Published in final edited form as:

Lifetime Data Anal. 2015 April ; 21(2): 218–240. doi:10.1007/s10985-014-9298-4.

Group sequential tests for long-term survival comparisons

Brent R. Logan and

Division of Biostatistics, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226-0509, USA

Shuyuan Mo

Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ, USA

Brent R. Logan: blogan@mcw.edu; Shuyuan Mo: Shuyuan.mo@novartis.com

Abstract

Sometimes in clinical trials, the hazard rates are anticipated to be nonproportional, resulting in potentially crossing survival curves. In these cases, researchers are usually interested in which treatment has better long-term survival. The log-rank test and the weighted log-rank test may not be appropriate or efficient to use here, because they are sensitive to differences in survival at any time and don't just focus on long-term outcomes. Also in a prospective clinical trial, patients are entered sequentially over calendar time, so that group sequential designs may be considered for ethical, administrative and economic concerns. Here we develop group sequential methods for testing the null hypothesis that the survival curves are identical after a prespecified time point. Several classes of tests are considered, including an integrated difference in survival probabilities after this time point, and linear or quadratic combinations of two component test statistics (pointwise comparisons of survival at the time point and comparisons of hazard rates after the time point). We examine the type I errors, stopping probabilities, and powers of these tests through simulation studies under the null and different alternatives, and we apply them to a real bone marrow transplant clinical trial.

Keywords

Crossing hazards; Crossing survival curves; Late survival difference; Group sequential test; Error-spending methods

1 Introduction

In clinical trials, the log-rank test is often used for comparing two survival curves, and it can attain the highest power under the proportional hazards alternative. However, sometimes the survival curves are anticipated to cross, and in this setting researchers are often interested in which treatment has better long-term survival. For example, in an international acute lymphoblastic leukemia (ALL) trial comparing allogeneic transplant versus autologous transplant/chemotherapy (Goldstone et al. 2008), allo transplants might be expected to have

higher mortality in the early time period due to graft-versus-host disease and other complications, while auto transplants might be anticipated to have higher mortality later on due to less protection against relapse from a graft versus leukemia effect. These different shapes of the hazard functions could lead to crossing survival curves, as happened in this trial (the Kaplan–Meier estimates of the survival probabilities can be seen in Fig. 1). More broadly, surgical versus non-surgical intervention trials may encounter a similar issue of anticipated crossing hazards due to differential timing of events. In scenarios like this, the log-rank test and the weighted log-rank test may not be appropriate or efficient, because they are sensitive to differences in survival at any time, and don't just focus on long-term outcomes.

Testing whether there are late survival differences between groups can be formulated as the hypotheses

$$H_0: S_1(t) = S_2(t) \text{ for all } t \geq t_0, \text{ vs. } H_A: S_1(t) \neq S_2(t) \text{ for some } t \geq t_0,$$

where t_0 is a prespecified time point. Logan et al. (2008) proposed several strategies for testing this null hypothesis. Here the parameter t_0 is chosen a priori to focus inference on the clinically relevant late portions of the survival curves. Ideally t_0 should be specified so that the curves cross prior to t_0 if at all, resulting in a more clear interpretation of the trial results. However, even if t_0 is misspecified, testing of these hypotheses still focuses inference and minimizes sensitivity to early differences which is associated with more standard survival analysis procedures. Currently, these strategies have been formulated for a fixed sample design. In a prospective clinical trial, patients are entered sequentially over calendar time, so that group sequential designs may be considered for ethical, administrative and economic concerns. Group sequential designs have been developed for many common survival tests. Many of these satisfy the independent increments structure across calendar times, including the log-rank test (Tsiatis 1982), the Cox model score process (Tsiatis et al. 1985; Biliak et al. 1997), the weighted log-rank test under certain weight conditions (Slud 1984; Gu and Lai 1991), and a pointwise comparison of survival probabilities at a fixed time (Jennison and Turnbull 1985; Lin et al. 1996). Alternatively, for weighted Kaplan–Meier test statistics proposed by Pepe and Fleming (1991; 1989), Li (1999) showed the asymptotic joint distribution across multiple calendar times is multivariate normal, though it does not follow the independent increments structure. Lee and Sather (1995) examined group sequential tests for parametric and nonparametric cure rate models, which may be appropriate when focusing on long-term survival through the proportion cured of disease. We developed group sequential test statistics and their joint distribution for many of the statistics proposed in Logan et al. (2008), and compared them with more standard group sequential test statistics using the log-rank test, weighted log-rank, and pointwise comparisons of the two survival curves. Note that in situations where one is interested in identifying long-term differences in survival, sufficient follow-up is needed for each patient, so there may be limited benefit in terms of patient accrual for incorporating group sequential designs. However, group sequential designs may still offer substantial savings in terms of the time until the research question can be addressed and the research findings disseminated, particularly for rare

diseases where accrual rates are slow and the total study duration is long. In Sect. 2 we derive the group sequential test statistics and their joint distributions. In Sect. 3 we examine the type I errors and powers of these tests through simulation studies under the null and different alternatives, and in Sect. 4 we return to the example of the ALL bone marrow transplant study and apply those test statistics to compare long-term survival for the allogeneic transplant group vs. the auto transplant/chemotherapy group.

2 Methods

In this section, we will first introduce notation based on counting processes, and then review group sequential methods for standard survival tests, including pointwise comparisons of survival and the log-rank test and weighted log-rank test. We will then review the methods in Logan et al. (2008) to test for a late difference in survival curves, and develop group sequential versions of these tests.

2.1 Notation and hypotheses

Suppose there are 2 groups, with n_1 patients in group 1, and n_2 patients in group 2. An individual patient j in group i , $i = 1, 2$, $j = 1, 2, \dots, n_i$, enters the study at calendar time τ_{ij} . He or she either dies at time $\tau_{ij} + T_{ij}$, or is censored at time $\tau_{ij} + C_{ij}$. The observed time for patient j in treatment group i at calendar time t is $X_{ij}(t) = \max\{\min(T_{ij}, C_{ij}, t - \tau_{ij}), 0\}$, and the event indicator is denoted by $I_{ij}(t) = I(T_{ij} \leq \min(t - \tau_{ij}, C_{ij}))$. For example, if a patient enrolls before calendar time t and is still alive at t , then the observed event time $X_{ij}(t)$ for this patient is $t - \tau_{ij}$, and the event indicator is 0 (censored). If a patient enrolls before calendar time t and dies before t , then the observed event time $X_{ij}(t)$ is T_{ij} , and the event indicator is 1 (event). If a patient hasn't entered the study by calendar time t , then the observed event time is 0, and the event indicator is 0, so he or she is excluded from any analyses at calendar time t .

