Biometrika (2004), **91,** 2, *pp.* 263–275

Printed in Great Britain

Sample-size formula for clustered survival data using weighted log-rank statistics

By RONALD E. GANGNON

Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 610 N. Walnut Street, Madison, Wisconsin 53726, U.S.A.

ronald@biostat.wisc.edu

and MICHAEL R. KOSOROK

Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, Wisconsin 53706, U.S.A.

kosorok@biostat.wisc.edu

Summary

We present a simple sample-size formula for weighted log-rank statistics applied to clustered survival data with variable cluster sizes and arbitrary treatment assignments within clusters. This formula is based on the asymptotic normality of weighted log-rank statistics under certain local alternatives in the clustered data context. We also provide consistent variance estimators. The derived sample-size formula reduces to Schoenfeld's (1983) formula for cases of no clustering or independence within clusters. Simulation results verify control of the Type I error and accuracy of the sample-size formula. Use of the sample-size formula in an event-driven clinical trial design is illustrated using data from the Early Treatment Diabetic Retinopathy Study.

Some key words: Clustered data; Local alternative; Log-rank statistic; Martingale residuals; Paired data; Proportional hazards; Sample size.

1. Introduction

Clustered survival data occur when a single type of event is assessed on two or more distinct, similar units within a common subject. For example, in ophthalmology, time to moderate vision loss could be assessed separately on both eyes of a person. In constrast, multivariate survival data occur when distinct events or repeated events occur to the same subject. For example, in cardiology, time to myocardial infarction and time to stroke could both be assessed on the same person. Clustered survival data may be distinguished from multivariate survival data by observing that a common marginal hazards model is likely to be appropriate for clustered survival data and inappropriate for multivariate survival data. The focus of this paper is on clustered survival data, although much of the methodology could be extended to multivariate survival data as well.

Several authors, including Mantel & Ciminera (1979), Woolson & Lachenbruch (1980), Wei (1980), O'Brien & Fleming (1987), Albers (1988) and Dabrowska (1990), have proposed nonparametric rank-based tests for survival differences in paired survival data. Dabrowska (1989), Huang (1999), Jung (1999) and Murray (2000) present adjusted variance estimators

for weighted log-rank statistics with paired survival data. Murray (2000) discusses this family of statistics in the context of group sequential monitoring of clinical trials. Murray (2001) considered nonparametric testing of weighted integrated survival differences in paired survival data. For non-paired clustered survival data, attention has principally been focused on parametric and semiparametric models. The marginal proportional hazards model of Wei et al. (1989) is especially attractive because the dependence structure within clusters need not be specified. Rosner & Glynn (1997) describe a model for analysing clustered ordinal data, which can be applied to clustered survival data. The dependence within clusters can also be modelled explicitly using parametric or semiparametric frailty models (Aalen, 1987).

In this paper, we propose a nonparametric method for detecting survival differences in clustered survival data based on weighted log-rank statistics. Simple sample-size formulae, analogous to Schoenfeld's (1983) formula, are derived for these statistics. These methods allow for fairly arbitrary weights, variable cluster sizes and arbitrary treatment assignments within clusters. In the case of a marginal proportional hazards model, these sample-size formulae are also applicable to tests and confidence intervals for the regression coefficient described in Wei et al. (1989). A key step in deriving these results is establishing asymptotic normality of the test statistic under local alternative hypotheses. We also resolve the tail instability problem for dependent failure times (Bilias et al., 1987) with minimal assumptions on the form of their joint distribution. A consistent variance estimator for the proposed test statistic is also provided. The sample-size formula as well as both the log-rank statistic and its estimated variance can be easily calculated using existing S-Plus and SAS software.

In § 2, the data structure and model assumptions are presented. The weighted log-rank statistic for clusters is described in § 3, and its asymptotic distribution along with a consistent variance estimator is given. Sample-size formulae are provided in § 4. Simulation results in § 5 verify the properties of the cluster log-rank test and sample-size formula under both null and alternative hypotheses and demonstrate the superiority of this approach to common alternative analyses such as time-to-first-event when treatments are assigned to clusters and ignoring clustering with paired survival data. In § 6, we present an example using the data from the Early Treatment Diabetic Retinopathy Study to demonstrate the implementation of an event-driven design with paired survival data. A discussion follows in § 7.

2. THE DATA STRUCTURE AND THE MODEL

The observed data $\{(X_{ijk}, \delta_{ijk}), k = 1, \ldots, m_{ij}, j = 1, 2, i = 1, \ldots, n\}$ consist of n independent clusters, two treatments, and m_{ij} individuals within cluster i and treatment j; m_{ij} may be zero. We have $X_{ijk} = T_{ijk} \wedge C_{ijk}$ and $\delta_{ijk} = 1\{X_{ijk} = T_{ijk}\}$, where T_{ijk} is a time-to-event of interest, C_{ijk} is a right-censoring time, $x \wedge y$ denotes the minimum of x and y, and $1\{A\}$ denotes the indicator of A.

We assume that $\{T_{ijk}, k = 1, \ldots, m_{ij}\}$ and $\{C_{ijk}, k = 1, \ldots, m_{ij}\}$ are independent within each cluster and treatment combination, $j = 1, 2$ and $i = 1, \ldots, m$. Failure and censoring times may otherwise be dependent within a cluster. Although the distribution functions involved may depend on n, we will sometimes suppress this dependence for clarity. Let $\tilde{\pi}_j^n(t) \equiv n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_{ij}} E1\{C_{ijk} \geqslant t\}$ be the average number of individuals per cluster assigned to treatment j not yet censored at time $t-$, where we define any summation from 1 to m_{ij} to be zero if $m_{ij} = 0$. We also assume that $\tilde{\pi}_j^n$ converges uniformly to $\tilde{\pi}_j$ $(j = 1, 2)$ and

that cluster sizes are bounded, that is $0 \leqslant m_{ij} \leqslant m_0 < \infty$, $i = 1, \ldots, n$, $j = 1, 2$, and that $\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} m_{ij} \in (0, m_0)$. Let $\tilde{I} \equiv \{t \geqslant 0 : \tilde{\pi}_1(t) \wedge \tilde{\pi}_2(t) > 0\}$ denote the interval of observation permitted by censoring.

