BIOMETRICS 57, 361–368
June 2001

Using Weighted Kaplan–Meier Statistics in Nonparametric Comparisons of Paired Censored Survival Outcomes

Susan Murray

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

SUMMARY. This research introduces methods for nonparametric testing of weighted integrated survival differences in the context of paired censored survival designs. The current work extends work done by Pepe and Fleming (1989, *Biometrics* **45,** 497–507), which considered similar test statistics directed toward independent treatment group comparisons. An asymptotic closed-form distribution of the proposed family of tests is presented, along with variance estimates constructed under null and alternative hypotheses using nonparametric maximum likelihood estimates of the closed-form quantities. The described method allows for additional information from individuals with no corresponding matched pair member to be incorporated into the test statistic in sampling scenarios where singletons are not prone to selection bias. Simulations presented over a range of potential dependence in the paired censored survival data demonstrate substantial power gains associated with taking into account the dependence structure. Consequences of ignoring the paired nature of the data include overly conservative tests in terms of power and size. In fact, simulation results using tests for independent samples in the presence of positive correlation consistently undershot both size and power targets that would have been attained in the absence of correlation. This additional worrisome effect on operating characteristics highlights the need for accounting for dependence in this popular family of tests.

KEY WORDS: Clinical trial; Correlated times-to-event; Two-sample test; Years-of-life-saved statistic.

1. Introduction

Paired censored survival data arise in a variety of clinical trial settings. For instance, one of the primary goals of the Early Treatment Diabetic Retinopathy Study (ETDRS, 1991a,b) was to determine the best time in the course of diabetic retinopathy to initiate laser photocoagulation surgery. This study enrolled 3711 patients with mild to severe nonproliferative or early proliferative diabetic retinopathy in both eyes and randomized alternate eyes to early photocoagulation or deferral of photocoagulation until such time when high-risk proliferative retinopathy was detected. The major endpoint of interest was time to severe vision loss, where this loss was defined as visual acuity less than 5/200 at two consecutive visits. Even after 9 years of follow-up, the study had a large proportion (94.5%) of censored outcomes among the 3711 pairs, making standard paired tests for uncensored data inappropriate.

More complex censored survival data structures have arisen in dental research on sealants done at the University of Michigan School of Dentistry (Feigal et al., 2000). Dental scientists have long endorsed the use of sealants as a measure for preventing dental caries, especially in the occlusal or grinding surface of molars, where teeth are most susceptible. The greatest risk of sealant failure occurs in newly erupted teeth, where moisture threatens the sealant process. Hence, researchers in the area of improving sealant protection enroll patients with newly erupted paired molars in the lower jaw for randomization to opposing sealant application treatments and measure the time to sealant failure. Occasionally, only one of the target molars in a patient has erupted at the time they come into the study so that singleton molars are occasionally randomized to treatment. Should the companion molar erupt during the course of the study, it could also potentially be entered into the protocol with a different entry time than its counterpart, resulting in differential censoring within a matched pair.

Several authors, including Woolson and Lachenbruch (1980), Wei (1980), O'Brien and Fleming (1987), Dabrowska (1989, 1990), Jung (1999), and Huang (1999), have presented work in relation to nonparametric testing for survival differences in paired censored survival data using a variety of sign-based or rank-based tests. Related rank-based tests have been developed by Holt and Prentice (1974) and Lee, Wei, and Amato (1992) in the context of paired proportional hazards models. Murray (2000) also studied paired weighted log-rank tests in the context of sequential monitoring of paired censored survival data. Woolson and O'Gorman (1992) provide a useful summary of much of this literature.

An alternative to rank-based methods, advocated by Pepe and Fleming (1989), looks at differences between integrated

weighted survival curves. The resulting test statistic, which was developed in the context of independent samples, has been shown to have higher power to detect survival differences than rank-based methods when underlying hazards cross during the study period. When hazards are proportional in nature, the power of the Pepe–Fleming test, while not always as high as rank-based tests, is often comparable. Investigators are particularly attracted to this test statistic when unweighted survival curves are integrated across the study period due to a related interpretation of average years of life saved on study using the superior treatment.

This research presents a nonparametric method for detecting survival differences in paired censored survival data using differences in weighted integrated survival curves as in Pepe and Fleming (1989). The related family of tests presented here takes into account the dependence between estimated survival curves that tend to vary in tandem in the presence of positively correlated data. The proposed family of test statistics will allow different and potentially dependent censoring distributions for pair members under comparison and will allow for singleton pair members to contribute information to the test statistic in the absence of a counterpart. In addition, this work will make available confidence intervals relating to the average improvement in survival time on study between treatment groups under comparison, an interpretable and useful measure in describing treatment benefit to nonstatisticians.

In Section 2, notation relating to this data structure is presented and the paired Pepe–Fleming statistic is introduced. Associated closed-form asymptotic variances for the family of tests, which may be estimated using nonparametric maximum likelihood estimates of the closed-form quantities, are presented under null and alternative hypotheses. Simulation results in Section 3 verify type I error operating characteristics and provide evidence that power grows with the dependence between paired endpoints. In addition, some consequences of ignoring the paired nature of the data are highlighted in this section. An example relating to the ETDRS is located in Section 4. A discussion follows in Section 5.