Let

$$\tilde{N}_{ij}(s) = I(T_{ij} \leq s)$$

be the unobservable counting process for the event in the absence of censoring. We can write the observed counting process as

$$N_{ij}(s, t) = I(X_{ij}(t) \leq s) \Delta_{ij}(t) = \int_0^s I_{ij}(u, t) d\tilde{N}_{ij}(u),$$

where

$$I_{ij}(s, t) = I\{C_{ij} \wedge (t - \tau_{ij}) \geq s\}, \quad i=1, 2$$

for patient j in group i at calendar time t and event time s . The martingale of $\tilde{N}_{ij}(s)$ is expressed as

$$M_{ij}(s) = \tilde{N}_{ij}(s) - \int_0^s I\{T_{ij} \geq u\} d\Lambda_i(u).$$

Let $Y_{ij}(s, t) = I(X_{ij}(t) > s) = I_{ij}(s, t)I(T_{ij} > s)$ be the indicator that patient j in group i is at risk at calendar time t and event time s . We can also define

$$N_i(s, t) = \sum_{j=1}^{n_i} N_{ij}(s, t),$$

and

$$Y_i(s, t) = \sum_{j=1}^{n_i} Y_{ij}(s, t)$$

as the total number of observed events and patients at risk, respectively, in treatment group i .

The Kaplan–Meier estimator of the survival function at calendar time t for group i at event time s can be expressed by

$$\hat{S}_i(s, t) = \prod_{u \leq s} \left\{ 1 - \frac{dN_i(u, t)}{Y_i(u, t)} \right\},$$

and the variance estimate for fixed time t is given by the counting process form of Greenwood's formula (Greenwood 1926)

$$\hat{\sigma}_{KM,i}^2(s, t) = \hat{S}_i(s, t)^2 \int_0^s \frac{J_i(u, t) dN_i(u, t)}{Y_i(u, t) (Y_i(u, t) - dN_i(u, t))},$$

where $J_i(s, t) = I(Y_i(s, t) > 0)$, and $0/0 = 0$.

The Nelson–Aalen estimator of the cumulative hazard function at calendar time t for group i at event time s can be expressed as

$$\hat{\Lambda}_i(s, t) = \int_0^s \frac{J_i(u, t) dN_i(u, t)}{Y_i(u, t)},$$

and the variance estimate for fixed calendar time t is given by

$$\hat{\sigma}_{NA,i}^2(s, t) = \int_0^s \frac{J_i(u, t) dN_i(u, t)}{Y_i^2(u, t)}.$$

2.2 Group sequential weighted log-rank test

The most common approach to comparing the survival distributions of two groups is the log-rank test. With the counting process notation, at calendar time t , the weighted log-rank test statistic can be expressed by

$$L_{LR}(t) = \int_0^\tau q(u, t) \frac{Y_1(u, t) Y_2(u, t)}{Y(u, t)} \left[\frac{dN_1(u, t)}{Y_1(u, t)} - \frac{dN_2(u, t)}{Y_2(u, t)} \right],$$

where τ is the maximum study time, and $q(u, t)$ is the weight function. If the weight function $q(u, t) = 1$, we get the usual log-rank test.

The variance of the weighted log-rank test statistic can be written as:

$$\text{Var}(L_{LR}(t)) = \int_0^\tau q^2(u, t) \frac{Y_1(u, t) Y_2(u, t)}{Y(u, t)} \lambda(u) du.$$

The covariance between the log-rank tests at different calendar times $t < t^*$ is given by

$$\text{Cov}(L_{LR}(t), L_{LR}(t^*)) = \int_0^\tau q(u, t) q(u, t^*) \frac{Y_1(u, t) Y_2(u, t)}{Y(u, t)} \lambda(u) du.$$

If the weight function does not depend on calendar time, $q(u, t) = q(u)$, the statistic has the independent increments covariance structure and follows the canonical joint distribution described in Jennison and Turnbull (2000), with information asymptotically equivalent to

$$I_{LR}(t) = n \phi_{LR}(t),$$

where

$$\phi_{LR}(t) = \rho_1 \rho_2 \int_0^\tau q(u)^2 \frac{\pi_1(u, t) \pi_2(u, t)}{\pi(u, t)} \lambda(u) du,$$

$$\pi_i(s, t) = \lim_{n_i \rightarrow \infty} E(Y_i(s, t))/n_i, \quad \pi(s, t) = \lim_{n \rightarrow \infty} E(Y(s, t))/n, \quad \text{and} \quad \rho_i = \lim_{n \rightarrow \infty} n_i/n.$$

Therefore standard techniques for group sequential monitoring can be used. Note that the unweighted log-rank test compares the entire survival curves and is inefficient in the presence of crossing hazards. Even with a weight function favoring late differences in the

hazard functions, such as $q(u) = \hat{S}(u)^p (1 - \hat{S}(u))^q$ with $p = 0, q = 1$ proposed in Fleming and Harrington (1981) and Harrington and Fleming (1982), and used in later simulations, the test still compares the entire curves and does not allow for specific inference about the late region of the survival curves. The weighted log-rank test also does not provide a clinically interpretable parameter estimate, which can be used to indicate the direction of benefit. Particularly for the crossing hazards situation, the weighted average differences in the hazard function may not match the direction of benefit for the survival curves long-term, leading to difficulties in interpretation. The group sequential setting leads to further complications, since the weight functions themselves change over calendar time in the presence of nonproportional hazards.