We assume a contiguous sequence of models for the failure times to enable development of workable sample-size formulae. Except for the presence of clusters, this approach is similar to that taken by Schoenfeld (1983). For each sample size $n \geqslant 1$, we assume that the marginal distributions of failure times are identical within treatment $j = 1, 2$, with integrated hazard Λ_j^n having the following properties: for $j = 1, 2$,

$$\sup_{t\in\tilde{I}} \left| \frac{d\Lambda_j^n(t)}{d\Lambda_0(t)} - 1 \right| \to 0, \quad \sup_{t\in\tilde{I}} \left| \sqrt{n} \left\{ \frac{d\Lambda_1^n(t)}{d\Lambda_2^n(t)} - 1 \right\} - \phi(t)\{1 + \eta(t)\} \right| \to 0,$$

as $n \to \infty$, for some cumulative hazard Λ_0 with corresponding survival function S_0, where ϕ is either cadlag, i.e. right-continuous with left-hand limits, or caglad, i.e. left-continuous with right-hand limits, with bounded total variation, and where η is bounded with nonzero values only at the jump points of S_0, where there may be ties in the failure times. Both proportional hazards and proportional odds local alternatives satisfy the above requirements. For example, the proportional hazards local alternative for Λ_j^n is defined via its survival function $S_j^n = 1 - F_j^n$, with

$$S_j^n(t) = \exp\left[- \int_{[0,t]} \exp\{(-1)^{j-1}\phi(u)/(2\sqrt{n})\} dA_0(u) \right] \quad (j = 1, 2),$$

where $A_0 = -\log S_0$. Choosing $\phi = 1$ yields the unweighted proportional hazards alternative. When there are ties, $\eta = \log\{e^{-\Delta\Lambda_{j0}}/(1 - \Delta\Lambda_{j0})\}$, and thus it is necessary to assume that $\sup_{t\in\tilde{I}} \Delta\Lambda_0(t) < 1$.

We will use counting process notation (Fleming & Harrington, 1991, pp. 15–49; Andersen et al., 1993, pp. 48–59) throughout the paper. Define the at-risk processes

$$Y_{ijk}(t) \equiv 1\{X_{ijk} \geqslant t\}, \quad \bar{Y}_j \equiv \sum_{i=1}^{n} \sum_{k=1}^{m_{ij}} Y_{ijk},$$

and let $\pi_j^n(t) = n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_{ij}} E Y_{ijk}(t)$. Under the above assumptions, standard probability arguments yield $\sup_{t\in[0,\infty]} |n^{-1}\bar{Y}_j(t) - \pi_j(t)| \to 0$ almost surely, as $n \to \infty$, where $\pi_j(t) \equiv \tilde{\pi}_j(t) S_0(t-)$ $(j = 1, 2)$. We write $I \equiv \{t \geqslant 0 : \pi_1(t) \wedge \pi_2(t) > 0\}$ and $\tau \equiv \sup I$, and also assume that $\sup_{t\in\{[0,\infty]-I\}} \pi_1^n(t) \wedge \pi_2^n(t) = 0$ for all n large enough.

3. Weighted log-rank tests for clustered data

The weighted log-rank test statistic for clustered data that we propose is

$$H_n \equiv n^{-\frac{1}{2}} \int_0^\infty \hat{U}_n(s) \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} \left\{ \frac{d\bar{N}_1(s)}{\bar{Y}_1(s)} - \frac{d\bar{N}_2(s)}{\bar{Y}_2(s)} \right\}, \tag{1}$$

where

$$N_{ijk}(t) \equiv 1\{X_{ijk} \leqslant t, \delta_{ijk} = 1\} \quad (k = 1, \ldots, m_{ij}, i = 1, \ldots, n), \quad \bar{N}_j \equiv \sum_{i=1}^{n} \sum_{k=1}^{m_{ij}} N_{ijk} \quad (j = 1, 2)$$

are the counting processes of observed events, and where $\hat{U}_n \geqslant 0$ is either cadlag or caglad with uniformly bounded total variation. We assume that

$$\sup_{t\in K} |\hat{U}_n(t) - U(t)| \to 0 \tag{2}$$

in probability, for some function U and every closed subinterval $K \subset I$. Typically, $\hat{U}_n$ would be nonnegative so that H_n would be sensitive to ordered hazards alternatives. Note that setting $U = \phi$ does not guarantee optimality as it does in the independent case, since the optimal weight depends both on the dependency structure within clusters and on the marginal model (Jung, 1999).

Most of the standard weight functions, such as the $G^{\rho,\gamma}$ family of Fleming & Harrington (1981) based on either pooled Kaplan–Meier or at-risk estimators, satisfy the above criteria. For example, let $\hat{S}_n$ be the Kaplan–Meier estimator based on the pooled data. The following lemma states that this weight function will work for clustered data. The proof, along with all other proofs, is provided in the Appendix.

LEMMA 1. *If* $\hat{U}_n = B_n^{\rho}(1 - B_n)^{\gamma}$, *where* B_n *is either the left- or right-continuous version of either* $\hat{S}_n$ *or* $n^{-1}(\bar{Y}_1 + \bar{Y}_2)$, *and* $\rho, \gamma \in [0, \infty]$, *then* (2) *is satisfied for all closed* $K \subset I$ *and* $\hat{U}_n$ *has uniformly bounded total variation.*

For H_n to have a limiting distribution, we need to ensure the existence of a limiting variance. Interestingly, this, in combination with certain moment conditions, is all that is needed beyond the assumptions we have already stated. This is because our primary tool for establishing weak convergence is the Lindeberg–Feller central limit theorem, which requires no more of a statistic than that it be asymptotically equivalent to a sum of independent terms with a limiting variance that is bounded. Of course, the difficult part is establishing this asymptotic equivalence. Let

$$M_{ijk}(t) \equiv N_{ijk}(t) - \int_0^t Y_{ijk}(s) d\Lambda_j^n(s) \ \ (k = 1, \ldots, m_{ij}), \quad \bar{M}_{ij.} \equiv \sum_{k=1}^{m_{ij}} M_{ijk} \ \ (j = 1, 2, i = 1, \ldots, n).$$

Define

$$\sigma_n^2 \equiv n^{-1} \sum_{i=1}^{n} E\left[\int_0^{\infty} U(s) \left\{ \frac{\pi_2^n(s)}{\pi_1^n(s) + \pi_2^n(s)} d\bar{M}_{i1.}(s) - \frac{\pi_1^n(s)}{\pi_1^n(s) + \pi_2^n(s)} d\bar{M}_{i2.}(s) \right\} \right]^2.$$

The following theorem establishes the limiting distribution of H_n.