2. Paired Pepe–Fleming Statistics

Consider survival endpoints T_{ik_i} and corresponding censoring times U_{ik_i}, indexed by treatment group $i = 1, 2$ and individual $k_i = 1, \ldots, n_i$. An arbitrary dependence is allowed between T_{1k_1} and T_{2k_2} and between U_{1k_1} and U_{2k_2} for $k_1 = k_2 = k$, $k = 1, \ldots, n$, where $n \leq \min(n_1, n_2)$; i.e., the first n individuals from each group are allowed, and even expected, to be correlated to their counterpart in the other group. Aside from this dependence, the random variables T_{ik_i}, U_{ik_i}, $i = 1, 2$, $k_i = 1, \ldots, n_i$, are assumed independent. If all study participants belong to a complete matched pair, as in the ETDRS study, then $n_1 = n_2 = n$. However, the more general case is permitted where the individuals indexed with $k_i > n$ remain unpaired, a circumstance that occasionally arises, as in the cited dental research application. Marginally, the various random variables, $T_{1k_1}, T_{2k_2}, U_{1k_1}$, and U_{2k_2}, are assumed to be independent and identically distributed across the $k = 1 \ldots n_i$ individuals. Different distributions are allowed for the failure and censoring random variables according to treatment. Pair members where there is an absence of information on either the failure or the censoring time are assumed to be missing completely at random (MCAR).

Define the observable event times as $X_{ik_i} = \min(T_{ik_i}, U_{ik_i})$ with corresponding censoring indicators $\Delta_{ik_i} = I(T_{ik_i} < U_{ik_i})$. Let $S_i(t)$ denote the survival function relating to T_{ik_i}, $k_i = 1, \ldots, n_i$, in group i at time t and let $\hat{S}_i(t)$ denote its Kaplan–Meier estimate. Similarly, let $\lambda_i(u)$ denote the hazard for failure in treatment group i. Because of the dependence allowed within the n complete pairs, we will also require joint and conditional hazards to be defined for the paired random variables as they appear in the asymptotic closed-form variances to be derived. Define the joint and conditional hazards $\lambda_{12}(t_1, t_2) = \lim_{\Delta t_1, \Delta t_2 \to 0} P(t_1 \leq X_{1k} < t_1 + \Delta t_1, t_2 \leq X_{2k} < t_2 + \Delta t_2, \Delta_{1k} = 1, \Delta_{2k} = 1 \mid X_{1k} \geq t_1, X_{2k} \geq t_2)/(\Delta t_1 \Delta t_2)$, $\lambda_{1|2}(t_1 \mid t_2) = \lim_{\Delta t_1 \to 0} P(t_1 \leq X_{1k} < t_1 + \Delta t_1, \Delta_{1k} = 1 \mid X_{1k} \geq t_1, X_{2k} \geq t_2)/\Delta t_1$, and $\lambda_{2|1}(t_2 \mid t_1) = \lim_{\Delta t_2 \to 0} P(t_2 \leq X_{2k} < t_2 + \Delta t_2, \Delta_{2k} = 1 \mid X_{1k} \geq t_1, X_{2k} \geq t_2)/\Delta t_2$. Let $J(t) = 1$ if $\{\Sigma_{k_1=1}^{n_1} I(X_{1k_1} \geq t)\}\{\Sigma_{k_2=1}^{n_2} I(X_{2k_2} \geq t)\} > 0$ and $J(t) = 0$ otherwise. Assume a predictable weighting process, $\hat{w}(t)$, such that

$$\sup_{u \in [0,t)} |\hat{w}(u) - w(u)|$$

approaches zero in probability for a nonstochastic function $w(u)$ and vanishes for values of t where $J(t) = 0$. Now define the paired Pepe–Fleming family of tests as

$$\mathcal{T} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{1/2} \int_0^\infty \hat{w}(u) \left\{\hat{S}_1(u) - \hat{S}_2(u)\right\} du.$$

Further notation is required to describe the variance of $\mathcal{T}$. Hence, let $A_i(t) = \int_t^\infty w(u) S_i(u) du$ with estimate $\hat{A}_i(t) = \int_t^\infty \hat{w}(u) \hat{S}_i(u) du$. Define π_i as the probability of belonging to treatment group i with $\hat{\pi}_i = n_i/(n_1 + n_2)$ and define θ as the proportion of dependent observations in the two groups, with $\hat{\theta} = 2n/(n_1 + n_2)$. Finally, define $G_{12}(t_1, t_2) = P(X_{1k} \geq t_1, X_{2k} \geq t_2)\{P(X_{1k} \geq t_1) P(X_{2k} \geq t_2)\}^{-1}\{\lambda_{12}(t_1, t_2) - \lambda_{1|2}(t_1 \mid t_2)\, \lambda_2(t_2) - \lambda_{2|1}(t_2 \mid t_1)\lambda_1(t_1) + \lambda_1(t_1)\lambda_2(t_2)\}$. As n approaches infinity, $\mathcal{T}$ is asymptotically normal with variance

$$\sigma^2 = \sum_{i=1}^{2} \frac{\pi_1 \pi_2}{\pi_i} \left[\int_0^\infty \frac{\{A_i(u)\}^2 \lambda_i(u)}{P(X_{ik_i} \geq u)} du\right] - \theta \int_0^\infty \int_0^\infty A_1(u) A_2(v) G_{12}(u, v) dv du,$$

as shown in the Appendix. The first term in this expression corresponds to the variance of the original Pepe and Fleming (1989) statistic with independent treatment groups and the second term corrects for the dependence in the Kaplan–Meier curves. If the Kaplan–Meier curves are independent, the trailing term vanishes in this expression. If the Kaplan–Meier curves tend to vary in tandem due to an underlying positive correlation in the failure time random variables, then this trailing term causes the variance of the test statistic to shrink. Possible weighting choices include $\hat{w}(t) = J(t)\hat{P}(U_{1k_1} \geq t) \times \hat{P}(U_{2k_2} \geq t)/\{\hat{\pi}_1 \hat{P}(U_{1k_1} \geq t) + \hat{\pi}_2 \hat{P}(U_{2k_2} \geq t)\}$, which is similar to the weighting recommended by Pepe and Fleming, or alternatively $\hat{w}(t) = J(t)$, which reflects an interpretation according to years of life saved (YLS) on study.

