STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4505–4519
Published online 9 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2864

Analyzing survival curves at a fixed point in time

John P. Klein[1,*,†], Brent Logan[1], Mette Harhoff[2] and Per Kragh Andersen[2]

[1]*Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, U.S.A.*
[2]*Department of Biostatistics, University of Copenhagen, Ø. Farimagsgade 5, PB 2099, DK 1014 Copenhagen K, Denmark*

SUMMARY

A common problem encountered in many medical applications is the comparison of survival curves. Often, rather than comparison of the entire survival curves, interest is focused on the comparison at a fixed point in time. In most cases, the naive test based on a difference in the estimates of survival is used for this comparison. In this note, we examine the performance of alternatives to the naive test. These include tests based on a number of transformations of the survival function and a test based on a generalized linear model for pseudo-observations. The type I errors and power of these tests for a variety of sample sizes are compared by a Monte Carlo study. We also discuss how these tests may be extended to situations where the data are stratified. The pseudo-value approach is also applicable in more detailed regression analysis of the survival probability at a fixed point in time. The methods are illustrated on a study comparing survival for autologous and allogeneic bone marrow transplants. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: generalized linear models; pseudo-value approach; variance stabilizing transformation; Kaplan–Meier estimators; censored data

1. INTRODUCTION

In many applications of survival analysis techniques clinical investigators are interested in comparing two or more survival curves at a set of time points. These comparisons are often made in providing summary univariate statistics for a study prior to more extensive modelling. They may be made to provide a comparison of 'cure rates' of the disease under study. For example, one may

[*]Correspondence to: John P. Klein, Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226, U.S.A.
[†]E-mail: klein@mcw.edu

Contract/grant sponsor: National Cancer Institute; contract/grant number: R01 CA54706-10

Received 16 December 2004
Accepted 21 January 2007

want to compare cancer survival rates in two treatments at 5 years with the thought that 5-year survival means a ‘cure’ of the cancer. Comparisons at fixed points in time are often of interest when the survival curves of the treatments are known to cross. Crossing survival curves may be a consequence of crossing hazards and it is well known that for this situation many standard tests, such as the log-rank or Wilcoxon tests, will fail to pick up differences in survival curves. Non-parametric tests with higher power against crossing hazard alternatives include Kolmogorov–Smirnov and Cramér–von Mises types of tests [1, 2] and the median test of Brookmeyer and Crowley [3].

1.1. Example

As an example consider a study comparing the survival of patients given a matched sibling donor bone marrow transplant to an autologous transplant for leukaemia. The two survival curves can be expected to cross at some point in time. Typically, allogeneic transplants will tend to have a higher mortality rate in the first few months after treatment due to the toxicity of the high doses of chemotherapy used in the preparative regimes as well as the high mortality due to graft-*versus*-host disease. Patients with autografts do not develop graft-*versus*-host disease so their early mortality is lower, but they do not have the graft-*versus*-leukaemia effect found in allograft patients so they tend to have higher relapse-related mortality which occurs later in time. Direct comparisons of the entire curves are difficult because of the crossing hazards (and survival functions). Many studies comparing these two groups will base the comparison at a fixed point in time where it is expected that the two survival curves have flattened out.

Such an example is provided by a study comparing autologous and allogeneic bone marrow transplants for follicular lymphoma [4]. The sample contained 175 patients with an human leukocyte antigen-identical sibling allogeneic transplant and 596 patients with an unpurged autologous transplant. Of the patients receiving an allogeneic transplant 95 (54%) were female while 319 (53%) of the autologous patients were female. As shown in Table I there was a difference in the distribution of donor type over years of transplant, which may require adjustment for in comparing the two transplant types.

Of interest is a comparison of the disease-free survival probabilities (i.e. the probability a patient is alive without recurrence of disease) in the two treatment arms. Figure 1 shows the two disease-free survival functions, which appear to cross at about 1 year. A comparison of the two curves using the log-rank test has a p-value of 0.44 and using the Wilcoxon test a p-value of 0.17, suggesting that the two survival curves are the same. Of clinical interest is a comparison of the curves at 6 months and 1 year.

In most applications, a naive test based on the comparison of Kaplan–Meier estimators [5–7] of the survival function or Nelson–Aalen estimators of the cumulative hazard rate is used. While this

Table I. Distribution of types of transplant over the years 1990–1998.

Donor	1990	1991	1992	1993	1994
Allo	8 (5%)	11 (6%)	17 (10%)	13 (7%)	16 (9%)
Auto	13 (2%)	24 (4%)	55 (9%)	89 (15%)	91 (15%)
	1995	1996	1997	1998	
Allo	24 (14%)	24 (14%)	38 (22%)	24 (14%)	
Auto	91 (15%)	86 (14%)	73 (12%)	74 (12%)	

DOI: 10.1002/sim

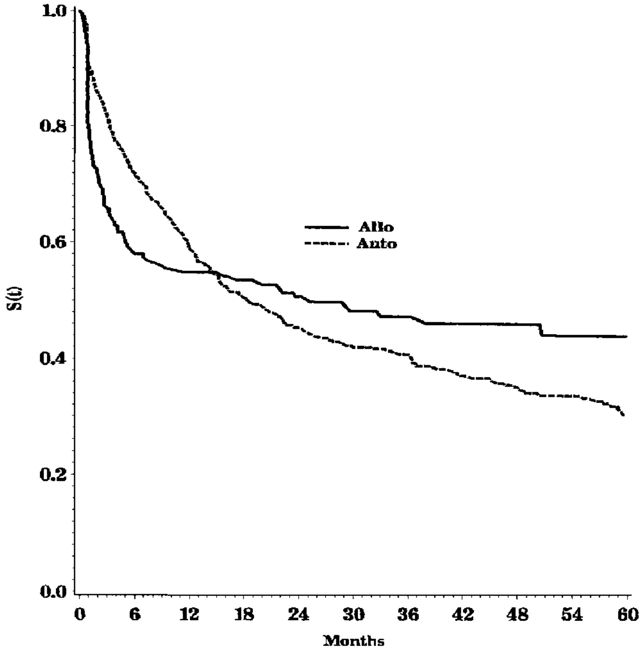

Figure 1. Comparison of survival curves for auto- and allo-transplanted leukaemia patients.