2.3 Group sequential pointwise comparison test statistic

Another important survival comparison commonly used is a comparison of survival probabilities at a single fixed time point. This could be used in the long-term survival comparison setting by choosing an appropriate late time point, although the restriction to a single time point may lose efficiency as described in Logan et al. (2008). Notice that the pointwise comparison of two survival curves $S_i(\tau_0)$ at time $\tau_0, i = 1, 2$ is equivalent to testing the null hypothesis $H_0: \Lambda_1(\tau_0) = \Lambda_2(\tau_0)$. Then the group sequential test statistic of the difference in Nelson-Aalen estimates at calendar time t is

$$L_{NA}(\tau_0, t) = \hat{\Lambda}_1(\tau_0, t) - \hat{\Lambda}_2(\tau_0, t).$$

The variance estimator can be expressed by

$$\sigma_{NA}^2(\tau_0, t) = \int_0^{\tau_0} \frac{dN_1(u, t)}{Y_1^2(u, t)} + \int_0^{\tau_0} \frac{dN_2(u, t)}{Y_2^2(u, t)}.$$

For two calendar time points $t < t^*$, the covariance of the group sequential Nelson–Aalen estimators at t and t^* as shown in Lin et al. (1996) can be expressed as

$$\begin{aligned} \text{Cov}(L_{NA}(\tau_0, t), L_{NA}(\tau_0, t^*)) &= \int_0^{\tau_0} \frac{\lambda_1(u) du}{Y_1(u, t^*)} + \int_0^{\tau_0} \frac{\lambda_2(u) du}{Y_2(u, t^*)} \\ &= \text{Var}(L_{NA}(\tau_0, t^*)). \end{aligned}$$

Since this covariance follows the independent increments structure and the statistics are asymptotically multivariate normal over a set of calendar times, the canonical joint distribution described in Jennison and Turnbull (2000) holds and the standard techniques for analyzing group sequential test statistics can be applied. In particular, the information for the difference in Nelson-Aalen estimates used in group sequential monitoring is asymptotically equivalent to

$$I_{NA}(\tau_0, t) = \frac{n}{\phi_{NA}(\tau_0, t)},$$

where

$$\phi_{NA}(\tau_0, t) = \left[\frac{1}{\rho_1} \int_0^{\tau_0} \frac{\lambda_1(u) du}{\pi_1(u, t)} + \frac{1}{\rho_2} \int_0^{\tau_0} \frac{\lambda_2(u) du}{\pi_2(u, t)} \right].$$

Here there is a clinically interpretable parameter estimate (the difference in cumulative hazards or survival probabilities at τ_0) associated with these tests, and the same parameter is being estimated at each time point.

2.4 Group sequential weighted Kaplan–Meier test

One strategy for comparing late differences in survival proposed in Logan et al. (2008) was a modification of the weighted Kaplan–Meier (WKM) test in Pepe and Fleming (1991; 1989), where the integral starts at a lower bound of t_0 to only include survival differences after t_0 . Li (1999) considered the joint distribution of the standard WKM test across calendar time in a group sequential design setting. Murray and Tsiatis (1999) considered an unweighted integrated difference in survival distributions (restricted mean survival, RMS) in the group sequential setting. The weight function used in the WKM test is primarily a tool to automatically discount parts of the curve where there is a lot of variability due to heavy censoring. However, this weight function can complicate interpretation of the test statistic. This discounting can alternatively be done by using an unweighted RMS statistic and limiting the integrated survival difference to an appropriate upper limit τ where there is sufficient data for estimation. By doing this, the clinical interpretation is more clear as the difference in mean survival time or life years between t_0 and τ . However, both statistics are complicated by use in a group sequential setting. With the WKM test, the weight function changes with calendar time, so that a different parameter is being estimated at each calendar time. With the RMS test, there may not be sufficient follow-up early to estimate the restricted mean survival over the region of interest, so one may need to increase the upper limit as calendar time progresses, thereby leading to changes in the parameter being estimated at each calendar time. The use of weights equal to 0 prior to t_0 focuses inference on late differences in survival curves, even if the weight function for the WKM test decreases at later time points as censoring increases. Also note that as t_0 approaches 0, the proposed statistics reduce to the usual Weighted Kaplan–Meier or restricted mean survival comparison over the entire curve; the specification of t_0 simply allows one to focus inference on late survival differences.

Following derivations from Li (1999), we modify the group sequential weighted Kaplan–Meier test to compare late differences in survival curves. The test statistic at calendar time t can be expressed as

$$WKM(t_0, t) = \sqrt{\frac{n_1 n_2}{n}} \int_{t_0}^t \hat{w}(u, t) \{ \hat{S}_1(u, t) - \hat{S}_2(u, t) \} du,$$

where $\hat{w}(s, t)$ is the estimated weight function to stabilize the integrated difference of $\hat{S}_1(s, t) - \hat{S}_2(s, t)$ under heavy censoring. If we define

$$G_i(s, t) = \sum_{j=1}^{n_i} I_{ij}(s, t),$$

a simple weight function satisfying the regularity conditions given by Li (1999) is

$$\hat{w}(s, t) = \frac{\hat{G}_1(s, t) \hat{G}_2(s, t)}{\hat{\rho}_1 \hat{G}_1(s, t) + \hat{\rho}_2 \hat{G}_2(s, t)},$$

where $\hat{G}_i(s, t)$ is an empirical estimate of $G_i(s, t)$. We will use this weight function in later simulations for group sequential weighted Kaplan–Meier test.

The statistic $WKM(t_0, t)$ at calendar time t follows an asymptotic Gaussian distribution with variance

$$\text{Var}(WKM(t_0, t)) = \rho_2 \sigma_1^2(t_0, t) + \rho_1 \sigma_2^2(t_0, t),$$

where

$$\sigma_i^2(t_0, t) = \int_0^t \frac{h_i^2(u, t)}{\pi_i(u, t)} \lambda_i(u) du,$$

$$\text{and } h_i(s, t) = \int_s^t I\{u > t_0\} w_i(u, t) S_i(u, t) du.$$

The group sequential weighted Kaplan–Meier test does not have an independent increments structure across calendar time. For $t < t^*$, the covariance of $WKM(t_0, t)$ and $WKM(t_0, t^*)$ is given by

$$\rho_2 \int_0^t \frac{h_1(u, t) h_1(u, t^*)}{\pi_1(u, t^*)} \lambda_1(u) du + \rho_1 \int_0^t \frac{h_2(u, t) h_2(u, t^*)}{\pi_2(u, t^*)} \lambda_2(u) du.$$

Asymptotic multivariate normality across multiple calendar times follows from the multivariate central limit theorem.

The corresponding expressions for the RMS statistic can be obtained by modifying the weight function as $w(s, t) = I(s \leq \tau(t))$, where $\tau(t)$ is the upper limit to the integral used at calendar time t .

