Published in final edited form as:
Biometrics. 2008 September ; 64(3): 733–740. doi:10.1111/j.1541-0420.2007.00975.x.

Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation

Brent R. Logan*, **John P. Klein**, and **Mei-Jie Zhang**
Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226, U.S.A.

Summary

In some clinical studies comparing treatments in terms of their survival curves, researchers may anticipate that the survival curves will cross at some point, leading to interest in a long-term survival comparison. However, simple comparison of the survival curves at a fixed point may be inefficient, and use of a weighted log-rank test may be overly sensitive to early differences in survival. We formulate the problem as one of testing for differences in survival curves after a prespecified time point, and propose a variety of techniques for testing this hypothesis. We study these methods using simulation and illustrate them on a study comparing survival for autologous and allogeneic bone marrow transplants.

Keywords

Censored data; Crossing hazard functions; Generalized linear models; Log-rank test; Pseudo-value approach; Weibull distribution; Weighted Kaplan–Meier statistic

1. Introduction

When comparing treatments in terms of their time-to-event distribution, there may be reason to believe that the survival curves will cross, and standard comparison techniques in such cases could lead to misleading results. Often researchers in such cases will focus on which treatment has a better long-term survival probability. In particular, this research is motivated by a common scenario in hematopoietic stem cell transplantation, illustrated using a study comparing autologous and allogeneic bone marrow transplants for follicular lymphoma (Van Besien et al., 2003). The sample contained 175 patients with an HLA-identical sibling allogeneic transplant and 596 patients with an unpurged autologous transplant. We are interested in comparing the disease-free survival (DFS) curves (i.e., the probability a patient is alive and disease free) between the two treatment arms. However, this comparison is complicated by the likely possibility that the hazard functions from these two treatments will cross at some point. Allogeneic transplants tend to have a higher mortality early due to the toxicity of the higher doses of chemotherapy used to ablate the immune system as well as graft-versus-host disease from the donor cells. However, the donor cells may provide a graft-versus-lymphoma effect resulting in less relapse of the primary disease in long-term survivors. In contrast, autologous transplants have lower early toxicity because patients do not experience

*email: blogan@mcw.edu.

6. Supplementary Materials
Web Appendix A, referenced in Section 2.3, is available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

graft-versus-host disease. However, these patients do not benefit from the protection against relapse from the graft-versus-lymphoma effect, so they tend to experience more relapses. These contrasting profiles are illustrated by the Kaplan–Meier curves for this dataset in Figure 1. The DFS of the allogeneic transplant arm drops quickly early, but then levels off, whereas the DFS of the autologous transplant arm decreases more slowly but does not plateau. The two curves appear to cross at about 1 year. The unweighted log-rank test will have poor power to detect such a difference in survival curves.

Generally, the question of interest is, which if any of the treatments yields better long-term survival? Several strategies for addressing this question are possible. One may pick a single long-term time point, and compare the survival estimates between the treatment groups at this single time point, as is discussed in Klein et al. (2007). However, there are some potential problems with this. First the results may be sensitive to the time point chosen. Second, this strategy ignores events occurring after the selected time point. For example, in a clinical trial comparing two treatments one might select 3-year survival as the primary endpoint. However, if everyone is followed up for 3 years and accrual occurs over a period of time such as 2 years, then there is substantial information on later events (between 3 and 5 years) for patients enrolled early in the trial. Therefore, selecting a single time point may be inefficient.

Another alternative would be to estimate simultaneous confidence bounds for the difference in survival curves (Parzen, Wei, and Ying, 1997; Zhang and Klein, 2001), which identify time regions where the two treatments are different. However, because of the large number of time points being considered and adjusted for, these tend to be quite wide and may be inefficient in determining late differences between treatments.

Another option would be the weighted log-rank test, with more weight placed on later time points to reflect interest in late events. For example, Fleming and Harrington (1981) proposed a class of weighted log-rank tests with a weight function equal to $\hat{S}(t)^{\rho}(1-\hat{S}(t))^{\gamma}$. Here setting $\rho = 0$ and $\gamma = 1$ would place more weight on late events and hence late differences in the hazard rates and/or the survival curves. However, even though the weight is placed appropriately, this test is still designed to test the null hypothesis that the entire survival curves are equal. If we are focused instead on late differences rather than the entire survival curve, even the weighted log-rank test may be overly sensitive to early differences in the survival curves. We will illustrate this point in simulations presented later in the article.

We propose a specific formulation of the hypothesis to focus on late differences in the survival curve. We assume that atime point t_0 can be prespecified, so that survival curves are presumed likely to cross prior to that time point if at all. The null hypothesis is $\mathrm{H}_0 : S_1(t) = S_0(t)$, for all $t \geq t_0$, where $S_1(t)$ and $S_0(t)$ denote the survival curves at time t for the treatment and control groups, respectively, versus the alternative, $H_1 : S_1(t) \neq S_0(t)$, for some $t \geq t_0$. This formulation allows us to specify exactly over what time range the comparison of treatments is of interest, for example, after t_0.

Note that this null hypothesis is equivalent to $\mathrm{H}_0 : \{S_1(t_0) = S_0(t_0)\} \cap \{\lambda_1(t) = \lambda_0(t), t > t_0\}$, where $\lambda_k(t)$ represents the hazard function at time t for group k, $k = 0, 1$. This formulation allows us to separate the hypotheses into two sub-hypotheses: the hypothesis of equality of survival at t_0 and the hypothesis of no difference in the hazard function after t_0. The composite hypothesis can then be tested using combinations of test statistics for each of the sub-hypotheses.

In the next section, we describe possible methods for testing this null hypothesis.

2. Methods

2.1 Notation

The data consist of $n_1 + n_0 = n$ subjects with event times t_j. Let the distinct event times be ordered such that $t_1 < \cdots < t_m$. At time t_j let d_{kj} denote the number of events and Y_{kj} denote the number at risk in the kth group, $k = 0, 1$.