test may have good properties for very large sample sizes its performance for small-to-moderate sample sizes may be suspect. As noted by Borgan and Liestøl [8] the sample size needed for the naive univariate confidence interval for a survival function to have the correct coverage probability is much larger than that needed for a suitably transformed confidence interval. Bie *et al.* [9] made a similar observation for the cumulative hazard function. It would seem that tests for the equality of the two survival functions based on transformed survival or cumulative hazard functions may perform better than the usual untransformed tests. These tests are considered in the next section where we also study a test derived from a censored data logistic regression technique based on using generalized estimating equations (GEEs) and pseudo-values from the data [10].

In Section 3, we compare these tests based on a Monte Carlo study. In Section 4, we discuss versions of these tests useful for stratified samples and we further elaborate on the pseudo-value approach and study how it may be applied for regression analysis of the survival probability at a fixed point in time. In Section 5, we revisit our example and in Section 6, we discuss how the tests can be extended to more than two samples.

2. TWO SAMPLE TESTS

We shall, for simplicity, consider the problem of comparing two survival curves at a single point in time. In the final section, we shall make some brief comments on the $K>2$ sample problem. The data from the kth sample, $k=1,2$, consist of n_k subjects with distinct event times t_{ki}. At time t_{ki}, let d_{ki} be the number of events and Y_{ki} the number at risk in the kth group. Using this notation

DOI: 10.1002/sim

the Kaplan–Meier estimators [5] are given by

$$\hat{S}_k(t)=\prod_{t_{ki}\leqslant t}\left(1-\frac{d_{ki}}{Y_{ki}}\right)$$

with estimated variance

$$\hat{V}(\hat{S}_k(t))=\hat{S}_k(t)^2\hat{\sigma}_k(t)^2$$

where

$$\hat{\sigma}_k(t)^2=\sum_{t_{ki}\leqslant t}\frac{d_{ki}}{Y_{ki}(Y_{ki}-d_{ki})}$$

(Greenwood's formula). The Nelson–Aalen estimators and associated estimated variances are given by

$$\hat{H}_k(t)=\sum_{t_{ki}\leqslant t}\frac{d_{ki}}{Y_{ki}}\quad\text{and}\quad\hat{V}(\hat{H}_k(t))=\sum_{t_{ki}\leqslant t}\frac{d_{ki}}{Y_{ki}^2}$$

The first test is the naive test of $H_0: S_1(t)=S_2(t)$ for some fixed t. This test is given by

$$X_1^2=\frac{(\hat{S}_1(t)-\hat{S}_2(t))^2}{\hat{S}_1(t)^2\hat{\sigma}_1(t)^2+\hat{S}_2(t)^2\hat{\sigma}_2(t)^2}\tag{1}$$

This simple test has an asymptotic chi-squared distribution with one degree of freedom when $S_1(t)=S_2(t)$; however, as we shall see in the next section, it requires large sample sizes to maintain its level and rejects too often when the null hypothesis is true. The analogous chi-squared test based on the Nelson–Aalen estimators is given by

$$Y_1^2=\frac{(\hat{H}_1(t)-\hat{H}_2(t))^2}{\hat{V}(\hat{H}_1(t))+\hat{V}(\hat{H}_2(t))}$$

Classes of test statistics are based on suitable transformations ϕ of either $\hat{S}_k(t)$ or $\hat{H}_k(t)$. Thus, by the delta-method, $\phi(\hat{S}_k(t))$ is asymptotically normal with mean $\phi(S_k(t))$ and estimated variance is given by

$$\hat{V}(\hat{S}_k(t))(\phi'(\hat{S}_k(t)))^2$$

A similar result holds for transformations of the Nelson–Aalen estimator. However, as tests based on the Nelson–Aalen estimator turned out not to outperform those based on $\phi(\hat{S}_k(t))$ they will not be considered further here. Tests based on a variance stabilizing transformation for the Kaplan–Meier estimator of the survival function are suggested by the better performance of transformed confidence intervals for the survival function [6–9]. Here we consider a number of transformations all leading to asymptotically χ_1^2-distributed statistics under the null hypothesis.

The first transformation is the logarithmic transformation for the survival function. This test, which should of course resemble that based on the Nelson–Aalen estimators, is given by

$$X_2^2=\frac{(\log(\hat{S}_1(t))-\log(\hat{S}_2(t)))^2}{\hat{\sigma}_1(t)^2+\hat{\sigma}_2(t)^2}\tag{2}$$

Another test is based on a $\log(-\log(\cdot))$ transformation for the survival function. This transformation has been found to be very useful in constructing confidence intervals and confidence bands for the survival function. The transformation gives about the correct coverage probability for a 95 per cent confidence interval for $S(t)$ based on as few as 25 observations with as much as 50 per cent censoring. The test of H_0 based on this transformation is:

$$X_3^2 = \frac{(\log(-\log(\hat{S}_1(t))) - \log(-\log(\hat{S}_2(t))))^2}{\hat{\sigma}_1(t)^2/(\log(\hat{S}_1(t)))^2 + \hat{\sigma}_2(t)^2/(\log(\hat{S}_2(t)))^2} \tag{3}$$

