Research Article

Statistics in Medicine

Received 16 October 2013, Accepted 26 October 2014 Published online 12 November 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6366

An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials

Valérie Garès,[a,b,c] Sandrine Andrieu,[a,c] Jean-François Dupuy[d] and Nicolas Savy[a,b*†]

The logrank test is optimal for testing the equality of survival distributions against a proportional hazards alternative. Under a late effects alternative, it is no longer appropriate, and one may turn to Fleming–Harrington's class of weighted logrank tests instead. In some settings, such as in preventive clinical trials where the statistical analysis has to be designed before the trial begins, it can be difficult to choose *a priori* between the logrank and Fleming–Harrington tests. A solution to this issue is provided. A decision rule is constructed for the problem of testing the equality of two survival distributions when the expected alternative may be one of the proportional hazards and late effects. A formula for computing the necessary sample size is obtained for this decision rule. A comprehensive simulation study is conducted to assess finite sample properties of the proposed test statistic. The proposed test improves both the logrank test and Fleming–Harrington's test for late effects. Finally, the methodology is illustrated on a data set in the field of prevention of Alzheimer's disease. **Copyright © 2014 John Wiley & Sons, Ltd.**

Keywords: hypothesis testing; survival data; maximum weighted tests; clinical trials

1. Introduction

Neurodegenerative dementias are a growing public health concern. For example, a recent study estimated the prevalence of Alzheimer's disease at 115.4 million people in 2050 [1]. There is currently no effective treatment for this pathology, which makes its prevention a priority. Prevention is feasible because of the long asymptomatic latent period of the disease. Some studies have shown that delaying Alzheimer's disease onset for a few years could substantially reduce the burden of dementia on society and public healthcare systems [2, 3].

A small number of clinical trials have been conducted to assess prevention treatments for Alzheimer's disease. Their evaluation criterion was a delayed appearance of the event 'develop dementia'. All these trials were analyzed using the logrank test [4] and concluded that the various treatments do not exhibit a significant effect (e.g., [5–8]). The logrank test allows testing of the equality of survival distributions from censored event time data. This test is optimal under the proportional hazards model [9] and thus is likely to be unsuitable for prevention clinical trials. Indeed, preventive treatments may require some exposure before an effect becomes visible. In this setting, we say that late effects occur. Late effects refer to the situation where the hazard (or survival) functions are mostly the same at the beginning of the trial and differ at a later stage. The proportional hazards assumption is unrealistic when late effects occur, and thus, the logrank test is not optimal. To overcome this problem, one idea is to plug a weight function $(W_n(s),\ s \in \mathbb{R}^+)$ (depending on the sample size n) in the logrank statistic. This yields the so-called weighted logrank tests. Specific weights are motivated by the kind of deviation from the null hypothesis

[a] *University of Toulouse III, Toulouse, F-31073, France*
[b] *Toulouse Institute of Mathematics, UMR C5583, Toulouse, F-31062, France*
[c] *INSERM, U1027, Toulouse, F-31073, France*
[d] *INSA of Rennes and IRMAR, Rennes, F-35708, France*
**Correspondence to: Nicolas Savy, Toulouse Institute of Mathematics, University of Toulouse III, UMR C5583, Toulouse, F-31062, France.*
[†] *E-mail: nicolas.savy@math.univ-toulouse.fr*

Copyright © 2014 John Wiley & Sons, Ltd.

(of equality of the survival functions) that one is interested in detecting. A large amount of literature has been devoted to these tests so far and numerous weights have been proposed:

- For detecting early differences between groups, the most widely used tests consider for $W_n(s)$: the number of subjects at risk at s (or its square root, [10, 11]) or Kaplan–Meier estimator $\hat{S}_n$ of the survival function under the null hypothesis [12, 13] or a power of this estimate (yielding the G^p family, see [14]). In [15], the authors adapted the G^p weight to rare events.
- For detecting lag alternatives, constant piecewise and linear piecewise weights were proposed in [16]. For detecting early and/or late effects, Fleming and Harrington [17] proposed the $G^{p,q}$ family based on the weight

$$W_n^{p,q}(s) = \left[\hat{S}_n(s)\right]^p \left[1 - \hat{S}_n(s)\right]^q . \tag{1}$$

  Wu and Gilbert [18] considered a quadratic weight based on $\hat{S}_n$ and an additional real parameter, and Wallenstein and Berger [19] considered a quadratic weight.

