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 (1999), among others, and sequential methods for comparing areas under survival curves were developed by Murray and Tsai (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 \( \tau \). 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 \( g = 1, 2 \), and \( t \) (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 \( P(E_{gl} > e) \), \( S_l(s) = P(T_{gl} > s) \), and \( C_l(c) = P(L_{gl} > c) \), respectively. For technical reasons, the correlation between paired survival times is assumed to be strictly less than 1.

At analysis time \( t \), \( n_g(t) = \sum_{i=1}^{n} I(E_{gi} \leq t) \) pair members \( g \) have entered the study, while \( X_{gi}(t) = \min(T_{gi}, L_{gi}, \max(t - E_{gi}, 0)) \) and \( \Delta_{gi}(t) = I(T_{gi} \leq \min(L_{gi}, t - E_{gi})) \) are the follow-up time and the censoring indicator, respectively; for pair \( l \), member \( g \). There are \( n_{g|1}(t_1, t_2) = \sum_{i=1}^{n_g(t)} I(E_{gi} \leq t_1, L_{gi} \leq t_2) \) pairs whose member \( g \) has entered the study by analysis time \( t_1 \) and member \( g_1 \) has entered by analysis time \( t_2 \). Let \( N_g(t, u) = \sum_{i=1}^{n_g(t)} I(X_{gi}(t) \leq u, \Delta_{gi}(t) = 1) \) and \( Y_g(t, u) = \sum_{i=1}^{n_g(t)} I(X_{gi}(t) \geq u) \) 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

\[
J(t, u) = I(0 \leq u \leq \tau)I(Y_1(t, u)Y_2(t, u) > 0),
\]

\[
p(t, u) = I(0 \leq u \leq \tau)I[P(X_{gi}(t) \geq u)P(X_{gi}(t) \geq u) > 0]
\]

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

At time \( t \),

\[
PWL(t) = \sqrt{n(t)} \int_{0}^{\infty} J(t, u)K(t, u)
\]

\[
\times \sum_{g=1}^{2} (-1)^{g+1} \{Y_g(t, u)\}^{-1}N_g(t, du),
\]

with

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

\[
K(t, u) = W_{pwlb}(t, u) \left\{ \prod_{g=1}^{2} Y_g(t, u) \right\} \left\{ \sum_{g=1}^{2} Y_g(t, u) \right\}^{-1},
\]

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

Also at time \( t \), \( PWKM(t) = n(t)\sqrt{t} \int_{0}^{\infty} J(t, u)\hat{W}_{pk}(t, u) \times \sum_{g=1}^{2} (-1)^{g+1} \hat{S}_g(t, u) du \), where \( \hat{S}_g(t, u) \) is the Kaplan–Meier (KM) estimator of \( S_g(u) \) based on \( t \) data and the weighting process \( \hat{W}_{pk}(t, u) \) converges in probability to a deterministic function \( w_{pk}(t, u) \) on \([0, \tau]\) with \( \hat{S}_g(t, u) = \sum_{i=1}^{n(t)} \hat{S}_g(t, u) \{\sum_{j=1}^{n(t)} \hat{S}_g(t, u)\}^{-1} \), where in the limiting distribution of the weighted function recommended by Pepe and Fleming (1989), or alternatively \( \hat{W}_{pk}(t, u) = 1 \), interpreted as paired years-of-life saved (PYLS) over \( \tau \) years of study.

For stochastically ordered survival curves, the null hypothesis is \( H_0: S_1(\cdot) = S_2(\cdot) = S(\cdot) \) on \([0, \tau] \). If \( t_1 < t_2 < \cdots < t_0 \) are successive analysis times such that the statistical information expected between them is sufficient to warrant additional analyses, then it follows that \( \{PWLR(t_1), PWKM(t_1), \ldots, PWLR(t_0), PWKM(t_0)\}^{T} \xrightarrow{D} N_{2D}(0_{2D}, \Sigma) \).