A third statistic is constructed based on an arcsine-square root transformation which has similar small sample coverage probabilities for confidence intervals for S as the log–log transformation [6–9]. Here the test statistic is

$$X_4^2 = \frac{\left(\arcsin\left(\sqrt{\hat{S}_1(t)}\right) - \arcsin\left(\sqrt{\hat{S}_2(t)}\right)\right)^2}{\hat{v}_1(t) + \hat{v}_2(t)} \tag{4}$$

where

$$\hat{v}_k(t) = \frac{\hat{S}_k(t)\hat{\sigma}_k(t)^2}{4(1-\hat{S}_k(t))}, \quad k = 1, 2$$

and the final transformation is the logit transformation with

$$X_5^2 = \frac{\left(\log\dfrac{\hat{S}_1(t)}{1-\hat{S}_1(t)} - \log\dfrac{\hat{S}_2(t)}{1-\hat{S}_2(t)}\right)^2}{\dfrac{\hat{\sigma}_1^2(t)}{(1-\hat{S}_1(t))^2} + \dfrac{\hat{\sigma}_2^2(t)}{(1-\hat{S}_2(t))^2}} \tag{5}$$

In the tests (1), (3), (4) and (5) we have normalized using a variance estimator that involves both $\hat{S}_1(t)$ and $\hat{S}_2(t)$. An alternative to this approach is to replace $\hat{S}_1(t)$ and $\hat{S}_2(t)$ with a common estimator, $\hat{S}_p(t)$ computed under the null, i.e. the Kaplan–Meier estimator based on the pooled sample of size $n_1 + n_2$.

When the censoring distributions in the two samples are equal one may go a step further and replace $\hat{\sigma}_k(t)$ by $((n_1+n_2)/n_k)\hat{\sigma}_p(t)$, where $\hat{\sigma}_p(t)$ is obtained from Greenwood's formula applied to the pooled sample. This leads to statistics

$$X_1^{*2} = \frac{(\hat{S}_1(t) - \hat{S}_2(t))^2}{\hat{S}_p(t)^2\hat{\sigma}_p(t)^2}\frac{n_1 n_2}{(n_1+n_2)^2} \tag{6}$$

$$X_2^{*2} = \frac{(\log(\hat{S}_1(t)) - \log(\hat{S}_2(t)))^2}{\hat{\sigma}_p(t)^2}\frac{n_1 n_2}{(n_1+n_2)^2} \tag{7}$$

DOI: 10.1002/sim

$$X_3^{*2} = \frac{(\log(-\log(\hat{S}_1(t))) - \log(-\log(\hat{S}_2(t))))^2}{\hat{\sigma}_p(t)^2/\log(\hat{S}_p(t))^2} \frac{n_1 n_2}{(n_1+n_2)^2} \tag{8}$$

$$X_4^{*2} = \frac{\left(\arcsin\left(\sqrt{\hat{S}_1(t)}\right) - \arcsin\left(\sqrt{\hat{S}_2(t)}\right)\right)^2}{\hat{v}_p(t)} \frac{n_1 n_2}{(n_1+n_2)^2} \tag{9}$$

where

$$\hat{v}_p(t) = \frac{\hat{S}_p(t)\hat{\sigma}_p(t)^2}{4(1-\hat{S}_p(t))}$$

and

$$X_5^{*2} = \frac{\left(\log\frac{\hat{S}_1(t)}{1-\hat{S}_1(t)} - \log\frac{\hat{S}_2(t)}{1-\hat{S}_2(t)}\right)^2}{\frac{\hat{\sigma}_p^2(t)}{(1-\hat{S}_p(t))^2}} \frac{n_1 n_2}{(n_1+n_2)^2} \tag{10}$$

For the naive statistic (1) the test (6) obtained in this way leads to that suggested by Sposto *et al.* [11] in another context.

A final pair of tests is based on a pseudo-value regression technique proposed by Andersen *et al.* [10] and Klein and Andersen [12]. Here we look at the pooled sample and compute the pooled sample Kaplan–Meier estimator, $\hat{S}_p(t)$, based on all n_1+n_2 observations and the Kaplan–Meier estimator based on the sample of size n_1+n_2-1 with the jth observation removed, $\hat{S}_p^{(j)}(t)$, $j=1,\ldots,n_1+n_2$. Define now the jth pseudo-value by

$$\hat{\theta}_j = (n_1+n_2)\hat{S}_p(t) - (n_1+n_2-1)\hat{S}_p^{(j)}(t), \quad j=1,\ldots,n_1+n_2 \tag{11}$$

If there is no censoring then $\hat{\theta}_j$ is simply the indicator that the jth patient was alive at time t. Note that as the pseudo-value approach is based on $\hat{S}_p(t)$ being a consistent estimator for the marginal survival distribution this method is expected to work better when censoring is the same in both groups.

The pseudo-values can be used in a generalized linear model to model the effects of covariates on outcome [10]. Let $g(\cdot)$ be a link function. Possible choices of the link function in models for the survival function are the logit link, $g(x)=\log(x/(1-x))$ or the complementary log–log function: $g(x)=\log(-\log(x))$. The complementary log–log function when applied to a survival function gives a proportional hazards representation at the single point in time t.

Given a $p+1$-vector of covariates Z_j for the jth patient in the pooled sample we assume a generalized linear model with

$$g(\theta_j) = \beta^{\mathrm{T}} Z_j, \quad j=1,\ldots,n_1+n_2$$

Here Z_j has a 1 in the first position to denote an intercept. We shall use the GEE approach [13] to estimate β. Let $\mu(\cdot)$ be the inverse function based on g. Define $\mathrm{d}\mu_j(\beta)$ to be the vector of partial derivatives of $\mu_j(\cdot)$ with respect to β. Let $V_j(\beta)$ be a working covariance matrix and let

$s_j = \mu(\beta^{\mathrm{T}} Z_j)$ be the model-based predicted value of $S(t \mid Z_j)$. The estimating equations to be solved are then

$$\begin{aligned} U(\beta) &= \sum_j \mathrm{d}\mu_j(\beta) V_j^{-1}(\beta)(\hat{\theta}_j - s_j) \\ &= \sum_j U_j(\beta) = 0 \end{aligned} \tag{12}$$