The choice of a given weight depends heavily on the postulated alternative hypothesis and misspecifying this alternative may imply a loss of power. Several tests have therefore been proposed to handle simultaneously a range of alternatives:

- linear combinations of weighted statistics were investigated in [20–22] and
- the maximum of several weighted logrank statistics was studied by [18, 22–24].

Several other strategies are available. In [25–27], authors proposed to estimate the weight from the data, and in [28, 29], authors considered the supremum over time of a weighted logrank statistic. To date, the most advanced work on this topic was provided by Kosorok and Lin [30], who considered a general class of function-indexed weighted logrank statistics. This class includes as special cases the supremum and infimum of $G^{p,q}$ statistics over p and/or q (respectively over time and p and/or q) and sums of these supremum and infimum.

Most of the aforementioned tests depend on some parameters that have to be fixed by the investigator, such as p and q. This is a problem in preventive randomized clinical trials where all parameters have to be fixed before the trials begin (and thus before any data-based information has become available to help choosing parameters). The tests mentioned earlier have not been discussed in this setting. Moreover, no formula is available for calculating the necessary sample size for most of them although this is a crucial aspect of the design of clinical trials. A sample size formula has been proposed for the supremum over time of a weighted logrank statistic [29, 31], but the fact that the precise set of alternatives for which this test has good power is currently unknown can make clinicians hesitate to use it. In [30], only intervals of values for p and q (and not the actual values) have to be selected. The problem is moved to specifying bounds for these intervals, which may be considered as a simpler issue. However, authors provide no indication on how to choose these bounds. Moreover, in practice, authors use a discrete approximation of the intervals for p and q, which raises the problem of choosing the resolution of these approximations. Finally, no sample size formula is provided for this class of tests.

In the present paper, we consider the setting of a preventive randomized clinical trial where an investigator wishes to test a treatment effect but has no firm prior knowledge of whether this effect should be of the proportional hazards form or a late effect. In this situation, it is difficult to choose *a priori* between a logrank test and a weighted logrank for late effects. We aim at providing a solution to this dilemma. Precisely, we construct a new statistic for testing the equality of two survival distributions when one suspects that the alternative is one of the proportional hazards or late effects. Our test statistic is constructed as the maximum of the logrank statistic and several Fleming–Harrington's statistics for late effect. This proposal is designed to have good power against both late effects and proportional hazards alternatives. As mentioned earlier, Kosorok and Lin [30] investigated a general class of weighted logrank statistics, which includes supremum (respectively infimum and sums of supremum and infimum) over time and/or p and q of $G^{p,q}$ statistics. However, almost none of these tests are powerful against both proportional hazards and late effects alternatives. From [30], the weighted logrank statistic $G^{0,1}$ and its supremum over time perform well against both alternatives but only if the delayed effect does not occur too late. Hence, these tests are not really appropriate in a genuine late effect setting.

Copyright © 2014 John Wiley & Sons, Ltd.

The test proposed in our paper also depends on some parameters that have to be fixed by the investigator, but we provide some indications on how to choose them. We derive the asymptotic distribution of the suggested test statistic under the null hypothesis of equality of the survival distributions between groups. Under the alternative hypothesis, we derive a weak convergence result, which yields a formula for computing the necessary sample size for the proposed test. We conduct a comprehensive simulation study to assess finite sample properties of the proposed statistic, and we compare its performance with the logrank test, to Fleming–Harrington's test for late effects and to the supremum logrank test. Finally, we illustrate the proposed methodology on the GuidAge study.