Terms in this asymptotic closed-form variance are easily estimated using either pooled estimates under the null hypothesis or unpooled estimates under the alternative hypothesis. To present these variance estimates, more notation is

required. Define $n^* = (n_1 n_2)/n$. Let $N_i(t) = \Sigma_{k_i=1}^{n_i} I(X_{ik_i} \leq t, \Delta_{ik_i} = 1)$ count the number of individuals from group i who fail at time t and let $Y_i(t) = \Sigma_{k_i=1}^{n_i} I(X_{ik_i} \geq t)$ count the number of individuals from group i who are still at risk for failure at time t, $i = 1, 2$. An unpooled estimate for $\lambda_i(t)dt$ is $\{Y_i(t)\}^{-1} dN_i(t)$. Let $Y_{12}(t_1, t_2) = \Sigma_{k=1}^{n} I(X_{1k} \geq t_1, X_{2k} \geq t_2)$ count the number of complete correlated pairs still at risk for failure at times t_1 and t_2 in treatment groups 1 and 2, respectively. Also, let $dN_{12}(t_1, t_2) = \lim_{\Delta t_1, \Delta t_2 \to 0} \Sigma_{k=1}^{n} I(t_1 \leq X_{1k} < t_1 + \Delta t_1, t_2 \leq X_{2k} < t_2 + \Delta t_2, \Delta_{1k} = 1, \Delta_{2k} = 1)$ count the number of individuals from complete pairs who failed at time t_1 for treatment 1 and failed at time t_2 for treatment 2. An estimate for $\lambda_{12}(t_1, t_2)dt_1 dt_2$ is $\{Y_{12}(t_1, t_2)\}^{-1} dN_{12}(t_1, t_2)$. Let $dN_{1|2}(t_1 \mid t_2) = \lim_{\Delta t_1 \to 0} \Sigma_{k=1}^{n} I(t_1 \leq X_{1k} < t_1 + \Delta t_1, X_{2k} \geq t_2, \Delta_{1k} = 1)$ count the number of complete correlated pairs who failed at time t_1 for treatment 1 and who are still at risk for failure at time t_2 for treatment 2. And let $dN_{2|1}(t_2 \mid t_1) = \lim_{\Delta t_2 \to 0} \Sigma_{k=1}^{n} I(t_2 \leq X_{2k} < t_2 + \Delta t_2, X_{1k} \geq t_1, \Delta_{2k} = 1)$ count the number of complete correlated pairs who failed at time t_2 for treatment 2 and who are still at risk for failure at time t_1 for treatment 1. Hence, estimates for $\lambda_{1|2}(t_1 \mid t_2)dt_1$ and $\lambda_{2|1}(t_2 \mid t_1)dt_2$ are $\{Y_{12}(t_1, t_2)\}^{-1} dN_{1|2}(t_1 \mid t_2)$ and $\{Y_{12}(t_1, t_2)\}^{-1} dN_{2|1}(t_2 \mid t_1)$, respectively. An unpooled estimate for $P(X_{ik} \geq t)$ is $Y_i(t)/n_i$. Also $P(X_{1k} \geq t_1, X_{2k} \geq t_2)$ is estimated with $Y_{12}(t_1, t_2)/n$. Incorporating the above estimates, an unpooled estimate for $G_{12}(t_1, t_2)dt_1 dt_2$ is $\hat{G}_{12}(t_1, t_2)dt_1 dt_2 = n^* \{Y_1(t_1) Y_2(t_2)\}^{-1} Y_{12}(t_1, t_2)[\{Y_{12}(t_1, t_2)\}^{-1} dN_{12}(t_1, t_2) - \{Y_{12}(t_1, t_2) Y_2(t_2)\}^{-1} dN_{1|2}(t_1 \mid t_2) dN_2(t_2) - \{Y_{12}(t_1, t_2) Y_1(t_1)\}^{-1} \times dN_{2|1}(t_2 \mid t_1) dN_1(t_1) + \{Y_1(t_1) Y_2(t_2)\}^{-1} dN_1(t_1) dN_2(t_2)]$. Notice that $\hat{G}_{12}(t_1, t_2)dt_1 dt_2$ uses all available data to estimate marginal hazards and probabilities, while joint and conditional quantities are estimated with complete pairs only.

Hence, an unpooled variance estimate for σ^2 that could be used in either hypothesis testing or confidence interval construction is $\hat{\sigma}^2 = \Sigma_{i=1}^{2} \hat{\pi}_{3-i} [\int_0^\infty n_i \{Y_i(u)\}^{-2} \{\hat{A}_i(u)\}^2 dN_i(u)] - \hat{\theta} \int_0^\infty \int_0^\infty \hat{A}_1(u) \hat{A}_2(v) \hat{G}_{12}(u, v) du dv$. In relation to defining a pooled variance estimate, let