THEOREM 1. *Under the stated model assumptions, and provided that* $\lim_{n \to \infty} \sigma_n^2 = \sigma^2 < \infty$, H_n *converges in distribution to a normal random variable with mean* μ *and variance* σ^2, *where*

$$\mu \equiv \int_0^{\infty} U(s) \phi(s) \{1 + \eta(s)\} \frac{\pi_1(s) \pi_2(s)}{\pi_1(s) + \pi_2(s)} d\Lambda_0(s).$$

The model assumptions in Theorem 1, provided in detail in § 2, are as follows: failure and censoring times are independent within clusters; individuals in the same treatment share the same marginal failure time distribution; and the marginal cumulative hazards satisfy certain conditions, which are met by proportional hazards and proportional odds local alternatives.

We now present a simple, consistent variance estimator for H_n. Let

$$\hat{M}_{ijk}(t) \equiv N_{ijk}(t) - \int_0^t Y_{ijk}(s) d\bar{N}_j(s) / \bar{Y}_j(s), \quad \check{M}_{ij.} \equiv \sum_{k=1}^{m_{ij}} \hat{M}_{ijk},$$

and define

$$\hat{\sigma}_n^2 \equiv n^{-1} \sum_{i=1}^{n} \left[\int_0^{\infty} \hat{U}_n(s) \left\{ \frac{\bar{Y}_2(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} d\check{M}_{i1.}(s) - \frac{\bar{Y}_1(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} d\check{M}_{i2.}(s) \right\} \right]^2.$$

THEOREM 2. *Under the conditions of Theorem* 1, $\hat{\sigma}_n^2 \to \sigma^2$ *in probability, as* $n \to \infty$.

4. Sample-size formulae

We will now derive a general sample-size formula under simplifying assumptions. Assume that the clusters are all of the same size m and that the marginal distributions for the censoring times are all identical. Assume that the baseline hazards are continuous and that the local alternative yields $\phi = \gamma U$, where $U : [0, \infty) \mapsto [0, \infty)$ is the limiting weight function for the chosen weighted log-rank statistic and $\gamma \in \mathbb{R}$. A marginal unweighted proportional hazards local alternative yields $U = 1$ for the unweighted log-rank statistic.

Assume that, within each cluster of size m, $m_1 = m(1+r)/2$ $(0 \leqslant r \leqslant 1)$ units are assigned to one treatment, and $m_2 = m(1-r)/2$ units are assigned to the other. If n_1 is the number of clusters in which treatment 1 is assigned to m_1 units, assume that $\lim_{n\to\infty} n_1/n = p_1 \in [0, 1]$, and define $p_2 = 1 - p_1$. Then $a_1 \equiv 1/2 + r(2p_1 - 1)/2$ and $a_2 \equiv 1/2 - r(2p_1 - 1)/2$ are the limiting proportions of units assigned to treatments 1 and 2, respectively. Let $M^U_{ijk}(t) = \int_0^t U(s) dM_{ijk}(s)$. Assume also that, for each $n \geqslant 1$, the correlation between $M^U_{ijk}(\infty)$ and $M^U_{ij'k'}(\infty)$ is ρ, for $j, j' = 1, 2$, and $k, k' = 1, \ldots, m_{ij}$, and provided that either $j \neq j'$ or $k \neq k'$. Let

$$D \equiv m^{-1} \int_0^\infty U^2(s) \{\pi_1(s) + \pi_2(s)\} d\Lambda_0(s),$$

and note that, for the log-rank test corresponding to $U = 1$, D is the marginal probability of observing an event.

The limiting distribution given in Theorem 1 can now be simplified to $\mu = \gamma m a_1 a_2 D$ and

$$\sigma^2 = m a_1 a_2 \left\{1 + \left(m \frac{p_1 p_2}{a_1 a_2} r^2 - 1\right)\rho\right\} D,$$

since $\lim_{n\to\infty} \text{var}\{M^U_{ijk}(\infty)\} = D$ for $j = 1, 2$. If we let Z_p be the pth quantile of a standard normal distribution, the number of clusters required to have a type II error rate of β for the alternative $\phi = \gamma U$ for a two-sided test of size α is

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{m D a_1 a_2 \gamma^2} \left\{1 + \left(m \frac{p_1 p_2}{a_1 a_2} r^2 - 1\right)\rho\right\}. \tag{3}$$

When $m = 1$ or $\rho = 0$, this formula simplifies to Schoenfeld's (1983) sample-size formula. If entire clusters are assigned to the same treatment, that is $r = 1$ and $a_1 = p_1$, the formula reduces to

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{m D a_1 a_2 \gamma^2} \{1 + (m-1)\rho\}.$$

If paired treatment assignments are made within clusters of size $m = 2$, the formula reduces to

$$n = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{D\gamma^2}(1 - \rho).$$

Alternatively, for the log-rank test corresponding to $U = 1$, these formulae may be rewritten in terms of the required number of events, $K = mnD$. For example, equation (3) is equivalent to

$$K = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{a_1 a_2 \gamma^2} \left\{1 + \left(m \frac{p_1 p_2}{a_1 a_2} r^2 - 1\right)\rho\right\}. \tag{4}$$

