $W_{pwlr}(t,u)$ converging uniformly in probability to a deterministic function $w_{pwlr}(t,u)$ on $[0,t]$. Weights such as $W_{pwlr}(t,u)=1$ or $W_{pwlr}(t,u)=Y_1(t,u)Y_2(t,u)\{n_1(t)n_2(t)\}^{-1}$

yield the paired LR and the paired Gehan test, respectively. By the weak law of large numbers, $\lim_{n_g(t)\to\infty} Y_g(t,u)\{n_g(t)\}^{-1} = S_g(u)H_g(t,u)$, where $H_g(t,u) = P(L_g \geq u, t - E_g \geq u | E_g \leq t)$ is the censoring survival function among group g members entered by analysis time t and S_g is the survivor function in group g. Denote $S_g(u)H_g(t,u)$ by $Q_g(t,u)$. Then, $\lim_{n_1(t),n_2(t)\to\infty}\{Y_1(t,u) + Y_2(t,u)\}\{n_1(t)+n_2(t)\}^{-1} = \sum_{g=1}^{2}\pi_g(t)Q_g(t,u)$ and $k(t,u) = \lim_{n_1(t),n_2(t)\to\infty} K(t,u) = w_{pwlr}(t,u)\times Q_1(t,u)Q_2(t,u)\{\pi_1(t)Q_1(t,u)+\pi_2(t)Q_2(t,u)\}^{-1}$ on $[0,t]$. With Λ_g being the group g true cumulative hazard, arguments as in Appendix A of Lee, Wei and Ying (1993) yield that asymptotically, $PWLR(t) = \sqrt{n^*(t)}\sum_{g=1}^{2}(-1)^{g+1}\int_0^\infty J(t,u)K(t,u)\{Y_g(t,u)\}^{-1}M_g(t,du) + \sqrt{n^*(t)}\int_0^\infty J(t,u)\times K(t,u)\{d\Lambda_1(u) - d\Lambda_2(u)\}$ is equivalent in distribution to $\sqrt{n^*(t)}\sum_{g=1}^{2}(-1)^{g+1}\int_0^\infty p(t,u)k(t,u)\times \{Q_g(t,u)n_g(t)\}^{-1}M_g(t,du) + \sqrt{n^*(t)}\int_0^\infty p(t,u)k(t,u)\{d\Lambda_1(u) - d\Lambda_2(u)\}$. From now on, $PWLR(t)$ will refer to this latter quantity.

Obtained as an integrated weighted difference of the Kaplan-Meier (KM) estimates computed at analysis time t, $PWKM(t) = \sqrt{n^*(t)}\int_0^\infty J(t,u)\hat{W}_{pwkm}(t,u)\{\hat{S}_1(t,u) - \hat{S}_2(t,u)\}du = \sqrt{n^*(t)}\sum_{g=1}^{2}(-1)^g\int_0^\infty J(t,u)\hat{W}_{pwkm}(t,u)\{S_g(u) - \hat{S}_g(t,u)\}du + \sqrt{n^*(t)}\int_0^\infty J(t,u)\hat{W}_{pwkm}(t,u)\times \{S_1(u) - S_2(u)\}du$, where $\hat{S}_g(t,u)$ denotes the KM estimator of $S_g(u)$ obtained using data available at analysis time t. The weighting process $\hat{W}_{pwkm}(t,u)$ converges in probability to a deterministic function $w_{pwkm}(t,u)$ on $[0,t]$. With $\hat{H}_g(t,u)$ being the KM estimator of $H_g(t,u)$ and $\hat{\pi}_g(t) = n_g(t)\{n_1(t)+n_2(t)\}^{-1}$ being the analysis time t sampling proportions, possible $\hat{W}_{pwkm}(t,u)$ choices are $\hat{H}_1(t,u-)\hat{H}_2(t,u-)\{\hat{\pi}_1(t)\hat{H}_1(t,u-) + \hat{\pi}_2(t)\hat{H}_2(t,u-)\}^{-1}$, 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. Again, arguments as in Lee, et. al (1993), yield that $PWKM(t)$ is asymptotically equivalent to the quantity obtained when $p(t,u)$ replaces $J(t,u)$. Lemma 2.4 from Gill (1983) leads to $\int_0^\infty p(t,u)w_{pwkm}(t,u)\{S_g(u) - \hat{S}_g(t,u)\}du = \int_0^\infty A_g(t,u)\hat{S}_g(t,u-)\{S_g(u)Y_g(t,u)\}^{-1}M_g(t,du)$, where $A_g(t,u) = \int_u^\infty p(t,y)w_{pwkm}(t,y)S_g(y)dy$, $0 \leq u \leq t$. With $\hat{S}_g(t,u-)$ estimating $S_g(u)$, $PWKM(t)$ is asymptotically equivalent to