2.5 Combination tests

Another strategy considered for long-term survival comparisons in Logan et al. (2008) was to break the overall null hypothesis $H_0 : S_1(t) = S_2(t)$ for all $t \geq t_0$ further into the intersection of two sub-hypotheses. Separate test statistics for each of these sub-hypotheses can then be combined into a single test of H_0 . Specifically, H_0 can be written as $\{H_{01} : S_1(t_0) = S_2(t_0)\} \cap \{H_{02} : \lambda_1(t) = \lambda_2(t), t > t_0\}$. Null sub-hypothesis H_{01} can be tested using the difference in the Nelson–Aalen estimators evaluated at time point t_0 , while null sub-hypothesis H_{02} can be tested using the left-truncated log-rank test starting at time point t_0 . Logan et al. (2008) evaluated the left truncated log-rank test alone and found that substantial loss of power could occur because it ignored survival differences accumulating prior to t_0 ; therefore we focus in this paper on linear or quadratic combinations of the component test statistics. One potential drawback of these combination tests is that they are restricted to testing only, and do not provide useful estimates of treatment effect. Also, the test is two-sided, leading to a conclusion about whether the survival curves are different after t_0 , but it does not provide directional inference since the survival difference at t_0 and the hazard ratio after t_0 may be in the opposite direction.

We have already discussed the joint distribution of the pointwise differences in the Nelson–Aalen estimates over calendar time in Sect. 2.3. Extending the weighted log-rank test statistics to compare late differences in hazard rates is trivial. The test statistic at calendar time t can be expressed as

$$L_{LR}(t_0, t) = \int_{t_0}^{\tau} q(u, t) \frac{Y_1(u, t)Y_2(u, t)}{Y(u, t)} \left[\frac{dN_1(u, t)}{Y_1(u, t)} - \frac{dN_2(u, t)}{Y_2(u, t)} \right] du.$$

with variance estimator

$$\hat{\sigma}_{LR}^2(t_0, t) = \int_{t_0}^{\tau} q(u, t)^2 \frac{Y_1(u, t)Y_2(u, t)dN(u, t)}{Y^2(u, t)}.$$

For time point $t < t^*$, one can show that the covariance between $L_{LR}(t_0, t)$ and $L_{LR}(t_0, t^*)$ above is

$$\text{Cov}(L_{LR}(t_0, t), L_{LR}(t_0, t^*)) = \int_{t_0}^{\tau} q(u, t)q(u, t^*) \frac{Y_1(u, t)Y_2(u, t)}{Y(u, t)} \lambda(u) du.$$

If the weight function $q(u, t) = 1$, we have the log-rank test statistic, and

$$\text{Cov}(L_{LR}(t_0, t), L_{LR}(t_0, t^*)) = \text{Var}(L_{LR}(t_0, t)),$$

yielding an independent information increments structure. By the central limit theorem, we can also see that $\{L_{LR}(t_0, t_1), \dots, L_{LR}(t_0, t_K)\}$ follows a multivariate normal distribution. The information associated with this modified weighted log-rank test statistic is asymptotically equivalent to

$$I_{LR}(t_0, t) = n\phi_{LR}(t_0, t),$$

where

$$\phi_{LR}(t_0, t) = \rho_1 \rho_2 \int_{t_0}^t q(u, t)^2 \frac{\pi_1(u, t) \pi_2(u, t)}{\pi(u, t)} \lambda(u) du.$$

One way to combine the test statistics $L_{NA}(t_0, t)$ and $L_{LR}(t_0, t)$ into a single test of H_0 is to use constant weights on the Z-scale as proposed in Logan et al. (2008),

$$C(t_0, t) = \frac{Z_{NA}(t_0, t) + Z_{LR}(t_0, t)}{\sqrt{2}},$$

where

$$Z_{NA}(t_0, t) = \frac{L_{NA}(t_0, t)}{\hat{\sigma}_{NA}(t_0, t)} \xrightarrow{D} \frac{\sqrt{n}L_{NA}(t_0, t)}{\sqrt{\phi_{NA}(t_0, t)}}$$

and

$$Z_{LR}(t_0, t) = \frac{L_{LR}(t_0, t)}{\hat{\sigma}_{LR}(t_0, t)} \xrightarrow{D} \frac{L_{LR}(t_0, t)}{\sqrt{n} \sqrt{\phi_{LR}(t_0, t)}}.$$

Using independence between $Z_{NA}(t_0, t)$ and $Z_{LR}(t_0, t)$, the covariance between the statistic at two different calendar times is then asymptotically

$$\begin{aligned} \text{Cov}(C(t_0, t), C(t_0, t^*)) &= \frac{1}{2} \text{Cov}(Z_{NA}(t_0, t), Z_{NA}(t_0, t^*)) + \frac{1}{2} \text{Cov}(Z_{LR}(t_0, t), Z_{LR}(t_0, t^*)) \\ &= \frac{1}{2} \left[\sqrt{\frac{I_{NA}(t_0, t)}{I_{NA}(t_0, t^*)}} + \sqrt{\frac{I_{LR}(t_0, t)}{I_{LR}(t_0, t^*)}} \right], \end{aligned}$$

which does not follow independent increments over calendar time points.

Another way to combine the two component tests is to use a partially grouped log-rank test, which was proposed for the group sequential setting in Sposto et al. (1997). The test statistic is

$$S_p(t_0, t) = \left[\frac{n_1 n_2}{n} (\hat{S}_2(t_0, t) - \hat{S}_1(t_0, t)) \right] + L_{LR}(t_0, t),$$

where $\hat{S}_i(t_0, t)$ is the Kaplan–Meier estimate at time point t_0 for group i .

Since

$$\hat{S}(t) - S(t) \approx -S(t)(\hat{\Lambda}(t) - \Lambda(t)),$$

then the covariance between test statistics at two different calendar times is

$$\text{Cov}(S_p(t_0, t), S_p(t_0, t^*)) = n^2 \rho_1^2 \rho_2^2 S^2(t_0) I_{NA}^{-1}(t_0, t^*) + I_{LR}(t_0, t),$$

which also does not have independent increments over calendar time points.

An alternative way to combine the two test statistics is to work on the score test scales. The non-standardized form of this test statistic is

$$U_N(t_0, t) = L_{NA}(t_0, t) / \hat{\sigma}_{NA}^2(t_0, t) + L_{LR}(t_0, t),$$

which is asymptotically equivalent to $L_{NA} I_{NA} + L_{LR}$. The covariance between these tests computed at two different calendar times is

$$\text{Cov}(U_N(t_0, t), U_N(t_0, t^*)) = I_{NA}(t_0, t) + I_{LR}(t_0, t) = \text{Var}(U_N(t_0, t)),$$

so that this test statistic has the independent increments structure.