In most settings, it will be difficult to specify ρ in advance, because ρ depends on the censoring distribution, and for paired data on the true effect of treatment, in addition to the dependence between the event times within the cluster. As a result of these concerns, this formula may be most useful in a maximum information trial (Lan & DeMets, 1989). In such a trial, one chooses to fix the maximum information,

$$K \Big/ \left\{1 + \left(m\frac{p_1 p_2}{a_1 a_2} r^2 - 1\right)\rho\right\},$$

at the end of the trial. The end of the trial is then defined as the time at which the observed information, based on the observed number of events and observed martingale correlation, equals or exceeds the maximum information requested. Equivalently, we can restate the information requirement in terms of numbers of events; that is, define the end of the trial as the time at which the observed number of events K equals or exceeds the correlation-adjusted number of events $K(\rho)$ based on equation (4). Here, ρ is evaluated from masked data, that is without access to treatment assignments; note that ρ is the intraclass correlation coefficient of $\int_0^\infty \hat{U}_n(s)d\hat{M}_{ijk}(s)$, which may be calculated using analysis of variance techniques. For the unweighted log-rank statistic, $\hat{M}_{ijk}(\infty)$ are available as martingale residuals from the Cox regression procedures in SAS, S-Plus and R.

In practice, initial sample size calculations would be based on crude estimates of the event rate and ρ. Simulation can be used to evaluate ρ for a variety of plausible failure time distributions, dependence structures and censoring distributions. The sample size should be based on a conservative choice for ρ, such as the largest ρ value for clustered data or the smallest ρ value for paired data. One could then either simply continue the trial until the target information, or equivalently correlation-adjusted number of events, is reached or make mid-course corrections, i.e. increase enrolment or length of follow-up, to achieve the required target information, or equivalently correction-adjusted number of events. An example illustrating the implementation of a maximum information trial with paired survival data is provided in § 6.

5. SIMULATION STUDIES

To assess the small-sample properties of the cluster log-rank test and to verify the usefulness of the above sample-size formula, we simulated a series of clinical trials with clustered survival data. In each simulation, we assumed uniform recruitment of clusters of two replicates over the first year of the trial with a fixed study length of three years, i.e. a maximum duration trial. The marginal distribution of the event times was exponential corresponding to constant hazard. The marginal overall event rate in the control group was set to 25%, 50% or 75% corresponding to an annual hazard rate of 0·115, 0·255 or 0·555. A marginal proportional hazards alternative for the treatment effect was used, with hazard ratios of 1·00, i.e. no treatment effect, 0·75 or 0·50. The dependence within a cluster was introduced using a gamma frailty with mean 1 and variance ϕ, with ϕ taking values 0·5, 1·0 or 2·0. The within-cluster martingale correlation ρ was evaluated through Monte Carlo integration using 10 000 simulations for each scenario; the Monte Carlo standard error for the resulting asymptotic power calculations is no greater than 0·0025 in any scenario. Note that ρ is the martingale correlation calculated assuming that the null hypothesis of no treatment effect is true. When treatments are assigned within clusters, the correlation ρ will be attenuated by the hypothesised treatment effect.

In the first set of simulations, treatments were assigned to whole clusters. Sample sizes of 50, 100 or 150 clusters per group were used, corresponding to 100, 200 or 300 replicates per group. For each set of parameter values, 10 000 simulations were performed. From each study, we calculated the rejection rates for a nominal two-sided 5% level test using the normal approximation for the cluster log-rank statistic, CLR, and the log-rank statistic based on time-to-first-event, TFE. Note that the cluster log-rank statistic uses the empirical variance estimate from Theorem 2 and does not require calculation of the within-cluster martingale correlation ρ. Results of these simulation studies are provided in Table 1. For comparison, the asymptotic power of the cluster log-rank test, AP, is also provided.

Both the cluster log-rank test and the log-rank test using time-to-first-event maintain empirical Type I error rates close to the desired 5% level. The cluster log-rank test consistently shows greater power than the log-rank test using time-to-first-event. This expected advantage

Table 1. *Simulated size and power of nominal* 5% *level tests for clustered survival data with treatments assigned to clusters of size* 2, *where* HR *denotes the hazard ratio,* r_0 *is the overall event rate,* ϕ *is the variance of the gamma frailty,* ρ *is the martingale correlation within clusters, and* N *is the number of clusters*

				HR = 1·0		HR = 0·75			HR = 0·5		
r_0	ϕ	ρ	N	CLR	TFE	CLR	AP	TFE	CLR	AP	TFE
0·25	0·5	0·118	50	0·052	0·053	0·145	0·146	0·132	0·514	0·528	0·467
			100	0·045	0·047	0·246	0·249	0·216	0·817	0·819	0·764
			150	0·048	0·050	0·352	0·348	0·308	0·933	0·940	0·902
	1·0	0·192	50	0·047	0·050	0·138	0·140	0·126	0·492	0·503	0·434
			100	0·052	0·052	0·241	0·236	0·208	0·790	0·794	0·714
			150	0·051	0·049	0·324	0·330	0·275	0·918	0·926	0·868
	2·0	0·351	50	0·048	0·049	0·118	0·128	0·102	0·434	0·455	0·370
			100	0·052	0·052	0·216	0·214	0·180	0·740	0·743	0·640
			150	0·053	0·050	0·300	0·298	0·241	0·895	0·893	0·816
0·50	0·5	0·191	50	0·057	0·054	0·238	0·240	0·181	0·796	0·807	0·679
			100	0·049	0·048	0·423	0·426	0·321	0·975	0·979	0·927
			150	0·050	0·047	0·573	0·584	0·448	0·998	0·998	0·990
	1·0	0·361	50	0·051	0·051	0·222	0·216	0·171	0·760	0·753	0·619
			100	0·049	0·052	0·389	0·382	0·287	0·968	0·963	0·896
			150	0·050	0·051	0·523	0·529	0·393	0·996	0·996	0·972
	2·0	0·534	50	0·050	0·051	0·193	0·196	0·142	0·698	0·703	0·543
			100	0·052	0·052	0·350	0·346	0·254	0·941	0·941	0·837
			150	0·053	0·051	0·487	0·482	0·352	0·991	0·991	0·948
0·75	0·5	0·294	50	0·054	0·050	0·326	0·321	0·216	0·934	0·926	0·789
			100	0·051	0·048	0·569	0·561	0·386	0·998	0·998	0·976
			150	0·055	0·055	0·750	0·735	0·539	1·000	1·000	0·998
	1·0	0·479	50	0·054	0·052	0·300	0·287	0·199	0·901	0·890	0·731
			100	0·052	0·051	0·510	0·507	0·346	0·996	0·995	0·953
			150	0·055	0·055	0·678	0·678	0·469	1·000	1·000	0·994
	2·0	0·707	50	0·055	0·048	0·256	0·255	0·182	0·852	0·843	0·676
			100	0·052	0·049	0·463	0·452	0·322	0·990	0·987	0·930
			150	0·053	0·051	0·621	0·615	0·439	0·999	0·999	0·989