GuidAge is a 5-year-long prospective prevention trial involving patients who spontaneously reported memory complaints [32]. The primary objective was to investigate the effect of a treatment called EGb761 on the conversion rate from memory complaints to Alzheimer's disease. The statistical analysis design was specified before the beginning of the trial and required data to be analyzed using the logrank test, which concluded that the treatment is ineffective. A re-analysis using Fleming–Harrington's test for late effect with $p = 0$ and $q = 3$ was conducted and concluded that the treatment is effective. This example illustrates how difficult it can be to choose the best test to use in a setting where late effects are suspected although they cannot be ascertained *a priori*.

The paper is organized as follows. In Section 2, we recall some useful background on weighted logrank tests (definition, asymptotics and asymptotic relative efficiency). In Section 3, we construct our test statistic, we investigate its asymptotic distribution, and we assess its performance via simulations. We also propose a sample size equation for this test and a numerical method for approximating the solution of this equation. In Section 4, we compare our test with the supremum over time logrank test. We analyze the GuidAge study in Section 5. A summary and some perspectives conclude the paper in Section 6.

2. Some background on logrank and weighted logrank tests

In this section, we set some notations, we describe the problem, and we briefly recall some results about weighted logrank tests and their asymptotic relative efficiency. These results will be useful in Section 3 (we refer the reader to [33] for a more detailed treatment of this material).

2.1. Weighted logrank tests: definition and asymptotic distribution

Let T be a non-negative random variable with cumulative distribution function F, survival function $S = 1-F$, hazard function λ and cumulative hazard function $\Lambda(\cdot) = \int_0^{\cdot} \lambda(s)ds$. T denotes the duration between a time origin and the time of occurrence of some event of interest. T is assumed to be right censored: we observe the event only if it occurs before a certain time C. C has distribution function G and is assumed independent of T. Assume that we observe n independent subjects. The ℓ-th subject has latent survival and censoring times T_ℓ and C_ℓ, respectively. The observations consist of the n couples $(X_\ell, \delta_\ell)_{\ell=1\ldots n}$, where $X_\ell = \min(T_\ell, C_\ell)$, $\delta_\ell = \mathbb{I}\{T_\ell \leqslant C_\ell\}$, and $\mathbb{I}$ is the indicator function. For any $t \geqslant 0$, we also define the random variables

$$N_n(t) = \sum_{\ell=1}^{n} \mathbb{I}\left\{X_\ell \leqslant t, \delta_\ell = 1\right\} \text{ and } Y_n(t) = \sum_{\ell=1}^{n} \mathbb{I}\left\{X_\ell \geqslant t\right\}.$$

$N_n(t)$ is the number of events at t, and $Y_n(t)$ is the number of subjects at risk at t^-. Finally, let τ denote the length of the study from the origin, and $\tau' = \inf_{t\geqslant 0}\{(1 - F(t))(1 - G(t)) = 0\}$. We assume that $\tau < \tau'$.

We consider a clinical trial with two arms, where n_T patients receive a drug (or treatment), and n_P patients receive a placebo. In what follows, all the random variables (and related quantities such as the distribution and survival functions) for the treatment (respectively placebo) group are upper-indexed by T (respectively P). With these notations, $n = n_P + n_T$, $N_n = N^P_{n_P} + N^T_{n_T}$ and $Y_n = Y^P_{n_P} + Y^T_{n_T}$.

Let $\{F_\theta : \theta \in \Theta \subset \mathbb{R}\}$ be a family of continuous cumulative distribution functions on $[0, \infty)$ indexed by the parameter $\theta \in \Theta$. The logrank statistic is a classical tool for testing the hypothesis of equality of the survival distributions in the two groups against the alternative of distinct distributions:

$$\mathcal{H}_0 : F^T = F^P = F_{\theta^0} \quad \text{against} \quad \mathcal{H}_1 : \left\{F^T = F_{\theta^T} \quad \text{and} \quad F^P = F_{\theta^P}\right\}. \tag{2}$$

Copyright © 2014 John Wiley & Sons, Ltd.