We can also modify the above test statistic by reweighting the two components according to the maximum information of each component at the final calendar time T . Then the resulting linear combination test statistic can be written by

$$U_S(t_0, t) = \frac{L_{NA}(t_0, t)}{\hat{\sigma}_{NA}^2(t_0, t)} \frac{1}{\sqrt{I_{NA}(t_0, T)}} + \frac{L_{LR}(t_0, t)}{\sqrt{I_{LR}(t_0, T)}}.$$

The covariance between the statistics at two different calendar times is

$$\text{Cov}(U_S(t_0, t), U_S(t_0, t^*)) = \frac{I_{NA}(t_0, t)}{I_{NA}(t_0, T)} + \frac{I_{LR}(t_0, t)}{I_{LR}(t_0, T)} = \text{Var}(U_S(t_0, t)).$$

so that the reweighted form of this combination also has independent increments over calendar time points.

This reweighted test statistic has potential advantages over C and U_N , since not only does it have an independent increments structure over calendar time points, but also it converges to the constant weight test (C) at the end of the study. However, it does require that the information for the two components at the final analysis is approximately known. This information would typically be used in the clinical trial design process. Finally, note that as t_0 approaches 0, the Sposto test and the non-standardized linear combination tests reduce to the standard log-rank test, while the others do not because they allocate a specific weight to the pointwise comparison of survival at t_0 .

2.6 Group sequential quadratic combination test

Logan et al. (2008) also proposed a quadratic form of the combination test based on the standardized statistics $Z_{NA}(t_0)$ and $Z_{LR}(t_0)$, as

$$Q(t_0) = Z_{NA}^2(t_0) + Z_{LR}^2(t_0)$$

for the fixed sample design. Under H_0 , it follows a χ_2^2 distribution.

Here we extend this test statistic to the group sequential design setting as

$$Q(t_0, t) = Z_{NA}^2(t_0, t) + Z_{LR}^2(t_0, t),$$

for calendar time t . The marginal distribution of $Q(t_0, t)$ for fixed calendar time under H_0 is also χ_2^2 .

Note that while $Z_{NA}(t_0, t)$ and $Z_{LR}(t_0, t)$ have the Markov property, the quadratic statistic $Q(t_0, t)$ does not. Therefore, we cannot use the methods in Jennison and Turnbull (1997) for group sequential χ^2 tests. Instead we use an error spending method to attribute the overall type I error α over multiple looks.

Suppose we have k looks, and let p_k be the type I error spent at the k^{th} look, and α_k be the cumulative type I error spent by the k^{th} look. For simplicity of notation, we write $Q^{(k)} = Q(t_0, t_k)$, $Z_{NA}^{(k)} = Z_{NA}(t_0, t_k)$, and $Z_{LR}^{(k)} = Z_{LR}(t_0, t_k)$. Then we have

$$P(Q^{(1)} > c_1 | H_0) = p_1 = \alpha_1, \\ P(Q^{(1)} < c_1, \dots, Q^{(k-1)} < c_{k-1}, Q^{(k)} > c_k | H_0) = p_k = \alpha_k - \alpha_{k-1}.$$

Let $R(c_k)$ be the rejection region of $Q_{(k)}$, defined by

$$R(c_k) = \{Z_{NA}^{(k)}, Z_{LR}^{(k)} | (Z_{NA}^{(k)})^2 + (Z_{LR}^{(k)})^2 > c_k\}.$$

and let $A(c_k)$ be the complementary acceptance region. The critical values c_k can be defined recursively as follows. The first critical value is $c_1 = \text{inv}(\chi_2^2(\alpha_1))$, while subsequent critical values satisfy

$$p_k = P(Q^{(1)} < c_1, \dots, Q^{(k-1)} < c_{k-1}, Q^{(k)} > c_k | H_0) \\ = \int_0^{c_1} \dots \int_0^{c_{k-1}} \int_{c_k}^{\infty} f(Q^{(1)}, \dots, Q^{(k)}) dQ^{(k)} \dots dQ^{(1)}$$

Due to independence between $Z_{NA}^{(k)}$ and $Z_{LR}^{(k)}$ as well as the Markov property for each component, we can write p_k as

$$\iint_{A(c_1)} \dots \iint_{A(c_{k-1})} \iint_{R(c_k)} f(Z_{NA}^{(1)}) f(Z_{LR}^{(1)}) f(Z_{NA}^{(2)} | Z_{NA}^{(1)}) f(Z_{LR}^{(2)} | Z_{LR}^{(1)}) \dots \\ f(Z_{NA}^{(k)} | Z_{NA}^{(k-1)}) f(Z_{LR}^{(k)} | Z_{LR}^{(k-1)}) dZ_{NA}^{(k)} dZ_{LR}^{(k)} \dots dZ_{NA}^{(1)} dZ_{LR}^{(1)}.$$

In practice, Monte Carlo integration can be easily used to calculate the critical values c_k at the k th look. This is implemented by simulating a sequence of pairs from the Markov conditional distributions:

$$Z_{NA}^{(k)} | Z_{NA}^{(k-1)} \sim N(\eta_{NA,k} Z_{NA}^{(k-1)}, 1 - \eta_{NA,k}^2),$$

and

$$Z_{LR}^{(k)} | Z_{LR}^{(k-1)} \sim N(\eta_{LR,k} Z_{LR}^{(k-1)}, 1 - \eta_{LR,k}^2),$$

where

$$\eta_{NA,k} = \frac{\sigma_{NA}(t_0, t_k)}{\sigma_{NA}(t_0, t_{k-1})}$$

and

$$\eta_{LR,k} = \frac{\sigma_{LR}(t_0, t_{k-1})}{\sigma_{LR}(t_0, t_k)}.$$

Then we obtain a Monte Carlo sample from the distribution of $Q^{(k)}|Q^{(k-1)}$ by

$$Q^{(k)} = (Z_{NA}^{(k)})^2 + (Z_{LR}^{(k)})^2.$$

Assuming B Monte Carlo samples of $Q_b^{(j)}$ for $b = 1, \dots, B$ and $j = 1, \dots, k$, by recursively solving the equation

$$p_k = \frac{1}{B} \sum_{b=1}^B I(Q_b^{(1)} \leq c_1) \cdots I(Q_b^{(k-1)} \leq c_{k-1}) I(Q_b^{(k)} > c_k)$$

we can get the critical value c_k . Alternatively, the critical value c_k is just the $1 - (a_k - a_{k-1})/(1 - a_{k-1})$ percentile of the total B sorted samples of $Q_b^{(k)}$ where the corresponding $Q_b^{(j)} \leq c_j, j = 1, \dots, k - 1$.