Let $\hat{\beta}$ be the solution to this equation and note that using results from Liang and Zeger [13], under standard regularity conditions, it follows that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with mean zero and a covariance that can be estimated consistently by a 'sandwich' estimator given by

$$\hat{\Sigma} = I(\hat{\beta})^{-1} \hat{V}(U(\hat{\beta})) I(\hat{\beta})^{-1}$$

where

$$I(\beta) = \sum_j \mathrm{d}\mu_j(\beta) V(\beta)^{-1} \mathrm{d}\mu_j(\beta)^{\mathrm{T}}$$

and

$$\hat{V}(U(\beta)) = \sum_j U_j(\beta) U_j(\beta)^{\mathrm{T}}$$

For the two-sample problem we let $Z_{j1} = 1$ for all j and $Z_{j2} = 1$ if the subject is in sample 1 and $Z_{j2} = 0$ if it is in sample 2. A test, based on $\hat{\beta}$ and $\hat{V}(\hat{\beta})$, of the hypothesis that the corresponding β_2 is 0, is a test of the equality at time t of the survival functions in the two samples. While any reasonable link function can be used we shall focus on the logit link. Two choices for the working covariance matrix are reasonable. The usual assumption in GEE is the identity matrix. An alternative is to use the fact that if we have no censoring the variance of the Bernoulli variable $\hat{\theta}_j$ is $S_j(1 - S_j)$. This second choice gives estimates and variances identical to what one would obtain from a logistic regression model when there is no censoring prior to t. Once the pseudo-values have been computed any GEE software, such as the SAS procedure GENMOD, can be used to find estimates of β. A SAS macro to compute the pseudo-values can be found on our website at www.biostat.mcw.edu.

3. COMPARISON OF TESTS

To compare the tests introduced in the previous section an extensive Monte Carlo study was performed. Randomly censored samples were generated from one of the two possible models. For the first model both samples were simulated from exponential populations with parameters chosen so that in the first sample the probability of survival at time 1 was 0.25, 0.50 or 0.75. The parameters in the other sample were chosen so the odds ratio of the survival function at time 1 was 1, 1.5 or 2. For the second set of models the two samples were generated from Weibull distributions with shape parameters 0.5 and 2, respectively. The scale parameters were picked to give the correct probabilities at time 1.

For both the exponential and Weibull distributions we generated one set of simulations with no censoring. For the exponential data we further generated exponential censoring times that were

DOI: 10.1002/sim

either (in half of the cases) the same in both samples or (in the other half of the cases) had a faster rate in one sample. For the Weibull cases a common exponential censoring rate was used in the two samples. However, because the shapes of the survival functions are different for the two groups in the Weibull situation, even in the null case, this led to different censoring patterns in the two samples. The overall censoring fraction in either setup was fixed at 20 or 50 per cent.

The total sample size, $N = n_1 + n_2$ was fixed at 30, 60, 90, 120, 150 or 300. For each value of N we generated 10 000 cases with $n_1 = n_2$ and 10 000 cases with $n_1 = 0.5n_2$. Since our analysis found no effect of different numbers in each arm but rather an overall sample size effect we pooled the results for the runs with equal and unequal n_k's in our final analysis. For the 12 statistics considered, the empirical rates of rejection for tests at the 5 per cent level were computed. All these power estimates are based on the samples for which $\hat{S}_k(1) \neq 0, 1$ for $k = 1, 2$. This problem occurs between 1 and 50 per cent of cases when $n = 30$; between 0 and 20 per cent of cases when $n = 60$; between 0 and 10 per cent of cases when $n = 90$ and in no case when n is greater than 100. For the pseudo-value methods this is less of a problem since the method only breaks down when there are no events (or all events) in both samples which occurs very infrequently. In all cases the inferences given below are based on at least 10 000 cases in each cell when the equal and unequal sample size cells are combined. To summarize the considerable simulation results, concerning both type 1 and type 2 error rates, we applied analysis of variance (ANOVA) techniques. For the type 1 error rate we defined the outcome variable, Y, for the ANOVA as the per cent rejection rate minus the nominal level of 5. In this way, good performance of the test is implied by numerically small estimates for the expectation $E(Y)$ in the ANOVA.

Since some of the tests considered are specifically designed for situations with equal censoring in the two samples, we first focused on this aspect by considering the following model for $E(Y)$:

$$E(Y) = \texttt{TEST} * \texttt{CENS} + \texttt{N} + \%\texttt{CENS} + \texttt{QUANTILE} \tag{13}$$

where the factor `TEST` has 12 levels, `CENS` has two: 'same' or 'different' while `N` has six and `%CENS` and `QUANTILE` have the three levels mentioned above. We fit the model without an intercept and normalize the effects of the last three factors to have a sum of zero since then the estimates for the factor `TEST` $*$ `CENS` have the interpretation as average deviations from the nominal 5 per cent level in the 24 combinations *adjusted for* the effects of the other three factors. (Note that for the factor `%CENS` it is a *weighted* sum of the coefficients which is 0 since this factor is not balanced.) We did not include a factor for distribution in (13) since the effect of this factor is not of primary interest (and since inclusion of that factor did not change the results much). Table II shows the results.

Table II. Average deviations from nominal 5 per cent level of 12 tests (`TEST`) adjusted, using ANOVA, for different sample sizes `N`, per cent censoring `%CENS` and `QUANTILE` at which the test was computed.