The Kaplan–Meier estimate of survival in group k is given by

$$\widehat{S}_k(t) = \prod_{t_j < t} \left(1 - \frac{d_{kj}}{Y_{kj}}\right).$$

The variance of the Kaplan–Meier estimate is estimated by Greenwood's formula given by

$$\widehat{\text{var}}\left\{\widehat{S}_k(t)\right\} = \widehat{S}_k(t)^2 \widehat{\sigma}_k(t)^2,$$

where

$$\widehat{\sigma}_k(t)^2 = \sum_{t_j \le t} \frac{d_{kj}}{Y_{kj}\left(Y_{kj} - d_{kj}\right)}.$$

The Nelson–Aalen estimate of the cumulative hazard function is

$$\widehat{\Lambda}_k(t) = \sum_{t_j \le t} \frac{d_{kj}}{Y_{kj}},$$

with variance estimated by

$$\widehat{\text{var}}\left\{\widehat{\Lambda}_k(t)\right\} = \sum_{t_j \le t} \frac{d_{kj}}{Y_{kj}^2}.$$

2.2 Comparisons Based on a Single Time Point

The simplest method for testing the null hypothesis that the survival curves after time t_0 are equal would be to compare the survival curves at a selected point $t' > t_0$, using the difference in Kaplan–Meier estimates of survival at t'. One can also construct a test statistic based on transformations of the survival probabilities at a fixed point in time, as described in Klein et al. (2007). Their recommendations were that the complementary log–log transformation of the survival probability works the best overall, resulting in the test statistic

$$Z_{\text{CLL}}(t') = \frac{\log\left[-\log\left\{\widehat{S}_1(t')\right\}\right] - \log\left[-\log\left\{\widehat{S}_0(t')\right\}\right]}{\sqrt{\widehat{\sigma}_1^2(t') / \left[\log\left\{\widehat{S}_1(t')\right\}\right]^2 + \widehat{\sigma}_0^2(t') / \left[\log\left\{\widehat{S}_0(t')\right\}\right]^2}}. \tag{1}$$

Alternatively, one could compare the cumulative hazard functions at a selected time point $t' > t_0$, using the Nelson–Aalen estimates at t', $\hat{\Lambda}_k(t')$. Tests based on the cumulative hazard function should behave similarly to those using a log transformation of the survival function.

2.3 Weighted Kaplan–Meier Test

One way to compare the entire survival curve after t_0 is to consider a modification of the weighted Kaplan–Meier statistic (Pepe and Fleming, 1989, 1991), where the integral is taken over the restricted range after t_0. The statistic is given by

$$W_{\mathrm{WKM}}(t_0)=\int_{t_0}^{t_m}\widehat{w}(t)\left\{\widehat{S}_1(t)-\widehat{S}_0(t)\right\}dt,$$

where $\hat{w}(t) = \{n_1\hat{G}_1(t) + n_0\hat{G}_0(t)\}^{-1}n\hat{G}_1(t)\hat{G}_0(t)$ and $\hat{G}_k(t)$ is the Kaplan–Meier estimate of the censoring distribution. Let ℓ denote the index of the event time such that $t_{\ell-1} \le t_0 < t_\ell$. The (unpooled) variance of this statistic can be estimated by

$$\widehat{\mathrm{var}}_{\mathrm{WKM}}(t_0)=\sum_{k=0}^{1}\left\{A_{k0}^2\sum_{j=1}^{\ell-1}\frac{d_{kj}}{Y_{kj}^2}+\sum_{j=\ell}^{m-1}A_{kj}^2\frac{d_{kj}}{Y_{kj}^2}\right\},$$

where $A_{k0}=\int_{t_0}^{t_m}\widehat{w}(t)\widehat{S}_k(t)\,dt$ and $A_{kj}=\int_{t_j}^{t_m}\widehat{w}(t)\widehat{S}_k(t)\,dt$. A sketch of the derivation of this variance expression is given in Web Appendix A. Then the standardized weighted K–M statistic follows a standard normal distribution under the null hypothesis and is given by

$$Z_{\mathrm{WKM}}(t_0)=\frac{W_{\mathrm{WKM}}(t_0)}{\sqrt{\widehat{\mathrm{var}}_{\mathrm{WKM}}(t_0)}}. \tag{2}$$

2.4 Tests Based on Pseudo-Value Observations

Another test is based on a pseudo-value regression technique proposed by Andersen, Klein, and Rosthoj (2003) and Klein and Andersen (2005). Originally applied in the context of regression models for multistate models and competing risks data, it can also be used in the simple survival comparison context. For a given time point τ_j, compute the pooled sample Kaplan–Meier estimator, $\hat{S}_p(\tau_j)$, based on all $n_1 + n_0$ observations and the Kaplan–Meier estimator based on the sample of size $n_1 + n_0 - 1$ with the ith observation removed, $\widehat{S}_p^{(i)}\left(\tau_j\right)$, for $i = 1, \ldots, n$. Define the ith pseudo-value at time τ_j by $\widehat{\theta}_{ij}=(n_1+n_0)\widehat{S}_p\left(\tau_j\right)-(n_1+n_0-1)\widehat{S}_p^{(i)}\left(\tau_j\right)$, for $i = 1, \ldots, n$.

To perform inference on survival curves after a fixed time t_0, we use the pseudo-values defined for event times $t > t_0$. Let τ_1 correspond to the earliest event time occurring after t_0, τ_2 correspond to the next earliest event time after t_0, and so forth, so that there are a total of m' such observed event times in the dataset. We consider a generalized linear model for the pseudo-values, given by $g(\theta_{ij}) = \alpha_j + \beta Z_i$, for $i = 1, \ldots, n$; $j = 1, \ldots, m'$, where Z_i is an indicator with value 1 if the patient is in the treatment group and 0 if they are in the control group. Then given that we are only considering pseudo-values for times $t > t_0$, the null hypothesis H_0 of equal survival curves after t_0 is equivalent to testing $H_0' : \beta=0$.