$\sqrt{n^*(t)}\left[\sum_{g=1}^{2}(-1)^g \int_0^\infty A_g(t,u)\{Y_g(t,u)\}^{-1}M_g(t,du) + \int_0^\infty p(t,u)w_{pwkm}(t,u)\{S_1(u)-S_2(u)\}du\right]$.
From now on, $PWKM(t)$ will stand for this equivalent quantity.

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

Results derived in Murray (2000), imply that within pair l, $g_1 \neq g_2$ and $t_i \leq t_j$, $E\{M_{g_1l}(t_i,du)\times M_{g_2l}(t_j,dv)\} = P\{X_{g_1l}(t_i) \geq u | E_{g_1l} \leq t_i\}P\{X_{g_2l}(t_j) \geq v | E_{g_2l} \leq t_j\}G_{g_1,g_2}\{(t_i,u),(t_j,v)\}dvdu$. Given the marginal martingale structure for group g, the use of the theory of stochastic integrals with respect to martingales facilitates the derivation of parts of the asymptotic variances-covariances for $PWLR(t_i)$ and $PWKM(t_j)$. Results from Gu and Lai (1991) indicate that $E\{M_g(t_i,du)M_g(t_j,du)\} = Y_g(t_i,u)\lambda_g(u)du$. These key results, as well as the multivariate central limit theorem, are used to compute the entries of $\mathbf{\Sigma}$ that have one of the following forms: $cov\{PWLR(t_i), PWLR(t_j)\}$, $cov\{PWKM(t_i), PWKM(t_j)\}$, $cov\{PWLR(t_i), PWKM(t_j)\}$ or $cov\{PWKM(t_i), PWLR(t_j)\}$, where $1 \leq i \leq j \leq D$.

If we let $\eta_{g,3-g}(t_i,t_j) = \sqrt{\pi_{3-g}(t_i)\pi_{3-g}(t_j)\pi_g(t_i|t_j)}$ and $p(t,u)k(t,u) = r(t,u)$, computations lead to, $cov\{PWLR(t_i), PWLR(t_j)\} = \sum_{g=1}^{2}[\eta_{g,3-g}(t_i,t_j)\int_0^\infty r(t_i,u)r(t_j,u)\{Q_g(t,u)\}^{-1}\lambda_g(u)du - \psi_{g,3-g}(t_i,t_j)\int_0^\infty\int_0^\infty r(t_i,u)r(t_j,v)G_{g,3-g}\{(t_i,u),(t_j,v)\}dvdu]$.

Also, $cov\{PWKM(t_i), PWKM(t_j)\} = \sum_{g=1}^{2}[\eta_{g,3-g}(t_i,t_j)\int_0^\infty A_g(t_i,u)A_g(t_j,u)\{Q_g(t_j,u)\}^{-1}\times \lambda_g(u)du - \psi_{g,3-g}(t_i,t_j)\int_0^\infty\int_0^\infty A(t_i,u)A(t_j,v)G_{g,3-g}\{(t_i,u),(t_j,v)\}dvdu]$.

Similarly, $cov\{PWLR(t_i), PWKM(t_j)\} = \sum_{g=1}^{2}[-\eta_{g,3-g}(t_i,t_j)\int_0^\infty r(t_i,u)A_g(t_j,u)\{Q_g(t_j,u)\}^{-1}\times \lambda_g(u)du + \psi_{g,3-g}(t_i,t_j)\int_0^\infty\int_0^\infty r(t_i,u)A_{3-g}(t_j,v)G_{g,3-g}\{(t_i,u),(t_j,v)\}dvdu]$.

Finally, $cov\{PWKM(t_i), PWLR(t_j)\} = \sum_{g=1}^{2}[-\eta_{g,3-g}(t_i,t_j)\int_0^\infty A_g(t_i,u)r(t_j,u)\{Q_g(t_j,u)\}^{-1}\times \lambda_g(u)du + \psi_{g,3-g}(t_i,t_j)\int_0^\infty\int_0^\infty A_g(t_i,u)r(t_j,v)G_{g,3-g}\{(t_i,u),(t_j,v)\}dvdu]$.