CLR, cluster log-rank test; TFE, log-rank test using time to first event only; AP, asymptotic power of cluster log-rank test based on equation (3)

presumably results from the more efficient use of available information on treatment differences. The empirical power of the cluster log-rank test is quite close to the asymptotic power calculations; the largest discrepancies are of the order of 1–2%. This observation suggests that, when treatments are assigned to clusters, the approximations used in its derivation are reasonably accurate, even in small or moderate samples.

Table 2. *Simulated size and power of nominal* 5% *level tests for paired survival data with treatments assigned within clusters of size* 2, *where* HR *denotes the hazard ratio,* r_0 *is the overall event rate,* ϕ *is the variance of the gamma frailty,* ρ *is the martingale correlation within clusters, and* N *is the number of clusters*

			HR = 1·0		HR = 0·75			HR = 0·5		
r_0	ϕ	N	CLR	OLR	CLR	AP	OLR	CLR	AP	OLR
			$\rho = 0{\cdot}118$		$\rho = 0{\cdot}097$			$\rho = 0{\cdot}085$		
0·25	0·5	50	0·043	0·037	0·099	0·106	0·086	0·311	0·355	0·302
		100	0·050	0·038	0·160	0·171	0·138	0·579	0·612	0·555
		200	0·047	0·036	0·288	0·297	0·258	0·872	0·888	0·857
			$\rho = 0{\cdot}192$		$\rho = 0{\cdot}187$			$\rho = 0{\cdot}151$		
	1·0	50	0·048	0·029	0·102	0·113	0·076	0·334	0·378	0·295
		100	0·048	0·030	0·177	0·185	0·128	0·612	0·645	0·559
		200	0·048	0·027	0·310	0·324	0·241	0·900	0·909	0·868
			$\rho = 0{\cdot}351$		$\rho = 0{\cdot}314$			$\rho = 0{\cdot}266$		
	2·0	50	0·045	0·016	0·110	0·127	0·061	0·377	0·426	0·283
		100	0·051	0·018	0·194	0·211	0·106	0·679	0·708	0·562
		200	0·048	0·016	0·371	0·373	0·235	0·934	0·944	0·882
			$\rho = 0{\cdot}191$		$\rho = 0{\cdot}187$			$\rho = 0{\cdot}155$		
0·50	0·5	50	0·047	0·027	0·168	0·188	0·129	0·628	0·661	0·573
		100	0·048	0·028	0·327	0·329	0·259	0·906	0·919	0·878
		200	0·051	0·030	0·565	0·574	0·485	0·997	0·997	0·995
			$\rho = 0{\cdot}361$		$\rho = 0{\cdot}325$			$\rho = 0{\cdot}287$		
	1·0	50	0·050	0·017	0·202	0·217	0·118	0·698	0·734	0·584
		100	0·052	0·016	0·370	0·384	0·231	0·945	0·955	0·899
		200	0·048	0·013	0·644	0·654	0·485	0·999	0·999	0·997
			$\rho = 0{\cdot}534$		$\rho = 0{\cdot}500$			$\rho = 0{\cdot}431$		
	2·0	50	0·050	0·005	0·255	0·277	0·080	0·812	0·825	0·598
		100	0·047	0·004	0·471	0·490	0·191	0·983	0·983	0·923
		200	0·046	0·003	0·783	0·782	0·480	1·000	1·000	0·999
			$\rho = 0{\cdot}294$		$\rho = 0{\cdot}274$			$\rho = 0{\cdot}239$		
0·75	0·5	50	0·049	0·020	0·285	0·291	0·186	0·871	0·881	0·809
		100	0·049	0·021	0·506	0·514	0·374	0·993	0·993	0·985
		200	0·057	0·022	0·797	0·805	0·691	1·000	1·000	1·000
			$\rho = 0{\cdot}479$		$\rho = 0{\cdot}452$			$\rho = 0{\cdot}392$		
	1·0	50	0·049	0·008	0·351	0·368	0·156	0·935	0·940	0·829
		100	0·051	0·006	0·618	0·632	0·359	0·998	0·999	0·992
		200	0·048	0·007	0·903	0·901	0·735	1·000	1·000	1·000
			$\rho = 0{\cdot}707$		$\rho = 0{\cdot}654$			$\rho = 0{\cdot}566$		
	2·0	50	0·047	0·000	0·508	0·533	0·103	0·990	0·986	0·873
		100	0·045	0·001	0·822	0·824	0·325	1·000	1·000	0·999
		200	0·052	0·000	0·985	0·983	0·772	1·000	1·000	1·000

CLR, cluster log-rank test; OLR, ordinary log-rank test ignoring clustering; AP, asymptotic power of cluster log-rank test based on equation (3)

In our second set of simulations, treatments were assigned within clusters, producing paired survival data. A sample size of 50, 100 or 200 clusters was used, corresponding to 50, 100 or 200 replicates per group. For each set of parameter values, 10 000 simulations were performed. From each study, we calculated the rejection rates for a nominal two-sided 5% level test using the normal approximation for the cluster log-rank statistic, CLR, and the log-rank statistic ignoring the clustering, OLR. Note that the cluster log-rank statistic uses the empirical variance estimate from Theorem 2 and does not require calculation of the within cluster martingale correlation ρ. Results of these simulation studies are provided in Table 2. For comparison, the asymptotic power of the cluster log-rank test, AP, is also provided as in Table 1.