Inference on β may be performed using generalized estimating equations (GEE; Liang and Zeger, 1986). Let $\mu(\cdot) = g^{-1}(\cdot)$ be the mean function. Define $d\mu_i(\beta, \alpha)$ to be the vector of partial derivatives of $\mu(\cdot)$ with respect to (β, α), where α is an m'-dimensional vector of intercepts at time τ_j, $j = 1, \ldots, m'$. Let V_i be a working covariance matrix. Express the pseudo-values and their expectations in vector notation as $\hat{\theta}_i = (\hat{\theta}_{11}, \ldots, \hat{\theta}_{1m'})$ and $\theta_i = (\theta_{11}, \ldots \theta_{1m'})$. Then the estimating equations to be solved are of the form

$$U(\beta,\alpha)=\sum_i d\mu_i(\beta,\alpha)' V_i^{-1}\left(\widehat{\theta_i}-\theta_i\right)=\sum_i U_i(\beta,\alpha)=0.$$

Let $(\hat{\beta}, \hat{\alpha})$ be the solution to this equation. Using results of Liang and Zeger (1986), under standard regulatory conditions, it follows that $\sqrt{n}\left\{\left(\widehat{\beta},\widehat{\alpha}\right)-(\beta,\alpha)\right\}$ is asymptotically multivariate normal with mean 0. The covariance matrix of $(\hat{\beta}, \hat{\alpha})$ can be estimated by the sandwich estimator $\hat{\Sigma}(\hat{\beta}, \hat{\alpha})$ where

$$\widehat{\Sigma}(\beta,\alpha)=\{I(\beta,\alpha)\}^{-1}\left\{\sum_i U_i(\beta,\alpha)\,U_i(\beta,\alpha)'\right\}\{I(\beta,\alpha)\}^{-1},$$

and

$$I(\beta,\alpha)=\sum_i d\mu_i(\beta,\alpha)\,V_i^{-1}d\mu_i(\beta,\alpha)'$$

is the model-based equivalent of the information matrix (Andersen et al., 2003).

When the number of time points or pseudo-values being included for each patient is large, this can present numerical difficulties in several aspects. Estimation of the parameters can be slow because there are a large number of parameters, and numerical algorithms must be used. Furthermore, calculation of the matrix $\hat{\Sigma}$ requires the difficult inversion of a high-dimensional matrix $I(\hat{\beta}, \hat{\alpha})$. One option is to consider alimited number of points (say 5 or 10) spread out equally on an event scale over the time period after t_0. An alternative is to use the generalized score statistic for β (Rotnitzky and Jewell, 1990; Boos, 1992), as considered in Lu (2006) for the pseudo-value regression context. The generalized score statistic for β when there is asingle dichotomous predictor can be shown to have a closed form, assuming an independent working correlation matrix and using the complementary log-log link function. Let $\tilde{\alpha_j} = \log\{-\log(\bar{\theta}_{\cdot j})\}$ be the solution for α_j in the estimating equation under the null hypothesis, where $\bar{\theta}_{1j}=n_1^{-1}\Sigma_{i:Z_i=1}\widehat{\theta}_{ij}, \bar{\theta}_{\cdot j}=n^{-1}\Sigma_i\widehat{\theta}_{ij}$, and $q_j = \bar{\theta}_{\cdot j}\log\bar{\theta}_{\cdot j}$. The generalized score statistic for testing $H_0:\beta = 0$ simplifies to

$$\begin{aligned}\chi^2_{\mathrm{PSV}}(t_0) &=\left\{\widehat{\Sigma}\left(0,\tilde{\alpha}\right)_{11}\right\}^{-1}\left[\left\{I^{-1}\left(0,\tilde{\alpha}\right)\right\}_{11}U\left(0,\tilde{\alpha}\right)_1\right]^2\\ &=\frac{n^2\left\{\sum_j n_1 q_j\left(\bar{\theta}_{1j}-\bar{\theta}_{\cdot j}\right)\right\}^2}{\sum_{j,j'}q_j q_{j'}\left[\sum_i\{n_0^2 Z_i+n_1^2(1-Z_i)\}\left(\widehat{\theta}_{ij}-\bar{\theta}_{\cdot j}\right)\left(\widehat{\theta}_{ij'}-\bar{\theta}_{\cdot j'}\right)\right]},\end{aligned} \tag{3}$$

where the matrix element $(\cdot)_{11}$ or vector element $(\cdot)_1$ refers to the β component. This statistic asymptotically follows a χ^2_1 distribution under the null hypothesis that $\beta = 0$. Note, however, that the method can be biased when the censoring distribution depends on covariates.

The pseudo-value regression technique offers several potential improvements over the other methods studied in this article. First, it allows for straightforward inclusion of additional covariates in the generalized linear model. Although other methods discussed here also can be extended to include additional covariates, the generalized linear model framework makes this extension very straightforward. Another advantage is that the pseudo-value regression approach allows one to model the effect of treatment as a time-dependent predictor. Even though we are attempting to eliminate the effect of early differences in outcome by comparing the survival curves after t_0, there is still the possibility that the effect of treatment is still not constant or even consistent after time t_0. Using a single parameter to describe the treatment effect may not be sensitive to these kinds of differences in the late survival curves, whereas allowing for a time-dependent effect in the generalized linear model will have better power to capture such a treatment effect. However, both of these extensions make the analysis more complex, because the simplified form of the generalized score test no longer holds. It is likely that these extensions would require one to use alimited number of time points after t_0, rather than all event times as is done here. We do not consider these extensions further in this article.

2.5 Combination Tests

Finally, we consider alternative test statistics based on the formulation of the overall hypothesis as an intersection of two sub-hypotheses, $H_0 = H_{01} \cap H_{02}$, given by $H_{01} : S_1(t_0) = S_0(t_0)$, and $H_{02} : \lambda_1(t) = \lambda_0(t)$, $t > t_0$. Hypothesis H_{01} can be tested using either a standardized difference in Kaplan–Meier estimates or alternatively a standardized difference in Nelson–Aalen estimates of the cumulative hazard function. Let $X_{NA}(t_0) = \hat{\Lambda}_1(t_0) - \hat{\Lambda}_0(t_0)$, and let $\widehat{\sigma}^2_{\text{NA}}(t_0) = \widehat{\text{var}}\left\{\widehat{\Lambda}_1(t_0)\right\} + \widehat{\text{var}}\left\{\widehat{\Lambda}_0(t_0)\right\}$. Then the test statistic for H_{01} is $Z_{NA}(t_0) = X_{NA}/\hat{\sigma}_{NA}$.

Hypothesis H_{02} can be tested using a log-rank test, starting from t_0, given by

$$X_{\text{LR}}(t_0) = \sum_{t_j > t_0} \frac{Y_{1j}Y_{0j}}{Y_j} \left\{ \frac{d_{1j}}{Y_{1j}} - \frac{d_{0j}}{Y_{0j}} \right\}.$$

The log-rank test has variance consistently estimated by

$$\widehat{\sigma}^2_{\text{LR}}(t_0) = \sum_{t_j > t_0} \frac{Y_{1j}Y_{0j}}{Y_j^2} \left(\frac{Y_j - d_j}{Y_j - 1} \right) d_j,$$

where d_j and Y_j are the total number of deaths and the total number at risk, respectively, in the pooled sample. Then the standardized test statistic is given by

$$Z_{\text{LR}}(t_0) = \frac{X_{\text{LR}}(t_0)}{\widehat{\sigma}_{\text{LR}}(t_0)}, \tag{4}$$

which asymptotically follows a standard normal distribution under H_0. One also could consider a weighted log-rank test of H_{02}; however, in simulations it made little difference, probably because we are only testing hazard rates after t_0, so we do not consider it further.