$$\tilde{A}(t) = \int_t^\infty \hat{w}(u) \tilde{S}(u) du$$

use the pooled Kaplan–Meier survival estimate in its integrand. Let $\bar{Y}(t) = Y_1(t) + Y_2(t)$ and $\bar{N}(t) = N_1(t) + N_2(t)$ so that a pooled estimate of $\lambda_i(t)dt$ is $\{\bar{Y}(t)\}^{-1} d\bar{N}(t)$. Let $\hat{H}_i(t)$ be the Kaplan–Meier estimate of the censoring survival function for group i. A pooled estimate for $P(X_{ik} \geq t)$ is $\hat{H}_i(t^-)\hat{S}(t^-)$. In a hypothesis testing framework, one may estimate σ^2 under the null hypothesis using the pooled estimate $\tilde{\sigma}^2 = \Sigma_{i=1}^{2} \hat{\pi}_{3-i} [\int_0^\infty \{\hat{H}_i(u^-) \hat{S}(u^-) \bar{Y}(u)\}^{-1} \{\tilde{A}(u)\}^2 \, d\bar{N}(u)] - \hat{\theta} \int_0^\infty \int_0^\infty \tilde{A}(u) \tilde{A}(v) \tilde{G}_{12}(u, v) dv du$, where $\tilde{G}_{12}(t_1, t_2) \, dt_1 dt_2 = Y_{12}(t_1, t_2) \{n \hat{S}(t_1^-) \hat{S}(t_2^-) \hat{H}_1(t_1^-) \hat{H}_2(t_2^-)\}^{-1} [\{Y_{12}(t_1, t_2)\}^{-1} \times dN_{12}(t_1, t_2) - \{Y_{12}(t_1, t_2) \bar{Y}(t_2)\}^{-1} dN_{1|2}(t_1 \mid t_2) \, d\bar{N}(t_2) - \{Y_{12}(t_1, t_2) \bar{Y}(t_1)\}^{-1} dN_{2|1}(t_2 \mid t_1) d\bar{N}(t_1) + \{\bar{Y}(t_1) \bar{Y}(t_2)\}^{-1} \times d\bar{N}(t_1) d\bar{N}(t_2)]$.

3. Simulations

In order to study finite sample properties of the test statistics, simulations were conducted for a variety of underlying correlation structures under null or alternative hypotheses using either $n = 50$ or $n = 100$ complete pairs of censored survival outcomes. In each simulation, bivariate log-normal failure distributions with correlation on the log scale, ρ_T, and bivariate log-normal censoring distributions with correlation on the log scale, ρ_U, were used independently in generating the observed paired data. Increasing levels of $\rho_T = \rho_U = (0.0, 0.3, 0.6, 0.9)$ were studied. In addition, finite sample properties for increasing ρ_T under common censoring times were studied by selecting $\rho_U = 1$. Log-scale means and variances for each of the two treatment group censoring times were 1.1 and 0.8, respectively. Under the hypothesis of no treatment difference, the paired failure times were generated with log-scale means and variances of 0.3 and 1.0, respectively. Under the alternative hypothesis, the paired log-scale means were taken to be 0.3 and 0.6, with remaining bivariate log-normal parameters unchanged. Along with the complete-case analysis, simulations were run to study the behavior of the test statistics when 25 additional singleton pair members per treatment group were available. The singletons were simulated using bivariate log-normal distributions with similar marginal distributions for failure and censoring times as described above but with zero correlation. Roughly one third of all failure times were censored. A type I error of 0.05 was employed in all simulations.

Size and power results are displayed in Table 1 for unpaired and paired Pepe–Fleming tests using $\hat{w}(t) = J(t)\hat{P}(U_{1k_1} \geq t)\hat{P}(U_{2k_2} \geq t)/\{\hat{\pi}_1 \hat{P}(U_{1k_1} \geq t) + \hat{\pi}_2 \hat{P}(U_{2k_2} \geq t)\}$ and pooled variance estimates. Observed type I errors for the paired tests using $n = 100$ are at desirable levels. Using $n = 50$, type I errors also look attractive except for the complete-case analysis at the highest level of correlation studied, $\rho_T = 0.9$, where the type I error seems to be slightly underestimated. Type I errors for $n = 50$ and $\rho_T = 0.9$ improved in the analysis that included 25 additional singletons. Results for type I error were similar using YLS weighting. Both pooled and unpooled variance estimates performed well in simulation. However, pooled variance estimates tended to slightly outperform unpooled variance estimates in terms of maintaining type I error, especially in the smaller sample sizes studied. Hence, pooled variance estimates will be recommended for hypothesis testing scenarios for purposes of constructing p-values and unpooled variance estimates will be recommended for constructing confidence intervals under the alternative hypothesis.

Regardless of whether pooled or unpooled estimates are employed and regardless of the weighting strategy used, the test sizes of the unadjusted original tests become increasingly conservative as the correlation in the censored survival endpoints grows and the dependence is not accounted for in the analysis. This is likely caused by the increased tendency of the Kaplan–Meier curves to vary in tandem for higher values of positive correlation. Without taking this into account, the test statistic has a very difficult time rejecting the null hypothesis. Figure 1 displays representative correlated Kaplan–Meier survival estimates under study for $n = 100$, where, in each panel of the figure, underlying marginal distributions of the curves under study are identical.

This lessened ability to reject the null hypothesis is also featured in power results under the alternative hypothesis in Table 1, where, for instance, power has a tendency to decrease with rising levels of correlation in the complete-case analyses.