The cluster log-rank test consistently maintains a Type I error rate close to the nominal 5% level. The log-rank test ignoring clustering, on the other hand, is consistently conservative, the degree of conservativeness increasing with the within-cluster martingale correlation. For large correlations, the Type I error rate for the log-rank test ignoring clustering is effectively zero. Consequently, the cluster log-rank test shows substantially greater power than the log-rank test ignoring clustering in most settings, emphasising the value of accounting for correlation within clusters. With paired data, the empirical power of the cluster log-rank test is quite similar to the asymptotic power calculations; most discrepancies are of the order of 1–2% with discrepancies of 4% in a few cases. This observation suggests that, when treatments are assigned within clusters, the sample-size formula is reasonably accurate, even for trials of small or moderate size.

6. Example: Early Treatment Diabetic Retinopathy Study

An example of a trial including clustered survival data is the Early Treatment Diabetic Retinopathy Study, which enrolled 3711 patients with non-proliferative or early-proliferative diabetic retinopathy in both eyes. Enrolment in the study lasted from April 1980 until July 1985. The final follow-up visit occurred in June 1989. The study included a multifactorial treatment design with several different endpoints. For illustrative purposes, we will consider only one of the questions of interest.

One eye per patient was randomised to early photocoagulation and the other to deferral of photocoagulation until the development of high-risk diabetic retinopathy. A principal endpoint of the study was time to severe visual loss or vitrectomy. Severe visual loss is defined as a Snellen visual acuity below 5/200 at two consecutive visits. The original study design was based on time to severe visual loss alone. However, because vitrectomy saved an unknown number of eyes from severe visual loss, the combined endpoint of severe visual loss or vitrectomy was used in subsequent reports (ETDRS Research Group, 1991b).

The design assumptions for the study included a five-year rate of severe visual loss or vitrectomy of 10% in eyes assigned to deferral and a five-year rate of 6% in eyes assigned to early photocoagulation (ETDRS Research Group, 1991a). These rates translate into a hazard ratio of 0·587, or $\gamma = -0{\cdot}533$. A two-sided Type I error rate of 1% was specified to account for the multiplicity of tests being performed. The desired power for the alternative was 98%. Based on the formulae in § 4·2, the required information to achieve the design requirements is $K/(1-\rho) = 303$, where K is the number of events and ρ is the within cluster, i.e. participant, martingale correlation. In a maximum information trial design, we define the end of the trial as the time at which the observed information, $K/(1-\rho)$, equals or exceeds the design requirement, 303. As noted previously, we can rewrite the information requirement in terms of numbers of events; that is, define the end

of the trial as the time at which the observed number of events, K, equals or exceeds the correlation-adjusted number of events, $K(\rho) = 303(1-\rho)$, where ρ, the current within cluster martingale correlation, is evaluated using masked data.

For illustration, we consider monitoring the Early Treatment Diabetic Retinopathy Study at six-month intervals starting on 9 April 1985. In Table 3, we present the available masked information at these time points. Based on the correlation-adjusted number of events $K(\rho)$, we would stop the trial at the third look on 9 April 1986. At that time, 202 events had been observed, the martingale correlation was 0·337, and the required number of events for 98% power was 201. As of 9 April 1986, the cluster log-rank statistic, favouring early photocoagulation, was 2·84 ($p = 0{\cdot}0084$), and the log-rank statistic ignoring clustering was 2·15 ($p = 0{\cdot}032$). The cluster log-rank statistic meets the specified 1% critical value at this time, while the log-rank statistic ignoring clustering does not.

Table 3. *Monitoring of the Early Treatment Diabetic Retinopathy Study data at six-month intervals starting on* 9 *April* 1985, *where N is the observed number of events*, ρ *is the current within-cluster martingale correlation, and* $K(\rho) = 303(1-\rho)$ *is the correlation-adjusted required number of events based on equation* (4)

Analysis date	K	ρ	$K(\rho)$	$K \geqslant K(\rho)$?	$K \geqslant K(0)$?
9 April 1985	125	0·401	182	No	No
9 October 1985	165	0·359	195	No	No
9 April 1986	202	0·337	201	Yes	No
9 October 1986	240	0·318	207	Yes	No
9 April 1987	276	0·314	208	Yes	No
9 October 1987	318	0·316	208	Yes	Yes
9 April 1988	352	0·331	203	Yes	Yes
9 October 1988	378	0·330	203	Yes	Yes
9 April 1989	388	0·325	205	Yes	Yes

If one ignored the correlation within clusters and stopped the trial after $K(0) = 303$ events were observed, the trial would continue until the sixth look on 9 October 1987. At that time, 318 events had been observed. Note that, on 9 October 1987, the martingale correlation had attenuated slightly to 0·316 and the required number of events for 98% power had increased slightly to 208 events. This emphasises the fact that the correlation-adjusted number of events is a moving target and cannot be easily specified in advance. As of 9 October 1987, the cluster log-rank statistic was 4·30 ($p = 0{\cdot}000017$) and the log-rank statistic ignoring clustering was 3·56 ($p = 0{\cdot}00037$). Both statistics easily meet the specified 1% significance level. Accounting for the within cluster correlation, in both the event-driven design, i.e. required number of events, and the analysis, reduced the study length by 1·5 years and required one-third fewer events, namely 201 as compared with 303.

7. DISCUSSION

The methods presented here can accommodate variable cluster sizes and any noninformative correlated censoring mechanism within clusters such as common censoring times, independent

censoring and so on. Extensions of these methods which incorporate stratification are given in an unpublished technical report by the authors from the Department of Biostatistics and Medical Informatics at the University of Wisconsin-Madison.