Note that $\sqrt{n}\left(X_{\text{NA}}(t_0), X_{\text{LR}}(t_0)\right)$ asymptotically follows a bivariate normal distribution under H_0 with mean (0, 0). The variance–covariance matrix of $(X_{NA}(t_0), X_{LR}(t_0))$ can be estimated by $\widehat{\Sigma} = \text{Diag}\left\{\widehat{\sigma}^2_{\text{NA}}(t_0), \widehat{\sigma}^2_{\text{LR}}(t_0)\right\}$. The off-diagonal elements of the covariance matrix are 0 because the events contributing to each statistic are occurring over nonoverlapping times; the Nelson–Aalen estimator only considers events occurring up until time t_0, whereas the log-rank test only considers events starting after t_0. An equivalent result is that $(Z_{NA}(t_0), Z_{LR}(t_0))$ follows a bivariate normal distribution with mean (0, 0) and variance–covariance matrix equal to the identity matrix. To test the composite null hypothesis H_0, we can consider a number of ways of combining $Z_{NA}(t_0)$ and $Z_{LR}(t_0)$, including linear combination tests, a quadratic form test, and a Union-Intersection (Roy, 1953) test. However, the Union-Intersection test is omitted for brevity because it did not perform as well as the quadratic test.

2.5.1 Linear combination tests—Linear combination tests consist of a linear combination of the test statistics for each component null hypothesis. These tests have the form

$$Z(t_0)=\frac{aX_{\mathrm{NA}}(t_0)+bX_{\mathrm{LR}}(t_0)}{\sqrt{a^2\widehat{\sigma}^2_{\mathrm{NA}}(t_0)+b^2\widehat{\sigma}^2_{\mathrm{LR}}(t_0)}}.$$

Several sets of weights are possible; we present one which performed well in simulations. If we set $a=\left\{2\widehat{\sigma}^2_{\mathrm{NA}}(t_0)\right\}^{-1/2}$ and $b=\left\{2\widehat{\sigma}^2_{\mathrm{LR}}(t_0)\right\}^{-1/2}$, this yields,

$$Z_{\mathrm{OLS}}(t_0)=\frac{1}{\sqrt{2}}\left\{Z_{\mathrm{NA}}(t_0)+Z_{\mathrm{LR}}(t_0)\right\}. \tag{5}$$

This statistic is analogous to the ordinary least squares (OLS) test of O'Brien (1984) for the multiple endpoint testing problem.

Sposto, Stablein, and Carter-Campbell (1997) proposed a partially grouped log-rank test, which is another special case of the linear combination test. Instead of testing H_{01} with a difference in Nelson–Aalen estimates at t_0, they used the difference in Kaplan–Meier estimates. They estimate the variance of this difference using the pooled samples, yielding a final statistic

$$Z_{\mathrm{SP,P}}(t_0)=\frac{\left[n_1n_0n^{-1}\left\{\widehat{S}_0(t_0)-\widehat{S}_1(t_0)\right\}\right]+X_{\mathrm{LR}}(t_0)}{\sqrt{n_1n_0\widehat{\mathrm{var}}\left\{\widehat{S}_p(t_0)\right\}+\widehat{\sigma}^2_{\mathrm{LR}}(t_0)}}, \tag{6}$$

where $\hat{S}_p(t_0)$ is the Kaplan–Meier estimate at t_0 based on the pooled sample of data. Under H_0 this statistic has a standard normal distribution. One could also consider a version of this test in which the variances of the first term (the survival difference at t_0) are not estimated using the pooled sample; in our simulations, this statistic performed almost identically to the pooled variance test, so we do not consider it further.

2.5.2 Quadratic tests—Next we consider quadratic forms of $(X_{\mathrm{NA}}(t_0), X_{\mathrm{LR}}(t_0))$, which result in a χ^2 test asymptotically. Here

$$\begin{aligned}\chi^2(t_0)\quad &=(X_{\mathrm{NA}}(t_0),X_{\mathrm{LR}}(t_0))'\widehat{\Sigma}^{-1}(X_{\mathrm{NA}}(t_0),X_{\mathrm{LR}}(t_0))\\ &=Z^2_{\mathrm{NA}}(t_0)+Z^2_{\mathrm{LR}}(t_0).\end{aligned} \tag{7}$$

This follows a χ^2_2 distribution under H_0.

3. Design of Simulation Study

A simulation study was designed to compare the performance of the different procedures in terms of their type I error rate and power to detect late differences in survival curves. We set it up based on comparing the survival curves between two equal-sized groups after t_0 = 24 months, and we assume that all patients have a maximum of 72 months of follow-up. In addition to censoring at the fixed time of 72 months, we also overlay an exponential censoring pattern prior to 72 months, with rates selected to induce one of three censoring percentages by 24 months: (1) 0% censoring in each group at 24 months, (2) 15% censoring in each group by 24 months, and (3) 10% censoring in the control group and 20% censoring in the treatment group by 24 months. The overall censoring percentage is between 35 and 60% depending on the scenario, and the censoring time was generated independently of the event time.

For the type I error rate simulations, total sample sizes of n = 100, 200, and 400 with equal sample size per group are studied, across four null hypothesis scenarios. To generate these scenarios we assume piecewise exponential survival curves, where the survival curves differ by 0%, 5%, 10%, or 15% at 8 months (scenarios A–D, respectively), but are equal at 24 months and beyond. These curves are shown in Figure 2a.

For the power simulations, total sample sizes of n = 400 are studied, across five alternative hypothesis scenarios. We generate these scenarios using Weibull survival curves. The curves are shown in Figure 2b–f for scenarios E–I, respectively. Scenario E refers to a proportional hazards situation with no crossing hazards or survival curves. Scenarios F and G refer to situations where the survival curves cross prior to t_0 but to differing degrees. Scenario H is a situation where the survival curves cross exactly at t_0, and in Scenario I the survival curves cross after t_0. The percentage of patients still at risk at time t_0, which impacts the power, varies between 30 and 66% depending on the scenario, and can be roughly obtained for a given scenario by combining the survival probability from Figure 2 at t_0 with the censoring percentage at t_0 (0%, 15%, or 10%, 20%).