	Naive		log		cloglog		arcsin$\sqrt{}$		logit		Pseudo-value	
`TEST`	X_1^2	X_1^{*2}	X_2^2	X_2^{*2}	X_3^2	X_3^{*2}	X_4^2	X_4^{*2}	X_5^2	X_5^{*2}	Id.	Bern.
`CENS`: different	2.24	−2.18	0.15	−2.01	−0.23	−1.76	0.59	−2.07	0.60	−1.36	−1.50	2.24
`CENS`: same	0.79	−1.41	−0.90	−0.85	−0.63	−0.77	0.17	−1.12	−0.93	−0.88	−1.52	0.79

First line: data with same censoring in the two samples, second line: data with different censoring in the two samples. Id., identity working covariance matrix; Bern, Bernoulli working covariance matrix.

Table III. Average deviations from nominal 5 per cent level of six tests (TEST) adjusted, using ANOVA, for QUANTILE at which the test was computed.

TEST	naive	log	cloglog	arcsin$\sqrt{}$	Logit	Pseudo-value
N = 30	2.10	−0.73	−1.33	0.31	−1.67	−1.07
N = 60	1.56	−0.14	−0.26	0.87	−0.73	−1.39
N = 90	1.32	−0.19	−0.03	0.72	−0.37	−1.06
N = 120	1.11	−0.05	−0.02	0.54	−0.11	−0.81
N = 150	0.95	−0.12	0.03	0.55	−0.19	−0.80
N = 300	0.59	−0.02	0.01	0.27	−0.09	−0.59
%CENS = 0	1.26	0.05	−0.12	0.55	−0.46	−0.03
%CENS = 20	0.92	−0.48	−0.47	0.26	−0.73	−0.77
%CENS = 50	1.57	−0.18	−0.23	0.76	−0.44	−1.82
	1.20	−0.28	−0.34	0.47	−0.60	−1.03

Upper panel: deviations given by sample size N (adjusted for per cent censoring %CENS) using model (14); middle panel: deviations given by per cent censoring %CENS (adjusted for sample size N) using model (15); lower panel: marginal effects of TEST from the model (16).

In this table, we see that the versions of the tests which assume equal censoring (X_1^{*2}–X_5^{*2}) do quite poorly when the censoring distributions are different in the two samples. Furthermore, these tests do not do consistently better when censoring patterns are in fact the same in both samples, even though there is an improvement over the unequal censoring case. We see that the same pattern holds for the pseudo-value test based on the identity working covariance matrix. In Table II we also note that the naive test using unequal variances rejects too often when the censoring patterns are different.

In Table III we illustrate how the type 1 error properties of these six tests depend on the factors N and %CENS by fitting models

$$E(Y) = \mathtt{TEST} * \mathtt{N} + \%\mathtt{CENS} + \mathtt{QUANTILE} + \mathtt{CENS} \tag{14}$$

and

$$E(Y) = \mathtt{TEST} * \%\mathtt{CENS} + \mathtt{N} + \mathtt{QUANTILE} + \mathtt{CENS} \tag{15}$$

respectively where TEST now has only six levels. For comparison we also fit the additive model

$$E(Y) = \mathtt{TEST} + \%\mathtt{CENS} + \mathtt{N} + \mathtt{QUANTILE} + \mathtt{CENS} \tag{16}$$

Comparison between the tests shows that all the transformed tests perform considerably better than the untransformed (naive) test. Of the transformed tests the arcsine-square root version tends to have a type 1 error that appears to be slightly elevated. The other tests are slightly conservative. The log and complementary log–log tests both perform quite well. The test based on the pseudo-value approach tends to be a bit conservative. As one would expect the performance is better as the sample size increases and tends to be best when there is no censoring.

Table IV shows the effects of the last four factors in the additive model (16). The effect of these factors on the type 1 error is relatively small except for sample size 30 which tends to have smaller type 1 errors. The table also suggests that the tests have larger type 1 errors when the censoring patterns prior to the time of interest are different.

DOI: 10.1002/sim

Table IV. Effects of N, %CENS, QUANTILE and CENS from the model (16).

N	30	60	90	120	150	300
	−0.38	0.01	0.09	0.13	0.09	0.05
%CENS	0	20	50			
	0.32	−0.18	−0.03			
QUANTILE	0.25	0.50	0.75			
	−0.13	0.12	0.01			
CENS	Same	Different				
	−0.24	0.24				

Table V. Average rejection rates for 12 tests (TEST) adjusted, using ANOVA, for different sample sizes N, per cent censoring %CENS and QUANTILE at which the test was computed.

	Naive		log		cloglog		arcsin$\sqrt{}$		Logit		Pseudo-value	
TEST	X_1^2	X_1^{*2}	X_2^2	X_2^{*2}	X_3^2	X_3^{*2}	X_4^2	X_4^{*2}	X_5^2	X_5^{*2}	Id.	Bern.
CENS: different	69.4	54.2	60.6	54.2	67.9	55.5	67.7	54.6	65.0	62.0	55.8	55.8
CENS: same	62.2	53.7	57.1	53.0	60.0	55.2	61.2	54.0	59.1	59.2	53.6	51.4

First line: data with same censoring in the two samples, second line: data with different censoring in the two samples. Id., identity working covariance matrix; Bern., Bernoulli working covariance matrix.

A similar approach was taken when studying the power of the tests when the odds ratio at time 1 was OR = 1.5 or 2. That is, we performed ANOVA of the response variable, Y, now defined as the per cent rejection rate and fit models like (13)–(16). Here, a good performance of a test is indicated by large estimates.

For the model (13) results when OR = 2 are presented in Table V. The rejection rates for OR = 1.5 were, on average, 18.8 lower. The estimates correspond to N = 300; for values of N = 30, 60, 90, 120, 150, respectively, the averages 43.7, 37.1, 31.1, 25.5, 20.3, respectively, should be subtracted.

From Table V it is seen that tests designed to be optimal for equal censoring in the two groups do not outperform the more general tests when there is the same censoring in both groups and for that reason we will keep the same six tests (including that based on pseudo-values using the Bernoulli covariance) in further comparisons. The results from fitting models (14), (15) and the additive model for these six tests (16) are shown in Table VI.