Thus, one may now obtain quantiles for any functional of $PWLR$ and $PWKM$ by means of Markov Chain Monte Carlo simulations. Suppose that in a hypothetical trial with D interim analyses, one wishes to obtain the group sequential stopping boundaries (GSSB) for a generic functional F. At the ith interim analysis, the quantity F, which we will refer to as F_i, $i = 1, \ldots, D$, may be the maximum or a linear combination of the absolute values of the standardized statistics. First, one generates N replications of a $2D$-dimensional zero-mean normal random vector with covariance matrix equal to that of $2D$ standardized $PWLR$ and $PWKM$ tests from the D interim analyses. For each such replication, compute F_i and its stopping boundary at analysis $i = 1, \ldots, D$, respecting the prespecified type I error α_i. Specifically, the boundary of F_1 at analysis time 1 is the α_1-th upper quartile of the N-dimensional vector of F_1 values. At analysis time $k = 2, \ldots, D$ form the vector of F_k values for which all of F_j values, $j = 1, 2, ..., k-1$, did not exceed their corresponding analysis time j GSSB. The k-th interim analysis GSSB for F_k is the $\{\alpha_k(1 - \sum_{j=1}^{k-1} \alpha_j)^{-1}\}$th upper quartile of this latter vector.

2.2 Estimation of Σ

Let $Y_{g_1,g_2}\{(t_i, u), (t_j, v)\} = \sum_{l=1}^{n_{g_1,g_2}(t_i,t_j)} I\{X_{g_1l}(t_i) \geq u, X_{g_2l}(t_j) \geq v\}$ 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 $N_{g_1,g_2}\{(t_i, du), (t_j, dv)\} = \sum_{l=1}^{n_{g_1,g_2}(t_i,t_j)} I\{X_{g_1l}(t_i) \in [u, u + du), X_{g_2l}(t_j) \in [v, v + dv), \Delta_{g_1l}(t_i) = 1, \Delta_{g_2l}(t_j) = 1\}$ 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 equal to $N_{g_1|g_2}\{(t_i, du)|(t_j, v)\} = \sum_{l=1}^{n_{g_1,g_2}(t_i,t_j)} I\{X_{g_1l}(t_i) \in [u, u + du), X_{g_2l}(t_j) \geq v, \Delta_{g_1l}(t_i) = 1\}$. An estimator for $P\{X_g(t_i) \geq u | E_g \leq t_i\}$ is $Y_g(t_i, u)\{n_g(t_i)\}^{-1}$, while $P\{X_{g_1}(t_i) \geq u, X_{g_2}(t_j) \geq v | E_{g_1} \leq t_i, E_{g_2} \leq t_j\}$ 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 $\lambda_g(u)du$, $\lambda_{g_1,g_2}\{(t_i,u),(t_j,v)\}dudv$ and $\lambda_{g_1|g_2}\{(t_i,u)|(t_j,v)\}du$ are available through $\{Y_g(t_j,u)\}^{-1}N_g(t_j,du)$, $N_{g_1,g_2}\{(t_i,du),(t_j,dv)\}\left[Y_{g_1,g_2}\{(t_i,u),(t_j,v)\}\right]^{-1}$ and $N_{g_1|g_2}\{(t_i,du)|(t_j,v)\}\left[Y_{g_1,g_2}\{(t_i,u),(t_j,v)\}\right]^{-1}$, respectively.

To estimate quantities involving terms not dependent on analysis time, such as $\lambda(u)$ or $S(u)$, given data availability at both $t_i < t_j$, one may use information available at t_j. For example, $A_g(t_i,u)$ can be estimated using $\int_u^\infty J(t_i,y)\hat{W}_{pwkm}(t_i,y)\hat{S}_g(t_j,y)dy$, where $\hat{S}_g(t_j,u)$ is the KM estimator of $S_g(u)$ based on data collected at the latter analysis time t_j. Note that the weighting terms in this expression still use only information available at the earlier analysis time t_i, since they are analysis time-dependent terms. A consistent estimator of $G_{g,3-g}\{(t_i,u),(t_j,v)\}dvdu$ can be obtained based on estimators of its components and $\hat{H}_g(t,u)$ is the KM estimator of $H_g(t,u)$. Thus, an estimator of the covariance matrix $\mathbf{\Sigma}$ is now available.