The covariance matrix \( \Sigma \) 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 \( U(0, 1) \), and survival times \((T_1, T_2)\) are correlated survival times generated from \( \text{AFT} \) models \( \log(T_g) = \mu_0 + 0.5Z_1 + 0.25W_g, g = 1, 2 \), 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 \( h_g(t \mid X_1, Y_1) = h_{0,g} \exp(\alpha_t X_1 + \beta Y_1), g = 1, 2 \), where \( h_{0,1} = 0.05, h_{0,2} = 0.02857, \alpha_1 = 0.2, \beta_1 = 0.2, \alpha_2 = 0.3, \beta_2 = 0.6 \), \( X_1, X_2 \) are independently \( \Gamma(0.2, 1) \), \( Y_1, Y_2 \) 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 \( T_g = -\left(h_{0,g}(t \mid X_1, Y_1)\right)^{-1} \times \log(1 - U_g) \), 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.

### Table 1

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

*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.

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...
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 who 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 2

**Paired and unpaired versions of (P)YLS and (P)L.R 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**

<table>
<thead>
<tr>
<th>Analysis time</th>
<th>Error spent</th>
<th>Paired</th>
<th>Unpaired</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PYLS PLR PEMAXb</td>
<td>YLS LR EMAXb</td>
</tr>
<tr>
<td>1</td>
<td>$2.85 \times 10^{-5}$</td>
<td>$-2.119$ $1.900$ $4.453$</td>
<td>$-1.169$ $1.441$ $4.051$</td>
</tr>
<tr>
<td>2</td>
<td>$1.42 \times 10^{-4}$</td>
<td>$-2.530$ $2.340$ $4.031$</td>
<td>$-1.970$ $1.810$ $3.663$</td>
</tr>
<tr>
<td>3</td>
<td>$5.74 \times 10^{-4}$</td>
<td>$-3.060$ $3.006$ $3.517$</td>
<td>$-2.453$ $2.437$ $3.287$</td>
</tr>
<tr>
<td>4</td>
<td>$1.18 \times 10^{-3}$</td>
<td>$-2.846$ $3.006$ $3.320$</td>
<td>$-2.727$ $2.472$ $3.106$</td>
</tr>
<tr>
<td>5</td>
<td>$1.31 \times 10^{-3}$</td>
<td>$-2.482$ $2.674$ $3.197$</td>
<td>$-1.928$ $2.165$ $2.940$</td>
</tr>
<tr>
<td>6</td>
<td>$2.34 \times 10^{-3}$</td>
<td>$-2.700$ $2.653$ $2.988$</td>
<td>$-2.095$ $2.140$ $2.722$</td>
</tr>
<tr>
<td>7</td>
<td>$1.33 \times 10^{-3}$</td>
<td>$-2.716$ $2.423$ $2.996$</td>
<td>$-2.074$ $1.909$ $2.715$</td>
</tr>
<tr>
<td>8</td>
<td>$2.27 \times 10^{-3}$</td>
<td>$-3.106$ $2.828$ $2.892$</td>
<td>$-2.412$ $2.284$ $2.590$</td>
</tr>
<tr>
<td>9</td>
<td>$8.29 \times 10^{-4}$</td>
<td>$-3.179$ $2.886$ $2.918$</td>
<td>$-2.490$ $2.348$ $2.585$</td>
</tr>
</tbody>
</table>

Bold font indicates that the PEMAXb boundary has been exceeded.