3 Simulation studies

In order to compare the performance of the group sequential test statistics mentioned in previous sections, we conducted simulation studies under three null hypothesis scenarios and 4 different alternative hypothesis scenarios. We assume patients are uniformly accrued over $A = 3$ and $A = 2$ years with total study time of $T = 5$ years. We used a cutpoint of $t_0 = 2$ years to define late survival. Simulations under H_0 featured an early difference in survival functions which disappears by time t_0 . These were obtained by generating survival curves from piece-wise Weibull distributions assuming different shape parameters α for the two groups before time t_0 , and the same α for the two groups after time t_0 (Fig. 2). Note that this definition of type I error is different than the usual one which is calculated assuming the survival curves are equal. This is used because of the focus on comparing survival curves after t_0 . For the alternative hypothesis scenarios, we generated survival curves from a Weibull distribution, with proportional hazards (alternative scenario 1), survival curves crossing at t_0 (alternative scenario 2), before t_0 (alternative scenario 3), and after t_0 (alternative scenario 4) (Fig. 3).

We planned four interim analyses with equal increments in information times (information fraction $f=0.25, 0.5, 0.75$ and 1) as well as equal increments in calendar times (calendar times = $2.75, 3.5, 4.25$ and 5 years). No additional censoring other than administrative censoring from study entry was used. Both O'Brien-Fleming and Pocock boundaries were analyzed. For test statistics which don't have independent information increments, Monte Carlo integration ($B = 2,000,000$ samples) was used to find the critical values under an error spending approach where the cumulative type I error spent at each of the 4 looks is

calibrated to the standardized linear combination test (L_S). For the null hypothesis scenarios, sample sizes of 100 and 300 per group were studied. For the alternative hypothesis scenarios, we used a sample size of 300 per group to reduce the impact of inflation of the type I error rate that could occur with small sample sizes. All simulation scenarios used 10,000 replications. Only the results with $A = 2$, equal calendar time increments, and an O'Brien-Fleming boundary are shown. Other results show similar findings.

Table 1 shows simulation results for the type I error of each group sequential test statistic under the three null hypothesis scenarios. Listed in the tables are the cumulative type I error across the 4 interim looks. The tests considered include the non-standardized linear combination test (L_N), the standardized linear combination test (L_S), the constant weight test (C), the Sposto et al. test (S_p), the quadratic combination test (Q), the log-rank test (LR), the weighted log-rank test (WLR), the weighted Kaplan–Meier test (WKM), the restricted mean survival test (RMS) with an upper limit $\pi(t)$ corresponding to the 85th percentile of the censoring distribution, and the pointwise comparison conducted at 2 years (NA(2)) and 3 years (NA(3)). Table 2 shows simulation results for the cumulative power by each interim look of each group sequential test statistic under the 4 alternative hypotheses scenarios.

For type I errors, we can see that for each scenario, the log-rank test and the weighted log-rank test don't control the type I error rate specifically for the test of late differences. This is not surprising, since they are testing for an overall difference in the survival curves, instead of testing for a late difference after t_0 as the other test statistics do. Another implication of this is that the log-rank and weighted log-rank tests will tend to hit the stopping boundary early, even though there is no long-term difference in survival curves, because they are sensitive to early differences. This may lead to premature conclusions about the study with insufficient follow-up. All other test statistics controlled the type I error rate when used with an O'Brien-Fleming type spending function.

Next we look at powers under different alternatives. Under scenario 1 (proportional hazards), the 11 tests are similar with 80–90 % power. This is important because even though the treatment differences are starting before t_0 , the tests are still sensitive to those differences and there is only a small loss of power compared to the log-rank test. Under scenario 2 (crossing at $t_0 = 2.0$ years), the weighted log-rank test does best (overall power 98%), followed by the quadratic combination test (overall power 93%), and the standardized linear combination test and constant weight test (overall power ≈ 74 %), leaving other tests far behind. Under scenario 3 (crossing before $t_0 = 2.0$ years), the weighted log-rank test does best (overall power 98 %), followed by the standardized linear combination test, constant weight test, Sposto test, and quadratic combination test. Under scenario 4 (crossing after $t_0 = 2.0$ years), the quadratic combination test does best (overall power 95 %), followed by the weighted log-rank tests and log-rank (overall powers 87 and 67 %), leaving other tests far behind (overall powers less than 32 %). Notice here the log-rank test tends to stop the study early, with 66 % probability of rejecting H_0 at the very first look, compared to its overall power of 67 %. However, it stops early in favor of the wrong treatment with worse long-term outcomes.

In summary, the standardized linear combination test and the quadratic combination test are comparable with other tests under proportional hazards scenarios, and they do better when survival curves cross before t_0 . The quadratic combination test does better than the standardized linear combination test when the survival curves cross at or after t_0 (and much better than other tests), while the linear combination test performs better than the quadratic combination test when the curves cross prior to t_0 . Both of them control the type I error under the early difference scenario very well when applied in a group sequential setting. Note that the log-rank test and the weighted log-rank test compare the entire curves. While the power for the weighted log-rank test is higher in some cases, inference is less specific about the actual effect of treatment on long-term survival outcomes. Thus these results are not directly comparable to those tests which specifically compare long-term survival. Finally, while the WKM and RMS tests have lower power than the combination tests, they have the advantage of being associated directly with a clinical parameter for estimation, namely the weighted or restricted mean survival difference after t_0 . The RMS test has higher power than the WKM test in most scenarios, likely due to the decreasing weight placed on later time points in the WKM test.