For comparison purposes, the log-rank test and the weighted log-rank test with Fleming–Harrington weights of $\rho = 0$, $\gamma = 1$ are also shown, even though they are not specifically suited for testing the null hypothesis of interest here. All simulations used 10,000 Monte Carlo samples.

To summarize the simulation results, we applied analysis of variance (ANOVA) techniques. For the type I error rate, we defined the outcome Y as the percent rejection rate minus the nominal rate of 5%, so that good performance is indicated by values of the expectation of Y near 0. For the power simulations, the outcome Y is defined as the percent rejection rate, so that good performance is defined by high values of the expectation of Y.

There were 11 test statistics considered in the simulations: three pointwise comparisons based on the complementary log–log transformation (1) evaluated at three different time points t' after t_0, the statistics given in equations (2)-(7), and the log-rank and weighted log-rank tests starting at time 0. In addition, we included in the ANOVA model factors for scenario (A–D; 4 levels), total sample size n (100, 200, and 400; 3 levels), and censoring pattern (0%, 15%, or (10%,20%); 3 levels). We also considered models in which each of these factors interacted with the main test statistic effect to examine performance of these statistics for specific levels of the other factors (lower order terms included when interaction is present). Specifically, we fit the following models:

$$E(Y_{tsnc}) = \mu_{ts} + \alpha_n + \beta_c, \tag{8}$$

$$E(Y_{tsnc}) = \mu_{tn} + \beta_c + \gamma_s, \tag{9}$$

$$E(Y_{tsnc}) = \mu_{tc} + \alpha_n + \gamma_s, \tag{10}$$

$$E\left(Y_{tsnc}\right)=\mu_t+\alpha_n+\beta_c+\gamma_s. \tag{11}$$

Here the subscripts *t*, *s*, *n*, and *c* refer to test (11 levels), scenario (4 levels), *n* (3 levels), and censoring pattern (3 levels), respectively, as described above. We fit the model without an intercept and normalized the effects of the other factors to have a sum of zero because then the estimates for the interaction terms have the interpretation as average deviations from the nominal level of 5% adjusted for the effects of the other factors. The results are shown for each test in Table 1 by scenario (model (8)), by total sample size *n* (model (9)), by censoring pattern (model (10)), and overall (model (11)).

We can see that most of the methods control the type I error rate accurately even for a total sample size of $n = 100$, whereas the $Z_{\mathrm{OLS}}(t_0)$ statistic has the most accurate control. Entries in the table where the type I error rate is more than 2 SEs from the nominal level are marked in bold. The weighted Kaplan–Meier statistic, $Z_{\mathrm{WKM}}(t_0)$, has an inflated type I error by about 1% for $n = 100$, although this inflation of the error rate dissipates with larger sample size. The $\chi^2(t_0)$ statistic is somewhat conservative for smaller sample sizes. Also, we point out that the log-rank and weighted log-rank tests starting at time 0 do not control the type I error rate for this hypothesis test for scenarios B–D, because of the early differences in the survival curves prior to $t_0 = 24$. The inflation of the type I error rate worsens with larger early differences in the survival curves (scenario D) and with larger sample sizes, because the test has higher power to detect these early differences. However, the log-rank and weighted log-rank procedures are not designed to test the null hypothesis H_0 of equal survival curves after t_0, so this result is not unexpected.

For the power results, we only considered one sample size (total $n = 400$) and the same censoring patterns as in the type I error simulations. The power of the various procedures is expected to depend heavily on the scenario; therefore to summarize the power results, we fit an ANOVA model with an interaction between the test statistic and the scenario, adjusting for censoring pattern,

$$E\left(Y_{tsc}\right)=\mu_{ts}+\beta_c. \tag{12}$$

Here the subscripts *t*, *s*, and *c* refer to test (11 levels), scenario (5 levels), and censoring pattern (3 levels), respectively.

The results are shown for each test and scenario combination in Table 2. Several general patterns emerge from examining this table. The pointwise comparisons based on the complementary log–log transformation, $Z_{\mathrm{CLL}}(t')$, are sensitive only to differences at that point t' and do not compare the entire curves. The pointwise comparison at 72 months has the highest power among pointwise comparisons because the largest differences in survival curves are seen there. Although the pointwise comparison at 72 months does well in scenarios F and G, it suffers from loss of power compared to some of the other methods for the proportional hazards situation (E) and when the differences at 72 months are not as pronounced (H–I). In these cases, a comparison of the entire curves using one of the other techniques can have more power.

The weighted Kaplan–Meier comparison, $Z_{\mathrm{WKM}}(t_0)$, and the pseudo-value technique, $\chi^2_{\mathrm{PSV}}(t_0)$, both aggregate differences in survival curves across the times after t_0, and as expected they perform well when those differences are in a consistent direction (scenarios E–G) and poorly when those differences are not in the same direction (scenario I) or when many of those early differences are very small (scenario H). Similarly, the linear combination tests

($Z_{\mathrm{OLS}}(t_0)$ and $Z_{\mathrm{SP,P}}(t_0)$) combine the test statistics before and after t_0 in a linear fashion and would be expected to perform well when those statistics have the same sign (E–G) and poorly when they have the opposite sign (I) or one of them has a zero mean (H). Among these, the OLS test $Z_{\mathrm{OLS}}(t_0)$ has the best power in scenarios (F–I) and has only slightly less power for scenario (E), and would be recommended.

The $\chi^2(t_0)$ combination test performs well for scenarios (H–I), but may be less efficient when the statistics have the same sign (E–G). However, it appears that the magnitude of the relative power loss for scenarios (E–G) is modest, so this may be a good general method. Note that the log-rank test starting at t_0, $Z_{\mathrm{LR}}(t_0)$, has high power for scenarios (G–I); this component of the combination tests is driving the power much more than the other component, and this disparity in the component statistics is what produces good power for the $\chi^2(t_0)$ combination test, as opposed to a linear combination test.

It is also worth noting that although the pseudo-value test $\chi^2_{\mathrm{PSV}}(t_0)$ had poor power for scenarios (H–I), it was implemented assuming a time-independent effect of treatment on survival. One advantage of these models is that one can test for whether there is a time-dependent effect and incorporate this into the models, thereby improving the sensitivity to the treatment effect captured in these scenarios. Further investigation of this is needed, however.