Table 1
Size and power results across different underlying correlation structures[a]

				Size		Power	
n	ρ_T	ρ_U	Additional singletons	PF	Paired PF	PF	Paired PF
100	0.0	0.0	0	0.052	0.048	0.4418	0.4370
		0.0	25	0.041	0.044	0.5230	0.5150
		1.0	0	0.043	0.045	0.4416	0.4352
		1.0	25	0.047	0.047	0.5236	0.5214
	0.3	0.3	0	0.027	0.044	0.4384	0.5396
		0.3	25	0.021	0.041	0.5420	0.6164
		1.0	0	0.024	0.049	0.4358	0.5484
		1.0	25	0.023	0.042	0.5356	0.6810
	0.6	0.6	0	0.011	0.050	0.4266	0.7080
		0.6	25	0.018	0.053	0.5384	0.7282
		1.0	0	0.006	0.041	0.4204	0.7308
		1.0	25	0.014	0.042	0.5372	0.7398
	0.9	0.9	0	0.000	0.048	0.3800	0.9734
		0.9	25	0.000	0.046	0.5492	0.9102
		1.0	0	0.000	0.048	0.3638	0.9780
		1.0	25	0.000	0.039	0.5418	0.9190
50	0.0	0.0	0	0.037	0.041	0.2510	0.2414
		0.0	25	0.038	0.039	0.3546	0.3482
		1.0	0	0.043	0.047	0.2528	0.2448
		1.0	25	0.038	0.041	0.3476	0.3448
	0.3	0.3	0	0.025	0.048	0.2242	0.2998
		0.3	25	0.034	0.047	0.3408	0.3968
		1.0	0	0.021	0.049	0.2148	0.2998
		1.0	25	0.036	0.051	0.3358	0.4012
	0.6	0.6	0	0.006	0.057	0.1672	0.4142
		0.6	25	0.009	0.051	0.3244	0.4730
		1.0	0	0.009	0.058	0.1688	0.4282
		1.0	25	0.015	0.047	0.3174	0.4826
	0.9	0.9	0	0.000	0.033	0.0832	0.7866
		0.9	25	0.000	0.051	0.2940	0.6344
		1.0	0	0.000	0.034	0.0738	0.8132
		1.0	25	0.006	0.045	0.2860	0.6388

[a] n, number of complete failure-time pairs generated from the bivariate log-normal distribution; ρ_T, correlation between failure times on the log scale; and ρ_U, correlation between bivariate log-normal censoring times on the log scale so that $\rho_U = 1$ gives common censoring times. In the additional singletons column, the complete pair analysis is denoted by 0; otherwise, the results are for the analysis with 25 singletons added per treatment group. The Pepe–Fleming test for independent treatment groups rejection rate is listed under PF, and the proposed test, which adjusts for the paired structure, is denoted as paired PF. One thousand and 5000 Monte Carlo simulations were used for size and power, respectively.

This phenomenon was repeated regardless of the weighting method chosen or the variance estimation method employed. However, using the proposed paired tests, which take into account the overall dependence structure in the data, significant increases in power were observed for increasing values of correlation in the failure times. Similar increases in simulated power were observed for each combination of weighting strategy and variance estimation procedure for the proposed paired test statistics.

With regard to the inclusion of singletons in an appropriate paired analysis, power was found to increase over the complete-case analysis for $\rho_T = 0.0, 0.3$, and 0.6, with gains tapering off as the level of correlation increased. For the highest level of correlation explored, $\rho_T = 0.9$, power did not improve but instead declined with the inclusion of the additional singletons. When deciding to use singletons in an analysis, there appears to be a trade-off between gaining additional sample size and gaining additional variability in the comparison of marginal Kaplan–Meier estimates. For extremely high levels of correlation, the use of additional singletons does not seem warranted with this statistic.

4. Application to Early Treatment Diabetic Retinopathy Study

As an illustrative example, the ETDRS comparison of time to severe vision loss, described in the Introduction, is again considered. Approximately 12 million Americans are affected by diabetes, the leading cause of blindness in working-age Americans, accounting for 12% of the new cases of blindness each year (Patz and Smith, 1991). In order to evaluate the best timing for laser photocoagulation surgery, 3711 patients had one eye randomized to receive early photocoagulation and

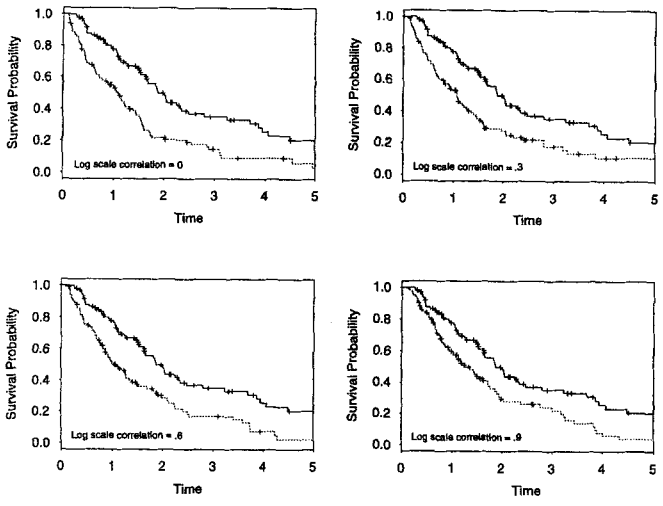

Figure 1. Survival curves constructed under the alternative hypothesis for increasing values of correlation ($n = 100$). The same random seed was used in generating each of the curves. The true underlying marginal distributions of the curves are identical in each panel, yet the more correlated curves appear closer together.