3 Simulation Study

Simulations are conducted 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 this latter case, we denote the statistics of interest by $EMAX$ rather than $PEMAX$. Both under the null and the alternative hypotheses, 1,000 Monte Carlo simulation runs consisting of 100 pairs of correlated piecewise exponential survival times are generated with correlation of approximately 0.25. We assume a common pair entry time to be uniform$(0, 0.25)$ years and conduct interim analyses at 0.5, 1 and 2 years. 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. Under either hypothesis, the hazard rates in both groups change at 1 and 1.65 years, with dynamics of the hazards further described in Figure 1.

To the best of our knowledge, in paired censored survival settings, there is no theoretical result to indicate a locally most powerful test. This particular formulation of stochastic ordering gives approximately PH at the early interim analyses, t_1 and t_2, and crossing hazards at the last

interim analysis t_3, where most of the type I error is spent (see Figure 1). We chose $PYLS$ to represent the $PWKM$ family because of its simple interpretation as the number of years-of-life-saved while on study, with the weighting of the area between the survival curves not involving the distribution of the censoring times. With differential weighting of the area between curves at analysis times, interpretation may be problematic, although the theory accommodates such choices. Also of major interest is to understand how correctly accounting for the paired nature of the data improves the operating characteristics of $PEMAX$.

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 over-conservativeness for all three tests. In the example simulated, $PYLS$ is expected to be more powerful than the PLR as the PH feature is lost over time, so it will be used as the reference test in order to describe the power loss percentages exhibited by the other tests. As expected, the paired versions of the tests observed are all more powerful that any of the those that ignore pairing. Using $PEMAX$, a 3.79% power gain over the disadvantaged test PLR is observed. In our experience, a similar phenomenon occurs as the hazard rates remain more proportional across all interim analyses, favoring the LR testing framework. That is, minimal losses of power using $PEMAX$ as opposed to the more powered test. Hence, the more robust $PEMAX$ is close in power when compared to the more powered of PLR and $PYLS$, when the alternative hypothesis is in doubt. Ignoring pairing induces more serious power losses, with the largest loss of 15.83% associated with the unpaired LR test as compared to the 8.69% loss when using the unpaired $EMAX$ and the 6.35% loss when using the unpaired YLS test.

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 end-point 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 15 February 1983 who were taking a placebo pill in a separate randomization process. Since 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 the data make this example suitable for analysis using group sequential methods and $PEMAX$ will be employed. A number of 9 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).

The GSSB for paired $PEMAX$ are then obtained via the algorithm described in Section 2.2, by producing 30,000 replications of an 18-dimensional zero-mean normal random vector, whose covariance matrix is that of the standardized PLR and $PYLS$ tests computed at each of the 9 interim analyses. Similarly, the GSSB for the unpaired $EMAX$ are obtained based on the standardized unpaired LR and YLS tests instead.

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

Unlike with independent groups, once correlation is involved it is hard to know which test is most powerful under PH or non-PH alternatives. One approach is to consider several tests simultaneously, thus covering more situations. Work by Oakes and Feng (2003) that combines stratified and unstratified PLR tests, suggests that the strength of within pair correlation, rather than the form of the alternative, may sometimes determine the more powerful test.

The newly proposed test, $PEMAX$, has several features that distinguish it from the individual tests. Although rank-based 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.

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 over-conservativeness 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 is paired.

ACKNOWLEDGEMENT

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.

REFERENCES

Chi, Y., Tsai, M. H. (2001). Some versatile tests based on the simultaneous use of weighted log-rank 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. *Mathematical Centre Tracts* **124**, Mathematisch Centrum, Amsterdam.

Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *Annals of Statistics* **11**, 49-58.

Gu, M. G., 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., Turnbull, B.W. (2000). Group sequential methods with applications to clinical trials. *CRC Press Inc.* (Boca Raton, FL)

Kosorok, M. R., 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., DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

Lee, E. W., Wei, L. J., Ying, Z. (1993). Linear regression analysis for highly stratified failure time data. *Journal of the American Statistical Association* **88**, 557-565.

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

Mantel, N. (1996). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163-170.

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.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487-493.

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

Oakes, D., Feng, C. (2003). Combining Stratified and Unstratified Log-Rank Tests for Matched Pairs Survival Data. *International Conference on Reliability and Survival Analysis 2003 (ICRSA 2003) Abstracts Booklet*, 7-8, University of South Carolina, Columbia, SC.