The power for the log-rank and weighted log-rank tests starting at time 0 is also shown for comparative purposes. As expected, the log-rank test has the highest power for the proportional hazards alternative (E), but performs poorly for the remaining crossing hazards alternatives (H–I). The weighted log-rank test has the highest power among all the tests for scenarios (F–H), but performs somewhat worse for E and I. Also, the weighted log-rank test can be overly sensitive to early differences in survival curves, and is not suitable for testing the null hypothesis H_0 of equal survival curves after t_0.

Overall, the $\chi^2(t_0)$ test from equation (7) has the highest power in scenarios H and I, and power that is not substantially worse than the other methods for scenarios E–G, and could be recommended as a general omnibus technique against a variety of alternative hypothesis scenarios. However, note that in scenarios E–G there is a survival difference at t_0 that is at least maintained or increased, implying a consistent and easily interpretable difference in long-term survival. On the other hand, in scenario I the survival curves come together and cross sometime after t_0. This makes it difficult to interpret which treatment is better, because it depends on when after t_0 you look and how much follow-up there is in the study. The power advantage of the $\chi^2(t_0)$ test for scenario I may not be worth the power loss for the more relevant and interpretable scenarios E–G. The OLS linear combination test $Z_{\mathrm{OLS}}(t_0)$ from equation (5) has high power for scenarios E–G, and is a better choice than the $\chi^2(t_0)$ test for identifying consistent differences in long-term survival after t_0.

4. Example

We now return to the motivating example comparing autologous and allogeneic bone marrow transplants for follicular lymphoma. We are interested in comparing the DFS curves (i.e., the probability that a patient is alive and disease free) between the two treatment arms, but in particular we are interested in comparing the DFS curves after 1 year, which should eliminate much of the anticipated differences in early mortality between autologous and allogeneic transplants. There is a modest amount (42%) of censoring present in the dataset.

The usual log-rank test gives $p = 0.443$, whereas a weighted log-rank test with Fleming–Harrington weights of $\rho = 0$, $\gamma = 1$ gives a p-value of <0.001. Each of the methods described above was applied using $t_0 = 12$ months. The pointwise comparison using the complementary log–log transformation is shown at $t' = 36$ months. The p-values are given in Table 3.

The table indicates that the pointwise comparison $Z_{\mathrm{CLL}}(36)$ does not find a significant difference in the survival estimates at 36 months. However, if you compare the entire survival curves after 12 months, the weighted Kaplan–Meier comparison, the pseudo-value approach, and the $Z_{\mathrm{SP,P}}(12)$ test all have nonsignificant p-values, whereas the $\chi^2(12)$ test and the linear combination test $Z_{\mathrm{OLS}}(12)$ have significant p-values.

In understanding the discrepancies in the results, we need to look at the shape of the survival curves and where they cross. The survival curves cross right around t_0, whereas the simple log-rank test of the hazard functions after t_0, $Z_{\mathrm{LR}}(12)$, is highly significant. This situation is most similar to simulation scenario H. The weighted Kaplan–Meier statistic, the pseudo-value approach, and the $Z_{\mathrm{SP,P}}(t_0)$ combination test all had low power in this scenario, whereas the $\chi^2(t_0)$ test and the $Z_{\mathrm{OLS}}(t_0)$ test had higher power. The latter two tests are more sensitive to the component test of the hazard functions after t_0. Using these latter tests, one would conclude that there is a significant difference in the survival curves after 12 months.

5. Discussion

We have considered a number of methods for comparing two survival curves after a prespecified time point, t_0. This situation may be of interest when the survival curves are expected to cross, so that we are only interested in late differences. Our simulations indicate that a simple pointwise comparison of the curves at some time after t_0 is sensitive to the time point chosen and may be inefficient in some settings because they ignore differences in the curves at other times after t_0. Use of the weighted log-rank test may be overly sensitive to early differences prior to t_0. Two of the methods studied stand out, both of which are based on a combination of a pointwise comparison of the survival curves at t_0 and a log-rank test after t_0. A $\chi^2(t_0)$ combination test given in equation (7) performs well as an omnibus test against a wide variety of alternative hypothesis scenarios. But if the interest is on identifying consistent differences in long-term survival after t_0, a simple equally weighted linear combination test $Z_{\mathrm{OLS}}(t_0)$ given in equation (5) has better power for this alternative hypothesis and is recommended. A pseudo-value regression approach was also discussed, which can be easily extended to account for covariates. Also, the regression model framework can be used to test whether the treatment effect is consistent across time, thereby allowing flexibility for dealing with different alternative hypotheses such as survival curves crossing after the prespecified time t_0. Further research is needed on adapting this approach.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This research was partially supported by a grant (R01 CA54706-10) from the National Cancer Institute.

References

Andersen PK, Klein JP, Rosthoj S. Generalized linear models for correlated pseudo-observations with applications to multi-state models. Biometrika 2003;90:15–27.

Boos D. On generalized score tests. American Statistician 1992;46:327–333.

Fleming TR, Harrington DP. A class of hypothesis tests for one and two samples of censored survival data. Communications in Statistics 1981;10:763–794.

Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudo-values of the cumulative incidence function. Biometrics 2005;61:223–229. [PubMed: 15737097]

Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. Statistics in Medicine 2007;26:4505–4519. [PubMed: 17348080]

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;78:13–22.

Lu, L. Explained variation in survival analysis and hypothesis testing for current leukemia-free survival. Medical College of Wisconsin; Milwaukee, Wisconsin: 2006. Ph.D. Dissertation

O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079–1087. [PubMed: 6534410]

Parzen MI, Wei LJ, Ying Z. Simultaneous confidence intervals for the difference of two survival functions. Scandinavian Journal of Statistics 1997;24:309–314.

Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. Biometrics 1989;45:497–507. [PubMed: 2765634]

Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. Journal of the Royal Statistical Society, Series B 1991;53:341–352.

Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika 1990;77:485–497.

Roy SN. On aheuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics 1953;24:220–238.