### Table 3

**Paired and unpaired versions of (P)L.R 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**

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

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


APPENDIX

The asymptotic proportion \( \pi_g(t) \) of pair members \( g \) entered by analysis time \( t \) is estimated by \( \hat{\pi}_g(t) = n_g(t) \left( \sum_{h=1}^{\pi_g} n_h(t) \right)^{-1} \).
For 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 $\hat{\pi}_g(t_1 | t_2) = n_g(t_1) / \{n_g(t_1) + n_{v-g}(t_1)\}$. 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 $\hat{\theta}_{g_1,g_2}(t_1, t_2) = 2n_{g_1,g_2}(t_1, t_2) / \{n_g(t_1) + n_{v-g}(t_1)\}$. 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 $\hat{\gamma}_{g_1,g_2}(t_1, t_2) = n_{g_1}(t_1) / \{n_g(t_1) + n_{v-g}(t_1)\}$. 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) / \{\pi_{g_1}(t_1)\pi_{g_2}(t_2)\}^{1/2} \{OR(\gamma_{g_1,g_2}(t_1, t_2))\}^{1/2}.\]

The marginal cause-specific hazard function for pair member $g$ at study time $0 \leq u \leq t$ is defined as $\lambda_g(u) = \lim_{\Delta u \to 0} \frac{1}{\Delta u} P(X_g(t) < u + \Delta u, \Delta_g(t) = 1 | X_g(t) \geq u)$. Using information on pair member $g_1$ at analysis time $t_1$ and pair member $g_2$ at analysis time $t_2$, the joint hazard at study time $0 \leq u \leq t$ for member $g_1$ and study time $0 \leq v \leq t_2$ for member $g_2$ is $\lambda_{g_1,g_2}(t_1, t_2, v) = \lim_{\Delta v \to 0} \frac{1}{\Delta v} P(X_g(t_1) < u + \Delta u, X_{g_2}(t_2) < v + \Delta v, \Delta_{g_1}(t_1) = 1, \Delta_{g_2}(t_2) = 1 | X_{g_1}(t_1) \geq u, X_{g_2}(t_2) \geq v)$. Let $R_{g_1,g_2}(\{t_1, u\}, \{t_2, v\}) = \lambda_{g_1|g_2}(\{t_1, u\}, \{t_2, v\})\lambda_{g_2}(v)$ $- \lambda_{g_1|g_2}(\{t_2, v\}, \{t_1, u\})\lambda_g(u) + \lambda_g(u)\lambda_{g_2}(v)$, $B_{g_1,g_2}(\{t_1, u\}, \{t_2, v\}) = P(X_g(t_1) \geq u, X_{g_2}(t_2) \geq v, E_g(t_1) \leq t_1, E_{g_2}(t_2) \leq t_2)$ $\times P(X_g(t_1) \geq u, E_{g_1}(t_1) \leq t_1)P(X_{g_2}(t_2) \geq v, E_{g_2}(t_2) \leq t_2)^{-1}$, and $G_{g_1,g_2} = R_{g_1,g_2}B_{g_1,g_2}$. $\xi(t_1, t_2)$ denotes $S(t_1)S(t_2)$ and, as a reminder, $H_g(t, u)$ denotes $P(\xi \geq u, t - E_g \geq u | E_g \leq t)$. For $0 \leq u \leq t$, define $A_g(t, u) = \int_0^u p(t, y)w_{puln}(t, y)S_g(y)dy$. A.1 Description of $\Sigma$

The entries of $\Sigma$ are of the form $\text{cov}\{PWLR(t_1), PWLR(t_2)\}$, $\text{cov}\{PWKM(t_1), PWKM(t_2)\}$, $\text{cov}\{PWLR(t_1), PWKM(t_2)\}$, or $\text{cov}\{PWKM(t_1), PWLR(t_2)\}$, where $1 \leq i \leq j \leq D$. If we let $\eta_{g_3,g}(t_1, t_2) = (\pi_{g_3-g}(t_1)\pi_{g_3-g}(t_2)\pi_{g_3}(t_1, t_2))^{1/2}$ and $p(t, u)k(t, u) = r(t, u)$, results detailed in a technical report by Andrei and Murray (2004) lead to $\text{cov}\{PWLR(t_1), PWLR(t_2)\} = \sum_{g_1} \eta_{g_3,g}(t_1, t_2)g_3 r(t_1, u)Q_g(t_1, u)^{-1} \lambda_g(u)du - \psi_{g_3,g}(t_1, t_2)\int_0^\infty r(t_1, u)r(t_2, v)G_{g_3,g}(t_1, u, \{t_2, v\})dvdu$. Also,$\text{cov}\{PWKM(t_1), PWKM(t_2)\} = \sum_{g_1} \eta_{g_3,g}(t_1, t_2)\int_0^\infty A_g(t_1, u)A_g(t_2, u)du - \psi_{g_3,g}(t_1, t_2)\int_0^\infty \int_0^\infty A(t_1, u)A(t_2, v)G_{g_3,g}(t_1, u, \{t_2, v\})dvdu$.