Observations from clusters of different sizes and types all contribute equally to the cluster log-rank statistic. The clustering of observations only has impact on the estimated variance of the statistic. However, in the context of paired data with some singleton observations, discussion in Manatunga & Oakes (1999) and Murray (2001) suggests that estimation and testing techniques should place higher value on more informative complete pairs, especially in the presence of high correlation. One approach might be to weight the clusters with the weights being proportional to the inverse of the within cluster variance. These weights would only depend on the martingale correlation, the cluster size and the treatment breakdown within the cluster.

The sample-size formula presented here is especially suitable for event-driven trial designs in which a specific amount of observed information, conveniently expressed as a correlation-adjusted required number of events, is targeted instead of a specific length of follow-up. Event-driven designs allow for mid-course corrections based on masked data to maintain power in the face of lags in recruitment and lower than expected event rates while still controlling the Type I error rates. In combination with the methods for group sequential monitoring presented by Murray (2000), all facets of trial monitoring used with non-clustered survival data are available for paired survival data. The extension of this group sequential monitoring framework to more general cases of clustered survival data is an area for future work.

Acknowledgement

R. E. Gangnon was supported by a contract with Novartis Pharma AG. M. R. Kosorok was supported by a grant from the U.S. National Cancer Institute.

Appendix

Proofs

Proof of Lemma 1. If we use product notation for the product integral,

$$\hat{S}_n(t) = \prod_{0 \leqslant s \leqslant t} \{1 - d\hat{\Lambda}_n(s)\},$$

where $\hat{\Lambda}_n(t) \equiv \int_{[0,t]} (d\bar{N}_1 + d\bar{N}_2)/(\bar{Y}_1 + \bar{Y}_2)$. We need to compare this with S_0 in two steps. First, define

$$\tilde{S}_n(t) \equiv \prod_{0 \leqslant s \leqslant t \wedge \tilde{T}_n} \{1 - d\tilde{\Lambda}_n(s)\},$$

where

$$\tilde{T}_n = \sup\{t : \bar{Y}_1(t) \wedge \bar{Y}_2(t) > 0\}, \quad \tilde{\Lambda}_j^n(t) \equiv \int_{[0,t]} (\bar{Y}_1 + \bar{Y}_2)^{-1} \times (\bar{Y}_1 d\Lambda_1^n + \bar{Y}_2 d\Lambda_2^n).$$

The Duhamel equation yields

$$\hat{S}_n(t \wedge \tilde{T}_n) - \tilde{S}_n(t \wedge \tilde{T}_n) = -\tilde{S}_n(t \wedge \tilde{T}_n) \int_{[0, t \wedge \tilde{T}_n]} \frac{\hat{S}_n(x-)}{\tilde{S}_n(x)} \left\{ \frac{d\bar{M}_1(x) + d\bar{M}_2(x)}{\bar{Y}_1(x) + \bar{Y}_2(x)} \right\}, \tag{A1}$$

where $\bar{M}_j(t) \equiv \bar{N}_j(t) - \int_{[0,t]} \bar{Y}_j(x) d\Lambda_j^n(x)$ $(j = 1, 2)$ and

$$\tilde{S}_n(t \wedge \tilde{T}_n) - S_0(t \wedge \tilde{T}_n)$$

$$= -S_0(t \wedge \tilde{T}_n) \int_{[0, t \wedge \tilde{T}_n]} \frac{\tilde{S}_n(x-)}{S_0(x)} \times \left[\frac{\bar{Y}_1(x)\{d\Lambda_1^n(x) - d\Lambda_0(x)\} + \bar{Y}_2(x)\{d\Lambda_2^n(x) - d\Lambda_0(x)\}}{\bar{Y}_1(x) + \bar{Y}_2(x)} \right]. \tag{A2}$$

We know that $\hat{S}_n$ and $\tilde{S}_n$ are both survival functions and that N_{ijk} and Y_{ijk} are both bounded monotone processes. Thus standard empirical process methods combined with integration by parts establish that both (A1) and (A2) converge to zero in probability, uniformly over $t \in K$, for any closed $K \subset I$. The result now follows since, with probability 1, $\tilde{T}_n \geqslant \sup K$ for all n large enough. □

Proof of Theorem 1. Let $\bar{M}_{.j.} \equiv \sum_{i=1}^{n} \bar{M}_{ij.}$. Then, from (1),

$$H_n = n^{-1/2} \int_I \hat{U}_n(s) \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} \left\{ \frac{d\bar{M}_{.1.}(s)}{\bar{Y}_1(s)} - \frac{d\bar{M}_{.2.}(s)}{\bar{Y}_2(s)} \right\}$$
$$+ n^{-1} \int_I \hat{U}_n(s) \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} \sqrt{n}\{d\Lambda_1^n(s) - d\Lambda_2^n(s)\}, \tag{A3}$$

for all n large enough. Since Y_{ijk} and N_{ijk} are monotone processes, they form 'manageable' arrays, and both Donsker and Glivenko–Cantelli results for independent but not identically distributed data (Pollard, 1990) will generally apply. These facts directly yield convergence of the second term on the right-hand side of (A3) to μ in probability.

Establishing convergence of the first term on the right-hand side of (A3) to a zero-mean Gaussian process with variance σ^2 is more complicated. For each $j = 1, 2$, we need to verify that

$$n^{-1/2} \int_I \left\{ \hat{U}_n(s) \frac{\bar{Y}_{j'}(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} - U(s) \frac{\pi_{j'}^n(s)}{\pi_1^n(s) + \pi_2^n(s)} \right\} d\bar{M}_{.j.}(s) = o_p(1) \tag{A4}$$

and that

$$\sup_{n \geqslant 1} \max_{1 \leqslant i \leqslant n} \max_{1 \leqslant m_{ij}} E \left| \int_I U(s) \frac{\pi_{j'}^n(s)}{\pi_1^n(s) + \pi_2^n(s)} dM_{ijk}(s) \right|^r < \infty \tag{A5}$$

for some $r > 2$, where $j' = 3 - j$. The result will then follow from the fact that $\sigma_n^2 \to \sigma^2$ and the Lindeberg–Feller central limit theorem.