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

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

Pepe, M., Fleming, T. R. (1989). Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data. *Biometrics* **45**, 497-507.

Sellke, T., 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. *Survival Analysis*, eds. Crowley, J. and Johnson, R.A., **Vol. 2**, 257-268.

Table 1

Size and power simulation results for paired and unpaired YLS, LR and $EMAX$ based on 1,000 replications with an overall type I error $\alpha = 0.05$. Power loss results (in percentages) are relative to the paired $PYLS$, which is more powered in this case.

	Paired			**Unpaired**		
Test	PYLS	PLR	PEMAX	YLS	LR	EMAX
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%

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 15 February 1983 that are taking a placebo pill

		Paired			**Unpaired**		
Analysis Time	Error spent	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

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 15 February 1983 that are taking an aspirin pill

		Paired			**Unpaired**		
Analysis Time	Error spent	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 $\leftarrow$	$2.34 * 10^{-3}$	-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

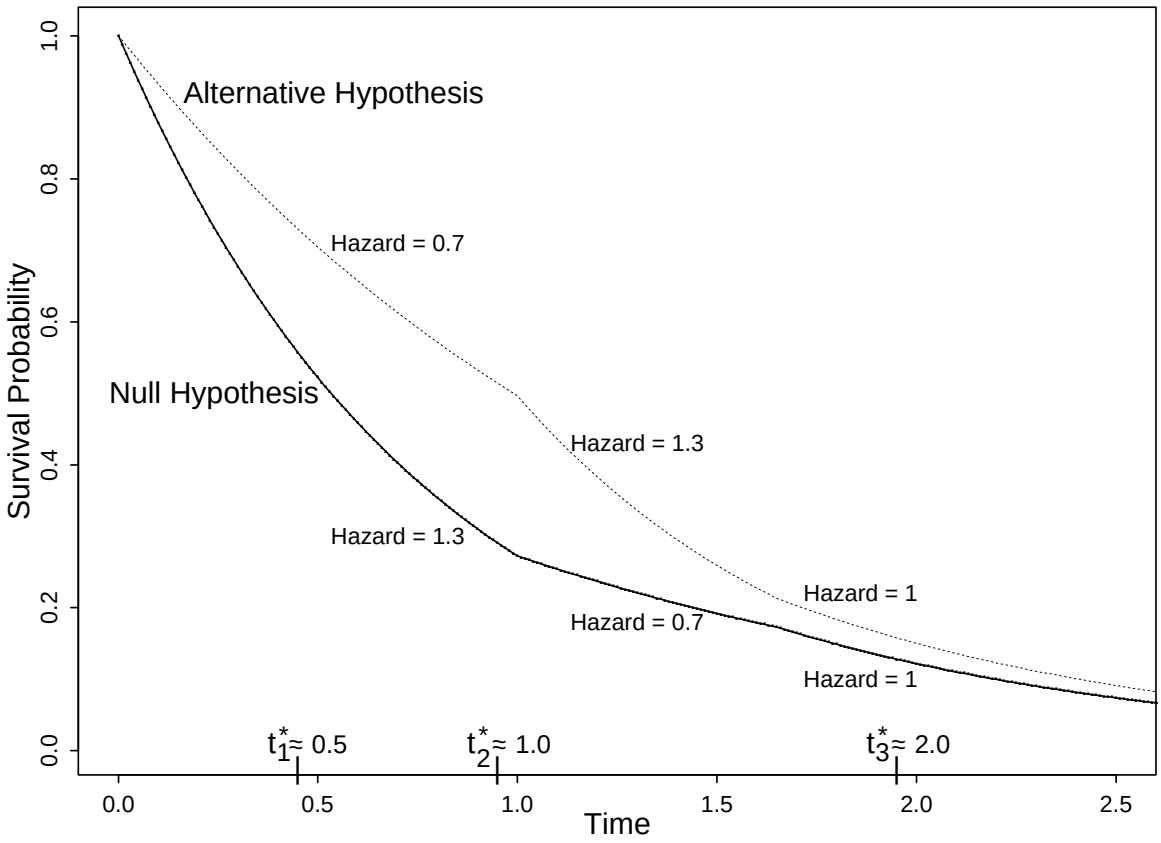

Figure 1: Null and alternative hypothesis simulation scenarios survival curves with superimposed hazard rates, where t_i^* indicates the configuration of the hazards likely to be observed prior to t_i at interim analysis $i = 1, 2, 3$, respectively.