4 Example

In this section we return to the example presented in the introduction section, and apply the proposed group sequential test statistics retroactively to this international ALL trial (MRC UKALLXII/ECOG E2993) discussed in Goldstone et al. (2008). In this study which compares allogeneic transplant (cells from donor) vs. autologous transplant (re-infusion of own cells)/chemotherapy, patients were recruited between 1993 and 2006. The total study duration is 14 years. With a focus on just Ph negative patients, there are 443 patients in the allo group, and 588 in the auto/chemo group. As described in the introduction, Fig. 1 shows the Kaplan-Meier estimates for the survival curves in the two groups. We can see from the plot that the survival curves of two groups cross between 2 and 3 years, then they come slightly closer again after 12 years although the risk set is small. In practice, t_0 would need to be prespecified prior to conduct of the trial, using clinical experience and external data to determine a target late time period of interest. However, we examine the performance of the various tests for a range of values of t_0 that might be used for this study from 2 to 4 years. Although we have the final dataset, we apply the methods of this manuscript as if we were conducting the trial with group sequential monitoring at 10 yearly interim analyses starting at year 5. The final information at the end of the study is based on the observed data and treated as if it were known, even though in practice this would need to be specified as part of the design. Rejection boundaries and test statistics at each look are shown in Fig. 4 for a select subset of procedures assuming $t_0 = 3$ (group sequential standardized linear and quadratic combination tests, the RMS test, and differences in NA estimates at 3 years). For $t_0 = 3$ years, this example is similar to alternative scenario 3 (survival curves cross before t_0), except that at the end the two groups come slightly closer to each other rather than continuing to separate.

From the boundary plots, we can see the standardized linear combination test and RMS tests stop at year 11 while the quadratic combination test and the Nelson–Aalen pointwise comparison at 3 years never reject H_0 . Therefore, instead of looking at the data at the end of

the study period (14 years) as for a fixed sample design, we can obtain the conclusion that there is a long-term survival difference (beyond 3 years) between the two groups much earlier (at 11 years), using the group sequential standardized linear combination test or the RMS test.

In Table 3 we show the results for all the proposed tests using several values of t_0 between 2 and 4 years to examine sensitivity of the procedures to different choices of t_0 . The example results generally follow the pattern seen in the simulation study. When $t_0 = 2$, so that the curves actually cross near t_0 , the quadratic test is most efficient stopping 7 years into the trial, while the standardized linear combination test and restricted mean survival test also perform well although stopping later. When $t_0 = 3$ or 4, most of the procedures stop to reject the null hypothesis 11 years into the trial. The quadratic test fails to reject the null hypothesis because it is less sensitive when t_0 is past the stopping point. The pointwise difference in NA estimates also performs worse because it ignores information past t_0 and is therefore less sensitive. Finally, note that some of the top performing methods (standardized linear combination test and RMS test) identify a significant treatment difference at a consistent interim analysis time point regardless of the prespecified t_0 , indicating that they are sensitive to long-term differences regardless of the time point used to define those long-term differences of interest.

5 Discussion

In order to test for late differences in survival curves and adapt to the accumulating information gathered during the period of the clinical trial, we derived group sequential linear and quadratic combination test statistics and extended the group sequential weighted Kaplan–Meier test to account for survival comparisons after a prespecified time point t_0 . We examined the performances of these various methods in terms of type I error and power through simulation studies, and showed that the standardized linear combination test and the quadratic combination test are comparable with other tests under proportional hazards scenarios, and they are superior in other settings. The quadratic combination test does better than the standardized linear combination test when the survival curves cross at or after t_0 , while the standardized linear combination test does better than the quadratic combination test when the survival curves cross before t_0 . Among those group sequential tests, the standardized and non-standardized linear combination tests are easier to conduct, since they have an independent increments structure over calendar time, which facilitates calculation of critical values. The weighted Kaplan–Meier or restricted mean survival test had lower power than the combination tests in some settings, but has the advantage of being tied to a parameter for estimation. For the constant weight test, the weighted Kaplan–Meier test, the restricted mean survival test, the Sposto et al. test, and the quadratic combination test, an error spending function needs to be used in order to calculate the corresponding rejection boundaries at each look. Although we showed that the tests perform well under different alternative scenarios, a time point t_0 still needs to be prespecified in order to conduct the analyses. The time point t_0 is chosen to define “long-term” survival benefit, and the appropriate choice depends on the nature of the different clinical studies. It should ideally be selected after potentially anticipated crossing hazard rates and survival curves, so that the difference in survival after t_0 are in a consistent direction and more easily interpretable. In

general, the performance of the various procedures may depend on the time point t_0 because their power depends on the survival differences after t_0 . However, even if t_0 is poorly specified, the procedures here are still less sensitive to early differences than more standard methods such as the log-rank test, and the methods allow the researchers to focus their inference on the part of the survival curve that is of primary interest. In order to compare long-term survival differences, the study period for these clinical trials are usually long, particularly in the transplant setting where crossing hazards may be anticipated, the diseases are rare and patient accrual may be slow. Therefore even for long-term survival comparisons group sequential testing can offer important benefits.

Acknowledgments

The authors would like to thank Dr. Susan Richards and Ms. Georgina Buck at the Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, for providing the deidentified dataset of the example used in the paper. This research was partially supported by a Grant (R01 CA54706-14) from the National Cancer Institute.

References

- Bilias Y, Gu M, Ying Z. Towards a general asymptotic theory for the Cox model with staggered entry. *Ann Stat.* 1997; 25:662–682.
- Fleming TR, Harrington DP. A class of hypothesis tests for one and two samples of censored survival data. *Commun Stat.* 1981; 10:763–794.
- Goldstone AH, Richards SM, Lazarus HM, Tallman MS, Buck G, Fielding AK, et al. In adults with standard-risk acute lymphoblastic leukemia, the greatest benefit is achieved from a matched sibling allogeneic transplantation in first complete remission, and an autologous transplanation is less effect than coventional consolidation/maintenance chemotherapy in all patients. *Blood.* 2008; 111:1827–1833. [PubMed: 18048644]
- Greenwood M. The natural duration of cancer. *Rep Public Health Med Subj.* 1926; 33:1–26.
- Gu M, Lai T. Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. *Ann Stat.* 1991; 19:1403–1433.
- Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika.* 1982; 69:133–143.
- Jennison C, Turnbull BW. Repeated confidence intervals for the median survival time. *Biometrika.* 1985; 72:619–625.
- Jennison C, Turnbull BW. Distribution theory of group sequential t , χ^2 and F tests for general linear models. *Seq Anal.* 1997; 16:295–317.
- Jennison, C.; Turnbull, BW. Group sequential tests with applications to clinical trials. Chapman and Hall/CRC; Boca Raton: 2000.
- Lee JW, Sather HN. Group sequential methods for comparison of cure rates in clinical trails. *Biometrics.* 1995; 51:756–763. [PubMed: 7662857]
- Li Z. A group sequential test for survival trials: an alternative to rank-based procedures. *Biometrics.* 1999; 55:277–283. [PubMed: 11318168]
- Lin DY, Shen L, Ying Z, Breslow NE. Group sequential designs for monitoring survival probabilities. *Biometrics.* 1996; 52:1033–1041. [PubMed: 8805766]
- Logan BR, Klein J, Zhang M-J. Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics.* 2008; 64:733–740. [PubMed: 18190619]
- Murray SA, Tsiatis AA. Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics.* 1999; 55:1085–1092. [PubMed: 11315052]
- 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. *J R Stat Soc B.* 1991; 53:341–352.