From Table VI we see that, for all tests considered, the power increases with the sample size and decreases with the amount of censoring as one would expect. For none of the tests is the power impressive for the quoted average values of the other factors, the effect of which in the additive model (16) are shown in Table VII. Comparing the tests we see the highest power for the naive test and that based on the arcsin square-root transformation (both of which, however, as shown above, are anti-conservative) but also tests based on the cloglog and logit transformations have comparably good power values. The (conservative) tests based on pseudo-values have the lowest power.

In summary, considering both type 1 and type 2 error rates the test with the best performance seems to be the one based on the cloglog transformation.

DOI: 10.1002/sim

Table VI. Average rejection rates for six tests (TEST) adjusted, using ANOVA, for QUANTILE at which the test was computed.

TEST	Naive	log	cloglog	arcsin$\sqrt{}$	Logit	Pseudo-value
N = 30	23.8	15.0	20.2	21.0	18.0	16.7
N = 60	31.1	22.6	29.4	29.8	26.5	21.6
N = 90	37.9	30.1	36.3	36.8	34.0	26.7
N = 120	43.8	37.1	42.2	42.7	40.8	31.7
N = 150	49.3	43.1	48.0	48.3	46.2	36.2
N = 300	68.4	65.0	67.4	67.6	66.5	55.2
%CENS = 0	76.1	68.6	74.8	74.8	72.8	63.0
%CENS = 20	70.0	64.2	68.2	68.9	66.6	60.1
%CENS = 50	63.6	56.0	61.3	62.0	59.2	52.8
	69.1	62.2	67.3	67.8	65.4	58.1

Upper panel: deviations given by sample size N (adjusted for per cent censoring %CENS) using model (14); middle panel: deviations given by per cent censoring %CENS (adjusted for sample size N) using model (15); lower panel: marginal effects of TEST from the model (16).

Table VII. Effects of N, %CENS, QUANTILE and CENS from the model (16).

	30	60	90	120	150	300
N	−45.9	−38.2	−31.4	−25.3	−19.8	0 (ref.)
%CENS	0	20	50			
	10.6	0.0	−7.1			
QUANTILE	0.25	0.50	0.75			
	−8.5	4.7	3.8			
CENS	Same	Different				
	3.9	−3.9				
OR	1.5	2				
	−19.8	0 (ref.)				

4. INCORPORATING EXPLANATORY VARIABLES

In many cases it is of interest to include explanatory variables, e.g. to use a stratified version of a test of equality of the survival curves at a fixed time. Multicentre clinical trials are often stratified on the centre or on a number of patient characteristics.

For each of the tests discussed in the previous two sections one can construct a stratified test. Stratified tests based on the Kaplan–Meier or Nelson–Aalen estimator are constructed as follows. Let $\phi(\cdot)$ be one of the transformations discussed in the previous section and let $\hat{S}_{ks}(t)$, $s=1,\ldots,m$, $k=1,2$, be the Kaplan–Meier estimate in the sth stratum for sample k. The test statistics are defined by

$$X^2_{\text{STRAT}} = \frac{\left(\sum_{s=1}^{m} \phi(\hat{S}_{1s}(t)) - \phi(\hat{S}_{2s}(t))\right)^2}{\sum_{s=1}^{m} \hat{V}(\phi(\hat{S}_{1s}(t))) + \hat{V}(\phi(\hat{S}_{2s}(t)))} \tag{17}$$

Tests based on the Nelson–Aalen estimators are obtained by replacing the Kaplan–Meier estimator with the Nelson–Aalen statistic in (17). Also tests [following (6)–(10)] using a pooled variance estimator within each stratum and assuming identical censoring patterns may be studied.

Stratified tests based on the pseudo-value approach are quite easy to conduct. Here, we compute pseudo-values based on the pooled Kaplan–Meier estimator ignoring the subject's treatment group and stratum. Let Z_{sj} be the treatment indicator for the jth subject in the sth stratum. We then fit a generalized linear model with $g(\theta_{sj}) = \alpha_s + \beta Z_{sj}$, where θ_{sj} is the pseudo-value for the jth subject in the sth stratum, using (11). Note that the pseudo-values are the same whether we stratify or not, but the model is different depending on if we wish to stratify or not.

4.1. Estimation

The tests in Section 2 may easily be supplemented with an estimate of the treatment effect at time t. Most obviously, this can be for the test based on the cloglog transformation where a log hazard ratio may be estimated as

$$\hat{\beta} = \log(-\log(\hat{S}_2(t))) - \log(-\log(\hat{S}_1(t))) \tag{18}$$

For the other transformations similar estimators can be set up for the treatment effect in the scale corresponding to that transformation, e.g. a log odds ratio for the logit transformation corresponding to (5).

If not only a treatment effect is of interest but also effects of other explanatory variables, a regression model for the t-year survival probability may be obtained from any standard hazard regression model like the Cox proportional hazards model or an additive hazard model. However, such a model would depend on assumptions on the hazard function for the entire time interval from 0 to t, e.g. on a proportional hazards assumptions for the Cox model and such assumptions seem unnecessary if interest is focused on only the value of the survival probability at t. On the other hand, for data without censoring the indicator of survival beyond time t would be observable for all individuals under study and a standard (e.g. logistic or cloglog) regression model for the expectation of this indicator could be used.

As mentioned above, the pseudo-value approach has the advantage that it is applicable not only for hypothesis testing but also for obtaining estimates in more detailed regression analyses. To study the behaviour of pseudo-value regression models for the t-year survival probability a small Monte Carlo simulation study was conducted.