For $r = 3$, it is not difficult to prove (A5). If the region of integration in (A4) is replaced by any closed subinterval $K \subset I$, the integral will converge to zero in probability by a minor modification of Lemma A·3 of Bilias et al. (1997), applying it for a second time if needed for $\hat{U}_n$. If $\tau \in I$, we are finished. If $\tau \notin I$, then let $\{t_n\} \in I$ be any increasing sequence with $t_n \to \tau$. The proof is complete if we can verify that

$$n^{-1/2} \int_{(t_{n,\tau})} \hat{U}_n(s) \frac{\bar{Y}_{j'}(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} d\bar{M}_{.j.}(s) = o_p(1),$$
$$n^{-1/2} \int_{(t_{n,\tau})} U(s) \frac{\pi_{j'}^n(s)}{\pi_1^n(s) + \pi_2^n(s)} d\bar{M}_{.j.}(s) = o_p(1).$$

The arguments for establishing these last two expressions are lengthy, and we omit them. The details are available from the authors. □

Proof of Theorem 2. For simplicity, we will suppress the time argument inside integrals over time. Fix $j \in \{1, 2\}$, and let $K \subset I$ be a closed subinterval. Standard empirical process arguments yield

$$n^{-1} \sum_{i=1}^{n} \left\{ \int_K \left(\hat{U}_n \frac{\bar{Y}_{j'}}{\bar{Y}_1 + \bar{Y}_2} - U \frac{\pi_{j'}^n}{\pi_1^n + \pi_2^n} \right) d\check{M}_{ij.} \right\}^2 = o_p(1).$$

Similarly,

$$n^{-1} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{m_{ij}} \int_K U \frac{\pi_{j'}^n}{\pi_1^n + \pi_2^n} Y_{ijk} \left(\frac{d\bar{N}_j}{\bar{Y}_j} - d\Lambda_j^n \right) \right\}^2 \leqslant m_0^2 \left\{ \sup_{t \in K} \int_{[0,t]} U \frac{\pi_{j'}^n}{\pi_1^n + \pi_2^n} \left(\frac{d\bar{N}_j}{\bar{Y}_j} - d\Lambda_j^n \right) \right\}^2 = o_p(1).$$

Since

$$\left|\sum_{i=1}^{n} a_i^2 - \sum_{i=1}^{n} b_i^2\right| \leqslant 2\left(\sum_{i=1}^{n} b_i^2\right)^{1/2}\left\{\sum_{i=1}^{n}(a_i-b_i)^2\right\}^{1/2} + \sum_{i=1}^{n}(a_i-b_i)^2,$$

for real numbers (a_i, b_i) $(i = 1, \ldots, n)$, we now have that

$$n^{-1}\sum_{i=1}^{n}\left(\int_K \hat{U}_n \frac{\bar{Y}_{j'}}{\bar{Y}_1+\bar{Y}_2} d\check{M}_{ij.}\right)^2 - n^{-1}\sum_{i=1}^{n}\left(\int_K U\frac{\pi^n_{j'}}{\pi^n_1+\pi^n_2} d\bar{M}_{ij.}\right)^2 = o_p(1).$$

If $\tau \in I$, we are finished. Otherwise, we need to use careful arguments to establish that the contributions in the tails are asymptotically negligible, as was done in Theorem 1. The details are available from the authors. □

References

Aalen, O. O. (1987). Two examples of modelling heterogeneity in survival analysis. *Scand. J. Statist.* **14**, 19–25.

Albers, W. (1988). Combined rank tests for randomly censored paired data. *J. Am. Statist. Assoc.* **83**, 1159–62.

Andersen, P. K., Borgan, O., Gill, R. D. & Keiding, N. K. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Bilias, Y., Gu, M. & Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *Ann. Statist.* **25**, 662–82.

Dabrowska, D. M. (1989). Rank tests for matched pair experiments with censored data. *J. Mult. Anal.* **28**, 88–114.

Dabrowska, D. M. (1990). Signed-rank tests for censored matched pairs. *J. Am. Statist. Assoc.* **85**, 478–85.

ETDRS Research Group (1991a). Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics. ETDRS Report Number 7. *Ophthalmology* **98**, 741–56.

ETDRS Research Group (1991b). Early photocoagulation for diabetic retinopathy. ETDRS Report Number 9. *Ophthalmology* **98**, 766–85.

Fleming, T. R. & Harrington, D. P. (1981). A class of hypothesis tests for one and two sample censored survival data. *Commun. Statist.* A **10**, 763–94.

Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Huang, Y. (1999). The two-sample problem with induced dependent censorship. *Biometrics* **55**, 1108–13.

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

Lan, K. K. G. & DeMets, D. L. (1989). Group sequential procedures: Calandar versus information time. *Statist. Med.* **8**, 1191–8.

Manatunga, A. K. & Oakes, D. (1999). Parametric analysis for matched pair survival data. *Lifetime Data Anal.* **5**, 371–87.

Mantel, N. & Ciminera, J. (1979). Use of log-rank scores in the analysis of litter-matched data on time to tumor appearance. *Cancer Res.* **39**, 4308–15.

Murray, S. (2000). Nonparametric rank-based methods for group sequential monitoring of paired censored survival data. *Biometrics* **56**, 984–90.

Murray, S. (2001). Using weighted Kaplan-Meier statistics in nonparametric comparisons of paired censored survival data. *Biometrics* **57**, 361–8.

O'Brien, P. C. & Fleming, T. R. (1987). A paired Prentice-Wilcoxon test for censored paired data. *Biometrics* **43**, 169–80.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Hayward, CA: Institute of Mathematical Statistics.

Rosner, B. & Glynn, R. J. (1997). Multivariate methods for clustered ordinal data with applications to survival analysis. *Statist. Med.* **16**, 357–72.

Schoenfeld, D. A. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* **39**, 499–503.

Wei, L. J. (1980). A generalized Gehan and Gilbert test for paired observations that are subject to arbitrary right censorship. *J. Am. Statist. Assoc.* **75**, 634–7.

Wei, L. J., Lin, D. Y. & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Assoc.* **84**, 1065–73.

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

[*Received July* 2001. *Revised August* 2003]