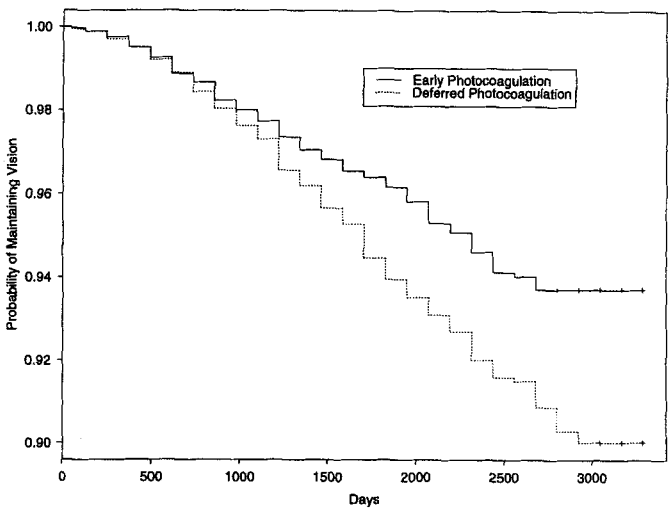

Figure 2. Probability of maintaining vision according to treatment. Correlation between the two Kaplan–Meier curves causes them to vary in tandem and appear more similar than they might if they were constructed from independent data sets.

the other eye received the photocoagulation surgery upon detection of high-risk retinopathy. Time to severe vision loss is positively correlated within an individual. Estimated probabilities of retaining acceptable vision by treatment are displayed in Figure 2. In addition to the paired censored data structure, the survival curves reveal an interval censored aspect to this data, a feature we will ignore for the purposes of this example.

At the time this study was originally published, methods for correctly handling paired censored times to severe vision loss were not available to the investigators; hence, they used standard methods for independent samples in designing and analyzing the ETDRS results while indicating that their analyses were likely to be conservative due to the underlying paired data structure.

In the following, 95% confidence intervals for the integrated weighted survival difference, $\Delta_S = \int_0^\infty w(u)\{S_1(u) - S_2(u)\}du = \{(n_1+n_2)/n_1n_2\}^{1/2}\mathcal{T}$, are based on unpooled variance estimates, and standardized test statistics, $\Delta_S \div[\{(n_1+n_2)n_1n_2\}^{1/2}\tilde{\sigma}]$, use pooled variance estimates. Treatment 1 refers to eyes randomized to immediate photocoagulation, with $n = n_1 = n_2 = 3711$. Using the Pepe–Fleming recommended weight as described in Section 2, the standardized test statistic comparing early to delayed photocoagulation was observed to be 3.75 ($p = 1.77 \times 10^{-4}$; $\hat{\Delta}_S = 18.40$; 95% CI for $\Delta_S = 8.81, 27.98$), where the study period spanned 9 years. Using the YLS weighting, the observed standardized test statistic was 4.64 ($p = 3.48 \times 10^{-6}$; $\hat{\Delta}_S = 50.44$; 95% CI for $\Delta_S = 29.22, 71.66$). The paired log-rank test, using pooled variance formulas as in Murray (2000), also detects significant differences in cumulative hazard functions favoring the early photocoagulation group ($p = 1.07 \times 10^{-6}$). Note that, using the YLS weights, $\hat{\Delta}_S$ may be interpreted as the estimated average difference in days of extended vision on the immediate photocoagulation therapy during the first 9 years on study, after which time both weighting functions became zero. For comparison, an analysis treating the eyes as independent results in larger p-values and wider confidence intervals. Results for the unpaired test using Pepe–Fleming-styled weights were 2.99 ($p = 1.39 \times 10^{-3}$; $\hat{\Delta}_S = 18.40$; 95% CI for $\Delta_S = 6.34, 30.45$) and using YLS weights were 3.79 ($p = 1.51 \times 10^{-4}$; $\hat{\Delta}_S = 50.44$; 95% CI for $\Delta_S = 24.38, 76.51$). At the time the ETDRS study results were published, emphasis was placed on improving early detection of diabetic retinopathy so that photocoagulation therapy could be initiated right away. Using new paired censored survival analysis tools, future studies designed similarly to the ETDRS can use fewer patient resources in achieving a particular desired power.

5. Discussion

This research extends the methods of Pepe and Fleming (1989) to the paired censored survival setting. Closed-form asymptotic variances of the adjusted test statistics are presented along with pooled and unpooled variance estimates. The methods presented are able to accommodate the gamut of uninformative paired correlated censoring structures ranging from common censoring times to independent censoring mechanisms. Also, single unpaired individuals may contribute to the marginal estimation of survival curves within the test statistic. Simulations in this research show that one may take advantage of the paired structure of censored survival data and often benefit from additional singleton structured data in an analysis, especially if the correlation in underlying survival times is mild to moderate. In cases where correlation in underlying survival times is extremely high, simulations indicate that a complete-case analysis using the proposed methods is preferable to incorporating singletons in the paired analysis. In addition, simulation results in Section 3 underscore the

disadvantages of ignoring pair matching in an analysis. Not only is type I error adversely conservative, but power in the presence of positively correlated paired data results in less power using traditional methods for independent samples than would be expected if the samples were truly independent.

Within each treatment group, individual independent and identically distributed observations contribute equally to the estimation of Kaplan–Meier survival curves whether these observations have a corresponding pair member or not. While gains in efficiency are afforded by accounting for correlation in the estimated curves through the variance of the statistic in this marginal analysis, further gains in efficiency may be obtained by taking additional advantage of the information in the paired event times in estimating the survival curves used in the paired Pepe–Fleming statistics. Results from Manatunga and Oakes (1999), in a slightly different context of bivariate observations where treatment does not necessarily differ within pairs, lend support to estimation and testing approaches that place higher value on more informative complete pairs, especially in the presence of high correlation. This intuition is further bolstered by simulation results in Section 3, which indicate that, for extremely high values of underlying correlation, the power of the complete pair analysis surpasses the power of the analysis that incorporates additional singleton values. Additional work in this area is needed in order to fully tap the statistical information in this data structure.