First, we simulated lifetime data with *proportional hazards*, that is satisfying the Cox regression model: $\alpha_j(t) = \alpha_0(t)\exp(\beta Z_j)$ for individual j, $j = 1, \ldots, N = 250$ with constant baseline hazard $\alpha_0(t) = 1$ and one binary covariate $Z_j = 0$ for half of the sample, $Z_j = 1$ for the other half. Furthermore, $\beta = 1$ or 0, and t was chosen as either the $p = 50$ or the 75 percentile in the distribution corresponding to the baseline hazard. Exponential censoring, $\approx$ 15 or 30 per cent at t, was imposed and 500 replications studied. Table VIII, upper part, shows the results comparing both with results from a standard Cox model using all the event times and with the 'direct' estimator (18). Obviously, both the pseudo-value regression model and the 'direct' estimator are less precise than fitting the true model based on all data but the estimates have little or no bias.

Next, for comparison we simulated lifetime data with *non-proportional* hazards. We used a Weibull model with different shape parameters: 0.5 and 2 in groups given by a binary covariate which again was 0 for half of the sample and 1 for the other half. The sample sizes were $N = 50$, 100, 250, 500. The difference in t-year survival probability between the two groups corresponded to $\beta = 1$ on the cloglog scale and t was the $p = 50$ or 75 percentile in the baseline distribution. Again, exponential censoring was imposed, $\approx$ 15, 30 per cent at t and 500 replications were generated. The results are shown in Table VIII, lower part and compared with results obtained

DOI: 10.1002/sim

Table VIII. Simulation results from analysis of Cox regression models, pseudo-values and 'direct' estimates (18).

				Cox model		Pseudo-values		Direct approach (18)	
N	p	β	Cens.	$\hat{\beta}$	ESD	$\hat{\beta}$	ESD	$\hat{\beta}$	ESD
250	75	0	15	−0.00274	0.142	−0.0053	0.166	−0.0053	0.166
250	75	0	30	−0.00534	0.140	−0.0081	0.166	−0.0081	0.166
250	50	0	15	0.00398	0.138	0.0082	0.189	0.0083	0.189
250	50	0	30	−0.00096	0.152	−0.0147	0.201	−0.0148	0.201
250	75	1	15	1.0003	0.146	1.0227	0.222	1.0190	0.195
250	75	1	30	1.0004	0.143	1.0150	0.224	1.0029	0.190
250	50	1	15	1.0042	0.155	1.0045	0.176	1.0045	0.177
250	50	1	30	1.0054	0.150	1.0130	0.181	1.0118	0.181
50	50	1	15	0.8194	0.382	1.0522	0.494	1.0252	0.406
50	75	1	15	0.1604	0.341	0.9089	0.366	0.7253	0.293
50	50	1	30	0.7533	0.431	1.0480	0.471	1.0474	0.465
50	75	1	30	0.1124	0.333	0.9000	0.381	0.6982	0.309
100	50	1	15	0.8552	0.265	1.0396	0.292	1.0382	0.289
100	75	1	15	0.1877	0.240	0.9839	0.297	0.9071	0.225
100	50	1	30	0.7265	0.291	1.0151	0.328	1.0105	0.314
100	75	1	30	0.1524	0.238	0.9750	0.301	0.9078	0.231
250	50	1	15	0.8557	0.164	1.0070	0.183	1.0066	0.183
250	75	1	15	0.2079	0.150	1.0122	0.209	1.0149	0.190
250	50	1	30	0.7503	0.175	1.0171	0.203	1.0106	0.202
250	75	1	30	0.1351	0.151	1.0001	0.230	0.9983	0.203
500	50	1	15	0.8606	0.118	1.0020	0.129	1.0019	0.129
500	75	1	15	0.2139	0.108	1.0339	0.190	1.0244	0.152
500	50	1	30	0.7504	0.122	1.0087	0.132	1.0085	0.131
500	75	1	30	0.1488	0.103	1.0291	0.158	1.0336	0.158

Upper panel: proportional hazards; lower panel: non-proportional hazards. ESD, estimated standard deviation.

by naively fitting the Cox model, which in this case is a wrong model, and with results for (18). Consequently, the results from the Cox model are quite unstable while the pseudo-value regression estimates are close to the true value of $\beta = 1$. Comparing the estimates based on pseudo-values and the direct estimates, the latter are more biased for small N when $p = 0.75$.

5. EXAMPLE

We now return to the comparison of the disease-free survival probabilities at 6 and 12 months between leukaemia patients given an allo- or auto-transplant. Ignoring the year of transplant the survival probabilities at 6 and 12 months, respectively, are: 0.721 and 0.592 in the auto group and 0.587 and 0.550 in the allo group. If one adjusts for the year of transplant then these probabilities are as given in Table IX.

Table X gives the chi-squared statistics and the p-values for the six tests studied above either ignoring the year of transplant or using a test stratified on the year of transplant. The tests seem to be in relatively close agreement. Using the pseudo-value approach at 6 months we find that the estimated odds ratio of disease-free survival for an auto patient relative to an allo patient

Table IX. Six- and 12-month survival probabilities by donor type and year of transplant.

		1990	1991	1992	1993	1994	1995	1996	1997	1998
Auto	6 months	0.250	0.546	0.471	0.529	0.563	0.625	0.583	0.709	0.622
Allo	6 months	0.615	0.667	0.836	0.730	0.714	0.770	0.647	0.637	0.779
Auto	12 months	0.250	0.455	0.471	0.462	0.500	0.526	0.583	0.709	0.579
Allo	12 months	0.385	0.542	0.673	0.627	0.571	0.621	0.558	0.514	0.652

Table X. Test results (chi-squared test statistics and p-values of equality of the two survival functions) using six different test statistics.