Sposto R, Stablein D, Carter-Campbell S. A partially grouped logrank test. Statistics in Medicine 1997;16:695–704. [PubMed: 9131757]

Van Besien K, Loberiza F, Bajorunaite R, et al. Comparison of autologous and allogeneic hematopoietic stem cell transplantation for follicular lymphoma. Blood 2003;102:3521–3529. [PubMed: 12893748]

Zhang M-J, Klein JP. Confidence bands for the difference of two survival curves under proportional hazards model. Lifetime Data Analysis 2001;7:243–254. [PubMed: 11677829]

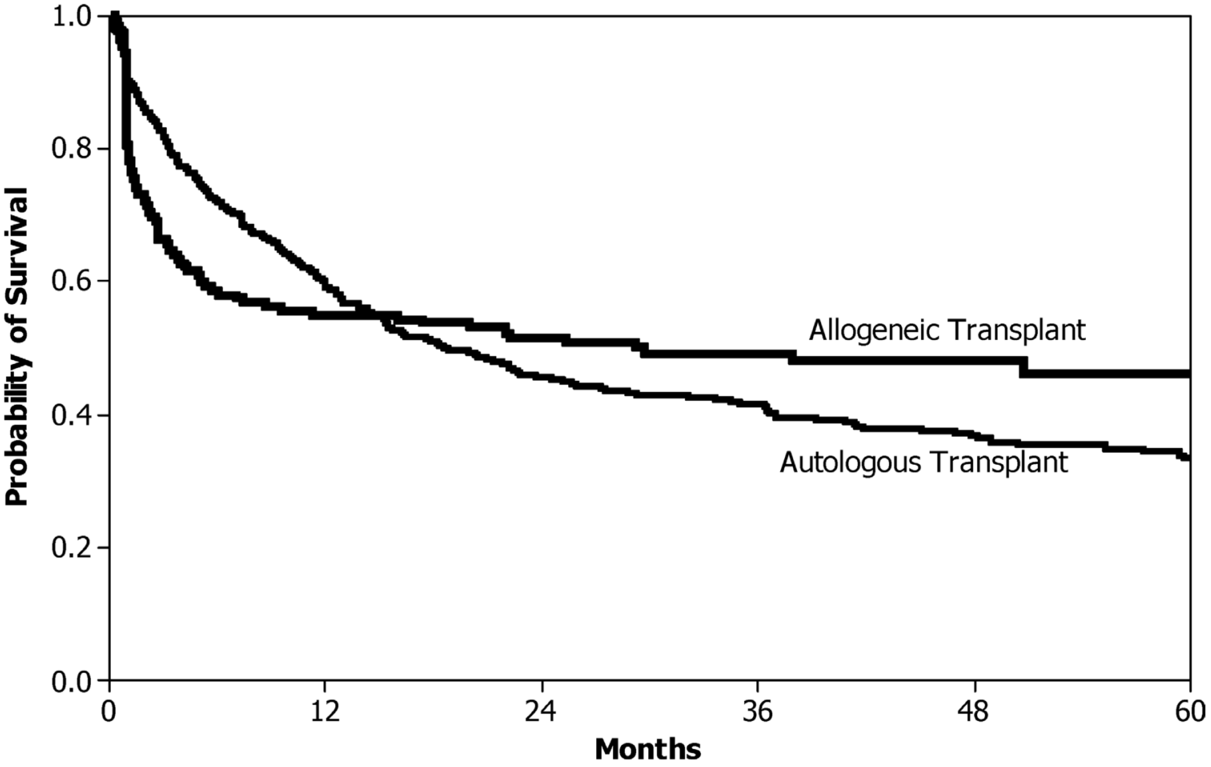
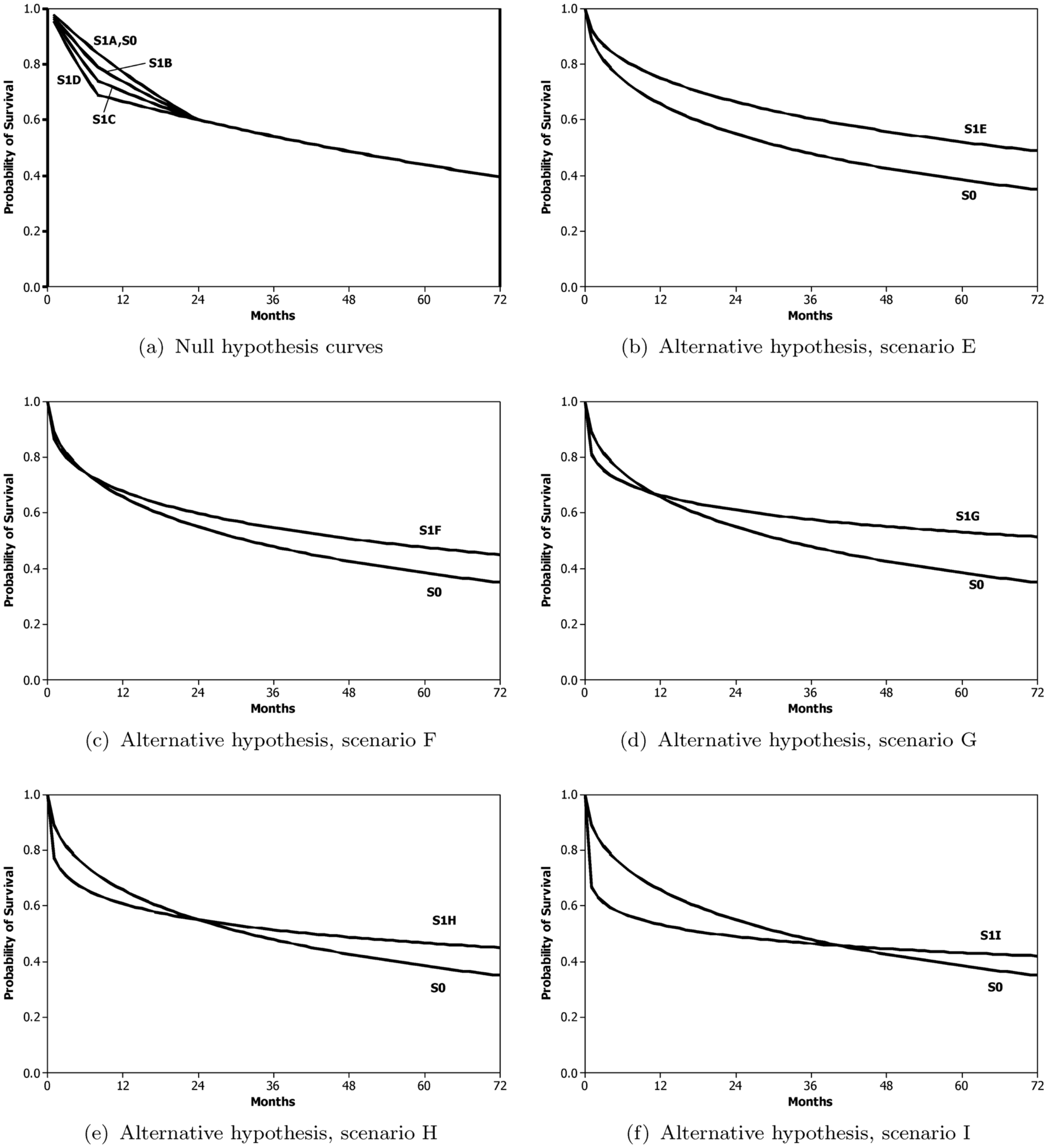