- Slud EV. Sequential linear rank tests for two-sample censored survival data. *Ann Stat.* 1984; 12:551–571.
- Sposto R, Stablein D, Carter-Campbell S. A partially grouped logrank test. *Stat Med.* 1997; 16:695–704. [PubMed: 9131757]
- Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *J Am Stat Assoc.* 1982; 77:855–861.
- Tsiatis AA, Rosner GL, Titchler DL. Group sequential tests with censored survival data adjusting for covariates. *Biometrika.* 1985; 72:365–373.

Fig. 1.
Kaplan–Meier estimates for survival curves in two groups

Fig. 2. Null hypothesis scenarios for simulation study. **a** Null hypothesis scenario 1. **b** Null hypothesis scenario 2. **c** Null hypothesis scenario 3

Fig. 3. Alternative hypothesis scenarios for simulation study. **a** Alternative hypothesis scenario 1. **b** Alternative hypothesis scenario 2. **c** Alternative hypothesis scenario 3 **d** Alternative hypothesis scenario 4

Fig. 4. Group sequential boundaries and test statistics applied to the ALL trial. **a** Boundary for standardized linear combination test. **b** Boundary for quadratic combination test. **c** Boundary for RMS test. **d** Boundary for NA pointwise comparison at 3 years

Table 1

Cumulative type I error rate for null hypothesis scenarios, using an OBF boundary with equal calendar time increments and accrual time $A = 2$

Test	Scenario 1		Scenario 2		Scenario 3	
	$n = 100$	$n = 300$	$n = 100$	$n = 300$	$n = 100$	$n = 300$
L_S	0.049	0.052	0.049	0.052	0.049	0.052
Q	0.048	0.053	0.048	0.053	0.048	0.053
S_p	0.050	0.052	0.049	0.052	0.050	0.052
C	0.050	0.051	0.049	0.051	0.049	0.051
WKM	0.048	0.051	0.046	0.051	0.046	0.050
RMS	0.053	0.053	0.052	0.052	0.051	0.052
L_N	0.047	0.050	0.045	0.050	0.046	0.051
LR	0.298	0.805	0.080	0.154	0.068	0.101
WLR	0.135	0.44	0.059	0.112	0.055	0.090
NA(2)	0.046	0.050	0.044	0.050	0.043	0.050
NA(3)	0.042	0.049	0.042	0.049	0.042	0.049

Table 2

Cumulative power by each interim analysis for alternative hypothesis scenarios, using an OBF boundary with equal calendar time increments and accrual time $A = 2$

Scenario	Test	Calendar time				
		2.75	3.5	4.25	5	
1	L_S	0.384	0.645	0.789	0.864	
	$\bar{Q}$	0.291	0.548	0.707	0.805	
	S_p	0.396	0.676	0.818	0.888	
	C	0.223	0.529	0.743	0.851	
	WKM	0.376	0.667	0.800	0.881	
	RMS	0.378	0.669	0.810	0.887	
	L_N	0.611	0.799	0.842	0.864	
	LR	0.702	0.798	0.854	0.892	
	WLR	0.338	0.550	0.704	0.804	
	NA(2)	0.629	0.791	0.808	0.808	
2	NA(3)	0.000	0.653	0.825	0.855	
	L_S	0.004	0.070	0.371	0.741	
	$\bar{Q}$	0.009	0.213	0.686	0.930	
	S_p	0.003	0.045	0.271	0.614	
	C	0.011	0.131	0.444	0.739	
	WKM	0.004	0.040	0.139	0.336	
	RMS	0.004	0.059	0.212	0.459	
	L_N	0.015	0.049	0.118	0.232	
	LR	0.166	0.169	0.195	0.289	
	WLR	0.041	0.387	0.829	0.975	
3	NA(2)	0.020	0.041	0.046	0.047	
	NA(3)	0.000	0.120	0.261	0.297	
	L_S	0.193	0.501	0.794	0.930	
	$\bar{Q}$	0.125	0.400	0.704	0.885	
	S_p	0.179	0.464	0.741	0.891	
	C	0.163	0.504	0.800	0.929	

Scenario	Test	Calendar time				
		2.75	3.5	4.25	5	
4	WKM	0.189	0.453	0.646	0.807	
	RMS	0.190	0.473	0.699	0.848	
	L_N	0.348	0.549	0.647	0.724	
	LR	0.150	0.324	0.506	0.653	
	WLR	0.436	0.778	0.931	0.981	
	NA(2)	0.353	0.494	0.516	0.516	
	NA(3)	0.000	0.572	0.745	0.779	
	L_S	0.028	0.034	0.086	0.311	
	$\bar{Q}$	0.053	0.328	0.740	0.946	
	S_p	0.039	0.051	0.072	0.188	
	C	0.005	0.015	0.090	0.280	
	WKM	0.024	0.044	0.052	0.076	
	RMS	0.027	0.039	0.052	0.097	
	L_N	0.134	0.186	0.188	0.193	
	LR	0.658	0.659	0.659	0.666	
	WLR	0.003	0.097	0.496	0.865	
	NA(2)	0.177	0.293	0.312	0.312	
	NA(3)	0.000	0.017	0.039	0.047	

Table 3

Calendar time at which each test procedure stops to reject the null hypothesis for the example dataset

Test	t_0		
	2	3	4
L_S	11	11	11
Q	7	NR	NR
S_p	NR	11	11
C	11	11	11
WKM	12	11	11
RMS	11	11	11
L_N	NR	13	11
$NA(t_0)$	NR	NR	12

NR means that the test statistic never rejected the null hypothesis