	At 6 months		At 12 months	
Method	Not stratified	Stratified	Not stratified	Stratified
Naive based on S	10.3 (0.0013)	12.6 (0.0004)	0.9 (0.336)	2.0 (0.161)
log transformed	9.0 (0.0022)	8.5 (0.0036)	0.9 (0.346)	1.7 (0.198)
cloglog transform	11.6 (0.0006)	14.1 (0.0002)	1.0 (0.329)	1.9 (0.168)
Arcsine-square root transform	10.7 (0.0011)	12.8 (0.0003)	0.9 (0.335)	1.9 (0.168)
Logit transform (S)	11.1 (0.0009)	18.9 (0.0001)	0.9 (0.334)	1.1 (0.272)
Pseudo-value method	11.1 (0.0009)	9.4 (0.0021)	0.9 (0.327)	0.6 (0.434)

ignoring stratification is 1.82 (95 per cent confidence interval 1.28–2.59) using the independence working covariance matrix, 1.82 (1.28–2.59) using the Bernoulli variance, and 1.77 (1.23–2.54) using the stratified version with an independence working covariance matrix. At 12 months the corresponding odds ratios are 1.19 (0.84–1.67), 1.19 (0.84–1.67) and 1.15 (0.81–1.64), respectively. Note the close agreement between the estimates using either the independent working matrix or the binomial working matrix. The test with the independence working matrix is obtained directly from `SAS PROC GENMOD` using a logit link and a normal error. A `REPEATED` statement forces the estimation of the correct standard errors.

6. DISCUSSION

We have been studying a number of methods for comparing two survival curves at a single point, t in time. The standard approach for making such a comparison, naively comparing the values of $\hat{S}_k(t)$ for $k=1,2$ using the Greenwood estimate of the variances turned out to be unsatisfactory in terms of type 1 error rates. Substantial improvement of the properties of the test was obtained using proper transformations, $\phi(\cdot)$, of the survival functions and the single transformation which seemed to outperform the others was $\phi=\text{cloglog}$. Another approach was tests based on a regression model for pseudo-observations.

Extensions to the $K>2$ sample case may also be of interest. For tests based on the Nelson–Aalen estimators for the cumulative hazards such extensions are special cases of the non-parametric tests studied by Andersen *et al.* [6, Section V.2]. Tests using the Kaplan–Meier estimators may be extended in a similar way. For example, a test based on the transformation $\phi()$ can be constructed using a quadratic form such as

$$X^2 = A\mathbf{\Sigma}^{-1}A^{\mathrm{T}}$$

where A is the vector $(\phi(\hat{S}_1(t))-\phi(\hat{S}_2(t)), \ldots, \phi(\hat{S}_1(t))-\phi(\hat{S}_K(t)))$ and $\mathbf{\Sigma}$ is the $(K-1)\times(K-1)$ matrix with diagonal elements $\hat{V}[\phi(\hat{S}_1(t))]+\hat{V}[\phi(\hat{S}_k(t))]$, $k=2,\ldots,K$, and off-diagonal elements $\hat{V}[\phi(\hat{S}_1(t))]$. Versions of these tests can be constructed using either the unpooled or pooled Kaplan–Meier curves to estimate the variance. The pseudo-value technique generalizes immediately to the $K>2$ sample situation using $K-1$ indicator variables.

While all methods, as seen in Section 4, can be extended to stratified analysis, only the pseudo-value approach accommodates the inclusion of several explanatory variables in the analysis. For this reason, the pseudo-value method is far the most general of those considered. However, for the simple and common two-sample comparison, tests based on a cloglog transformation of the Kaplan–Meier estimators is a simple and reliable method.

ACKNOWLEDGEMENTS

This research was partially supported by a grant (R01 CA54706-10) from the National Cancer Institute. It is, further, based on the joint work of an international research group studying Statistical Analysis of Complex Event History Data at the Centre of Advanced Study at the Norwegian Academy of Sciences and Letters in Oslo, Norway during the academic year 2005/06. We are grateful to anonymous referees for comments to an earlier version of the manuscript which greatly improved the presentation.

REFERENCES

1. Fleming TR, O'Fallon JR, O'Sullivan M, Harrington DP. Modified Kolmogorov–Smirnov test procedures with application to arbitrarily right censored data. *Biometrics* 1980; **36**:607–626.
2. Schumacher M. Two-sample tests of Cramér–von Mises and Kolmogorov–Smirnov type for randomly censored data. *International Statistical Review* 1984; **52**:263–281.
3. Brookmeyer R, Crowley JJ. A *k*-sample median test for censored data. *Journal of the American Statistical Association* 1982; **77**:433–440.
4. Van Besien K, Loberiza F, Bajorunatie R, Armitage J, Bashev A, Burns L, Freytes C, Gibson J, Horowitz M, Inwards D, Martino R, Molina A, Pavlovsky S, Pecora A, Schouten H, Shea T, Lazarus H, Rizo J, Voss J. Comparison of autologous and allogeneic hematopoietic stem cell transplantation for follicular lymphoma. *Blood* 2003; **102**:3521–3529.
5. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
6. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer: New York, NY, 1993.
7. Klein JP, Moeschberger ML. *Survival Analysis*: *Techniques for Censored and Truncated Data* (2nd edn). Springer: New York, NY, 2003.
8. Borgan Ø, Liestøl K. A note on confidence intervals and bands for the survival curve based on transformations. *Scandinavian Journal of Statistics* 1990; **17**:35–41.
9. Bie O, Borgan Ø, Liestøl K. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics* 1987; **14**:221–233.
10. Andersen PK, Klein JP, Rosthøj S. Generalized linear models for correlated pseudo-observations with applications to multi-state models. *Biometrika* 2003; **90**:15–27.
11. Sposto R, Stablein D, Carter-Campbell S. A partially grouped log rank test. *Statistics in Medicine* 1997; **16**:695–704.
12. Klein JP, Andersen PK. Regression modelling of competing risks data based on pseudo-values of the cumulative incidence function. *Biometrics* 2005; **61**:223–229.
13. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **78**:13–22.

DOI: 10.1002/sim