Similarly,$\text{cov}\{PWLR(t_1), PWLR(t_2)\} = \sum_{g_1} \eta_{g_3,g}(t_1, t_2)\int_0^\infty r(t_1, u)A_g(t_2, u)du - \psi_{g_3,g}(t_1, t_2)\int_0^\infty \int_0^\infty r(t_1, u)r(t_2, v)G_{g_3,g}(t_1, u, \{t_2, v\})dvdu$.

Finally,$\text{cov}\{PWKM(t_1), PWLR(t_2)\} = \sum_{g_1} \eta_{g_3,g}(t_1, t_2)\int_0^\infty A_g(t_1, u)r(t_2, u)du - \psi_{g_3,g}(t_1, t_2)\int_0^\infty \int_0^\infty A(t_1, u)r(t_2, v)G_{g_3,g}(t_1, u, \{t_2, v\})dvdu$.

A.2 Estimation of $\Sigma$

Let $Y_{g_1,g_2}(\{t_1, u\}, \{t_2, v\})$ be the number of pairs in which, at analysis time $t_1$, pair member $g_1$ is-at-risk at study time $u$ and at analysis time $t_2$, pair member $g_2$ is-at-risk at study time $v$. Define $N_{g_1,g_2}(\{t_1, du\}, \{t_2, dv\})$ to be the number of pairs in which group $g_1$ member, who has entered by analysis time $t_1$, fails at study time $u$ and group $g_2$ member, who has entered by analysis time $t_2$, fails at study time $v$. Finally, the number of pairs for which the group $g_1$ member, who has entered by analysis time $t_1$, is still-at-risk at study time $v$ is denoted by $N_{g_1|g_2}(\{t_1, du\}, \{t_2, v\})$. An estimator for $P(X_g(t_1) \geq u | E_g \leq t_1)$ is $Y_{g_1}(t_1, u)\times N_{g_1}(t_1) - 1$, while $P(X_g(t_2) \geq v | E_g \leq t_2, E_g \leq t_2)$ is estimable by $Y_{g_1,g_2}(\{t_1, u\}, \{t_2, v\})\{n_{g_1,g_2}(t_1, t_2)\}^{-1}$. Nelson–Aalen-type estimators of $\lambda_g(u)du$, $\lambda_{g_1,g_2}(\{t_1, u\}, \{t_2, v\})dvdu$, and $\lambda_{g_1|g_2}(\{t_1, u\}, \{t_2, v\})dvdu$ are available through $\{Y_g(t_1, u)^{-1}N_{g}(t_1, du), N_{g_1|g_2}(\{t_1, du\}, \{t_2, dv\})\times \{X_{g_1,g_2}(\{t_1, u\}, \{t_2, v\})\}^{-1}, and N_{g_1|g_2}(\{t_1, du\}, \{t_2, dv\})\}Y_{g_1,g_2}(\{t_1, u\}, \{t_2, v\})^{-1}$, respectively. A consistent estimator of $G_{g_3,g}(t_1, \{t_2, v\})dvdu$ can be obtained based on previously described estimators of its components. Thus, an estimator of the covariance matrix $\Sigma$ is now available. More details can be found in Andrei and Murray (2004).