Figure 1.
Kaplan–Meier estimate of DFS for follicular lymphoma example, by stem cell source.



(a) Null hypothesis curves

(b) Alternative hypothesis, scenario E

(c) Alternative hypothesis, scenario F

(d) Alternative hypothesis, scenario G

(e) Alternative hypothesis, scenario H

(f) Alternative hypothesis, scenario I

Figure 2.

Survival curves for treatment (S1) and control (S0) groups used in simulations. Curves for the null hypothesis simulations are shown in (a) for each of the four scenarios, and curves for the alternative hypothesis simulations are shown in (b)–(f) for the five scenarios.

Table 1

Average deviations from nominal 5% *level for* 11 tests adjusted using ANOVA. Deviations are given by scenario, by sample size, by censoring pattern, and overall using models (8–11), respectively. The last two rows refer to the log-rank (*LR*) *test and weighted log-rank* (*WLR*) *tests starting at time* 0. $t_0 = 24$.

		Scenario				Total sample size, *n*			Censoring			
Method	**Equation**	**A**	**B**	**C**	**D**	**100**	**200**	**400**	**0%**	**15%**	**(10%, 20%)**	**Overall**
$Z_{\mathrm{CLL}}(24)$	(1)	−0.05	−0.10	0.08	−0.04	−0.02	0.01	−0.06	0.03	−0.07	−0.03	−0.03
$Z_{\mathrm{CLL}}(48)$	(1)	**0.36**	−0.02	0.16	0.12	0.26	0.18	0.03	**0.47**	−0.04	0.05	0.16
$Z_{\mathrm{CLL}}(72)$	(1)	0.18	−0.02	0.15	0.07	0.24	0.01	0.04	−0.01	0.11	0.19	0.10
$Z_{\mathrm{WKM}}(t_0)$	(2)	**0.84**	**0.59**	**0.71**	**0.61**	**1.18**	**0.59**	0.29	**0.55**	**0.73**	**0.78**	**0.69**
$\chi^2_{\mathrm{PSV}}(t_0)$	(3)	0.24	−0.01	0.04	−0.04	0.12	0.04	0.01	0.04	0.09	0.04	0.06
$Z_{\mathrm{LR}}(t_0)$	(4)	0.21	0.09	−0.04	−0.09	0.08	−0.03	0.07	−0.03	0.05	0.10	0.04
$Z_{\mathrm{OLS}}(t_0)$	(5)	0.10	−0.12	0.01	−0.08	−0.03	−0.05	0.01	−0.12	−0.04	0.09	−0.02
$Z_{\mathrm{SP,P}}(t_0)$	(6)	0.21	0.10	0.15	0.11	0.27	0.08	0.07	0.06	0.19	0.18	0.14
$\chi^2(t_0)$	(7)	−0.13	−0.31	−0.26	**−0.40**	**−0.53**	−0.23	−0.07	**−0.44**	−0.18	−0.21	**−0.27**
Log rank		0.22	**0.38**	**1.40**	**3.04**	**0.67**	**1.11**	**2.00**	**0.77**	**1.40**	**1.61**	**1.26**
Weighted log rank $\rho = 0, \gamma = 1$		0.20	**0.34**	**1.33**	**2.87**	**0.37**	**0.95**	**2.24**	**1.23**	**1.22**	**1.12**	**1.19**

Table 2

Average rejection rates for 11 tests adjusted using ANOVA for censoring pattern. Rejection rates given by scenario using model (12). *The last two rows refer to the log-rank* (*LR*) *test and weighted log-rank* (*WLR*) *tests starting at time* 0. $t_0 = 24$.

		Scenario				
Method	**Equation**	**E**	**F**	**G**	**H**	**I**
$Z_{\mathrm{CLL}}(24)$	(1)	62.4	15.3	21.1	4.7	21.8
$Z_{\mathrm{CLL}}(48)$	(1)	70.1	32.9	65.1	21.5	6.8
$Z_{\mathrm{CLL}}(72)$	(1)	71.2	44.5	85.1	46.1	25.9
$Z_{\mathrm{WKM}}(t_0)$	(2)	75.8	35.0	66.3	20.3	6.0
$\chi^2_{\mathrm{PSV}}(t_0)$	(3)	74.8	32.0	61.2	16.4	4.8
$Z_{\mathrm{LR}}(t_0)$	(4)	30.7	36.5	85.4	71.7	82.6
$Z_{\mathrm{OLS}}(t_0)$	(5)	74.7	43.9	84.1	43.4	23.6
$Z_{\mathrm{SP,P}}(t_0)$	(6)	76.9	40.2	74.8	29.6	10.7
$\chi^2(t_0)$	(7)	67.2	36.7	83.1	61.1	81.0
Log rank		78.0	28.9	47.0	8.6	22.2
Weighted log rank $\rho = 0, \gamma = 1$		64.7	49.7	93.8	70.0	64.6

Table 3

P-values for tests comparing survival curves after 1 *year, applied to the follicular lymphoma dataset*

Method	Equation	*p*-value	Method	Equation	*p*-value
$Z_{\mathrm{CLL}}(36)$	(1)	0.091	$Z_{\mathrm{OLS}}(12)$	(5)	0.020
$Z_{\mathrm{WKM}}(12)$	(2)	0.068	$Z_{\mathrm{SP,P}}(12)$	(6)	0.055
$\chi^2_{\mathrm{PSV}}(12)$	(3)	0.131	$\chi^2(12)$	(7)	<0.001
$Z_{\mathrm{LR}}(12)$	(4)	<0.001			