The weight recommended by Pepe and Fleming (1989) deemphasizes areas under the survival curve where censoring is heavy and hence can be particularly effective in detecting early treatment differences when censoring is heavy in the tails. Use of the YLS weight in the test statistic is often an attractive choice for detecting differences later in the study period and gives an attractive interpretation that should appeal to nonstatistically minded collaborators. Because the proposed methods adapt an already popular family of test statistics, the process of transition to these more efficient tests in the paired censored survival setting should be straightforward once software is available. In addition, results from this research apply to the quality-adjusted survival setting, as discussed by Glasziou et al. (1998), when nonstochastic weights are chosen to reflect quality of life while on treatment. Hence. these methods also extend quality-adjusted survival analysis to the paired censored survival setting.

Acknowledgements

The author 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.

Résumé

Cette recherche introduit des méthodes non paramétriques pour tester la différence entre des courbes de survie pondérées dans le cas de planification concernant l'étude de survie appariées avec censures. Ce travail est une extension de celui qui a été réalisé par Pepe et Fleming (1989) qui ont utilisé des statistiques de test analogues dans le but de comparer des groupes de traitements indépendants. Une distribution de forme asymptotiquement fermée est présentée pour la famille des tests proposés. On donne également des estimateurs de la variance, construits sous les hypothèses nulle et alternative, à partir des estimateurs non paramétriques du maximum de vraisemblance des quantités de forme fermée. La méthode décrite permet de prendre en compte, dans la statistique de test, l'information additionnelle relative à des sujets n'ayant pas de membre associé à leur paire dans l'appariement. Ceci se fait, grâce à des scénarios d'échantillonnage où les singletons ne sont pas particulièrement soumis à des biais de sélection. Des simulations effectuées pour divers types de dépendance potentielle entre les données de survie censurées et appariées, montrent l'existence de gain substantiel, quand on prend en compte dans l'analyse la structure de dépendance sur les observations. Ignorer la nature appariée des données conduit à des tests par trop conservateurs en terme de puissance et de taille. En effet, les résultats des simulations mettant en œuvre des tests considérant les échantillons comme indépendants, alors qu'il existe une corrélation positive, atteignent des objectifs de taille et de puissance notablement en dessous de ceux que l'on aurait pu obtenir en l'absence de corrélation. Ces effets supplémentaires mettent bien en lumière la nécessité de prendre en compte la dépendance des observations dans cette famille de tests si populaire.

References

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2,** 437–453.

Dabrowska, D. (1989). Rank tests for matched pair experiments with censored data. *Journal of Multivariate Analysis* **28,** 88–114.

Dabrowska, D. (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 7. *Ophthalmology* **98,** 741–756.

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

Feigal, R. J., Musherure, P., Gillespie, B., Levy-Polack, M., Quelhas, I., and Hebling, J. (2000). Improved sealant retention using bonding agents: A clinical study of two-bottle and single-bottle systems. *Journal of Dental Research* **79,** 1850–1856.

Glasziou, P., Cole, B., Gelber, R., Hilden, J., and Simes, R. J. (1998). Quality-adjusted survival analysis with repeated quality-of-life measures. *Statistics in Medicine* **17,** 1215–1229.

Holt, J. and Prentice, R. (1974). Survival analysis in twin studies and matched pair experiments. *Biometrika* **61,** 17–30.

Huang, Y. (1999). The two-sample problem with induced dependent censorship. *Biometrics* **55,** 1108–1113.

Jung, S. (1999). Rank tests for matched survival data. *Lifetime Data Analysis* **5,** 67–79.

Lee, E., Wei, L. J., and Amato, D. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art,* J. P. Klein and P. K. Goel (eds.), 237–247. Dordrecht: Kluwer Academic.

Manatunga, A. and Oakes, D. (1999). Parametric analysis for

matched pair survival data. *Lifetime Data Analysis* **5,** 371–387.

Murray, S. (2000). Nonparametric rank-based methods for group sequential monitoring of paired censored survival data. *Biometrics* **56,** 984–991.

Murray, S. and Cole, B. (2000). Variance and sample size calculations in quality of life adjusted survival analysis (Q-TWiST). *Biometrics* **56,** 266–275.

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

Patz, A. and Smith, R. (1991). The ETDRS and Diabetes 2000. *Ophthalmology* **98,** 739–740.

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

Wei, L. J. (1980). A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *Journal of the American Statistical Association* **75,** 634–637.

Woolson, R. and Lachenbruch, P. (1980). Rank tests for censored matched pairs. *Biometrika* **67,** 597–606.

Woolson, R. and O'Gorman, T. (1992). A comparison of several tests for censored paired data. *Statistics in Medicine* **11,** 193–208.

Received October 1999. *Revised July* 2000.
Accepted September 2000.

Appendix

Define the martingale $M_i(t) = N_i(t) - \int_0^t \lambda_i(u)Y_i(u)du$ with respect to the filtration containing all available censoring and survival data for the endpoint corresponding to group i prior to time t, $i = 1, 2$. Because the filtrations for $M_1(t)$ and $M_2(t)$ are dependent and nonnested, covariances relating to these martingales will be directly derived rather than conditioning first on a common filtration as is done in standard martingale theory. Toward this end, it is useful to note that $M_i(t) = N_i(t) - \int_0^t \lambda_i(u)Y_i(u)du$ can be rewritten as $\Sigma_{k_i=1}^{n_i}\{I(X_{ik_i} \le t, \Delta_{ik_i} = 1) - \int_0^t \lambda_i(u)I(X_{ik_i} \ge u)du\} = \Sigma_{k_i=1}^{n_i} M_{ik_i}(t)$, which is a sum of independent and identically distributed quantities. Let $\hat{\Lambda}_i(t)$ be the Nelson–Aalen hazard estimate at time t and let $\Lambda_i(t)$ be the cumulative hazard for failure by time t. For the moment, focus on the term $(n^*)^{1/2}\int_0^t\{Y_i(u)\}^{-1}dM_i(u)$, which is equal to $(n^*)^{1/2}\{\hat{\Lambda}_i(t) - \Lambda_i(t)\}$ for values of t where $J(t) > 0$. After an application of the martingale central limit theorem (or Lenglart's inequality) similar to that used in the appendix of Murray and Cole (2000), this term has the same limiting distribution as

$$Z_i(t) = \sqrt{n^*}\left[n_i^{-1}\sum_{k_i=1}^{n_i}\int_0^t \{P(X_i \ge u)\}^{-1}dM_{ik_i}(u)\right].$$

Since dependent terms between $Z_1(t_1)$ and $Z_2(t_2)$ involve only terms with $k_1 = k_2 = k \le n$, the multivariate central limit theorem identifies the covariance of $Z_1(t_1)$ and $Z_2(t_2)$ as

$$\begin{aligned}
&\operatorname{cov}\left[\int_0^{t_1}\{P(X_1 \ge u)\}^{-1}dM_{1k}(u),\right.\\
&\qquad\left.\int_0^{t_2}\{P(X_2 \ge v)\}^{-1}dM_{2k}(v)\right]\\
&= \mathrm{E}\left(\left[\int_0^{t_1}\{P(X_1 \ge u)\}^{-1}dM_{1k}(u)\right]\right.\\
&\qquad\left.\times\left[\int_0^{t_2}\{P(X_2 \ge v)\}^{-1}dM_{2k}(v)\right]\right),
\end{aligned}$$

which, after some calculation, becomes

$$\int_0^{t_1}\int_0^{t_2} G_{12}(u,v)dvdu. \tag{1}$$

The above result, which pertains to covariances between dependent Nelson–Aalen hazard estimates in this general setting, may be further utilized in understanding the asymptotic behavior for dependent Kaplan–Meier estimates. A result derived by Breslow and Crowley (1974) shows that

$$\sqrt{n^*}\left[\hat{S}_i(t) - \exp\{-\hat{\Lambda}_i(t)\}\right] \xrightarrow{\mathcal{P}} 0.$$

Hence, for values t_1 and t_2, where $\min\{Y_1(t_1), Y_2(t_2)\} > 0$, the delta method in combination with (1) gives

$$\begin{aligned}
&\operatorname{cov}\left[\sqrt{n^*}\left\{\hat{S}_1(t_1) - S_1(t_1)\right\}, \sqrt{n^*}\left\{\hat{S}_2(t_2) - S_2(t_2)\right\}\right]\\
&\quad = S_1(t_1)S_2(t_2)\int_0^{t_1}\int_0^{t_2} G_{12}(u,v)dvdu.
\end{aligned}$$

Finally,

$$\begin{aligned}
\operatorname{var}(T) &= \operatorname{var}\left[\{n_1n_2/(n_1+n_2)\}^{1/2}\right.\\
&\qquad\left.\times\int_0^{\infty}\hat{w}(u)\left\{\hat{S}_1(u) - \hat{S}_2(u)\right\}du\right]\\
&\approx \sum_{i=1}^{2}\pi_{3-i}\int_0^{\infty}\{A_i(u)\}^2\lambda_i(u)\{P(X_{ik_i} \ge u)\}^{-1}du\\
&\quad - \theta\operatorname{cov}\left\{\sqrt{n^*}\int_0^{\infty} w(t_1)\hat{S}_1(t_1)dt_1,\right.\\
&\qquad\left.\sqrt{n^*}\int_0^{\infty} w(t_2)\hat{S}_2(t_2)dt_2\right\}\\
&= \sum_{i=1}^{2}\pi_{3-i}\int_0^{\infty}\{A_i(u)\}^2\lambda_i(u)\{P(X_{ik_i} \ge u)\}^{-1}du\\
&\quad - \theta\int_0^{\infty}\int_0^{\infty} w(t_1)w(t_2)\\
&\qquad\times\operatorname{cov}\left\{\sqrt{n^*}\hat{S}_1(t_1), \sqrt{n^*}\hat{S}_2(t_2)\right\}dt_1dt_2\\
&= \sum_{i=1}^{2}\pi_{3-i}\int_0^{\infty}\{A_i(u)\}^2\lambda_i(u)\{P(X_{ik_i} \ge u)\}^{-1}du\\
&\quad - \theta\int_0^{\infty}\int_0^{\infty} w(t_1)w(t_2)S_1(t_1)S_2(t_2)
\end{aligned}$$

$$\times \int_0^{t_1} \int_0^{t_2} G_{12}(u,v)dvdudt_1dt_2$$

$$= \sum_{i=1}^{2} \pi_{3-i} \int_0^{\infty} \{A_i(u)\}^2 \lambda_i(u) \{P(X_{ik_i} \geq u)\}^{-1} du$$

$$- \theta \int_0^{\infty} \int_0^{\infty} A_1(u)A_2(v)G_{12}(u,v)dvdu$$

$$= \sigma^2,$$

as presented in Section 2.