At time t, the logrank statistic can be written as

$$\mathrm{LR}_n(t) = \int_0^t \left(\frac{n_P + n_T}{n_P n_T}\right)^{1/2} \frac{Y^P_{n_P}(s) Y^T_{n_T}(s)}{Y_n(s)} \left[\frac{dN^P_{n_P}(s)}{Y^P_{n_P}(s)} - \frac{dN^T_{n_T}(s)}{Y^T_{n_T}(s)}\right]. \tag{3}$$

The logrank test is optimal for testing $\mathcal{H}_0$ against a proportional hazards alternative [17]. When early or late differences between groups exist, one turns usually to weighted logrank tests, which are obtained by plugging a weight function in (3). If (W_n) is a sequence of adapted, bounded and non-negative predictable processes, a weighted logrank test is defined as

$$\mathrm{LR}_{W_n}(t) = \int_0^t W_n(s) \left(\frac{n_P + n_T}{n_P n_T}\right)^{1/2} \frac{Y^P_{n_P}(s) Y^T_{n_T}(s)}{Y_n(s)} \left[\frac{dN^P_{n_P}(s)}{Y^P_{n_P}(s)} - \frac{dN^T_{n_T}(s)}{Y^T_{n_T}(s)}\right].$$

The logrank statistic LR_n is a particular case of LR_{W_n}, with $W_n(t) = 1$ for every t. Now, assume that $\frac{n_P}{n} \xrightarrow[n\to\infty]{} \frac{1}{2}$, $\frac{n_T}{n} \xrightarrow[n\to\infty]{} \frac{1}{2}$ and that there exists a function $w \in \mathbb{D}$ satisfying $W_n(s) \xrightarrow[n\to\infty]{a.s.} w(s)$ (with $\mathbb{D}$ the Skohorod space of càdlàg functions). If $t \in \mathbb{R}$, let

$$\pi^P(t) = \lim_{n\to\infty} \frac{Y^P_{n_P}(t)}{n_P},\ \pi^T(t) = \lim_{n\to\infty} \frac{Y^T_{n_T}(t)}{n_T},\ \pi(t) = \lim_{n\to\infty} \frac{Y_n(t)}{n} = \frac{1}{2}(\pi^P(t) + \pi^T(t)) \text{ and } k(t) = w(t)\frac{\pi^P(t)\pi^T(s)}{\pi(t)}.$$

Then under $\mathcal{H}_0$, as $n \to \infty$, the process LR_{W_n} converges weakly to a mean-zero Gaussian process with covariance function

$$\sigma^2 : (t_1, t_2) \to \int_0^{t_1 \wedge t_2} w^2(s) \frac{\pi^P(s)\pi^T(s)}{\pi(s)} d\Lambda_{\theta^0}(s).$$

Under $\mathcal{H}_1$, as $n \to \infty$, $\mathrm{LR}_{W_n} - \sqrt{n}\mu^{\mathbb{G}}_{(\theta^T, \theta^P)}$ converges weakly to a mean-zero Gaussian process with covariance function $(\sigma^{\mathbb{G}}_{(\theta^T,\theta^P)})^2 = (\sigma^P_{\theta^P})^2 + (\sigma^T_{\theta^T})^2$, where

$$\mu^{\mathbb{G}}_{(\theta^T,\theta^P)} : t \to \int_0^t \frac{1}{2} k(s) \left(d\Lambda_{\theta^P}(s) - d\Lambda_{\theta^T}(s)\right) \text{ and } \left(\sigma^k_{\theta^k}\right)^2 : (t_1, t_2) \to \int_0^{t_1 \wedge t_2} \frac{1}{2} \frac{k^2(s)}{\pi^k(s)} d\Lambda_{\theta^k}(s),\ k = T, P.$$

Finally, a consistent estimator of the asymptotic variance of the weighted logrank statistic $\mathrm{LR}_{W_n}(\tau)$ [17] is given by

$$\hat{\sigma}^2_{W_n} = \frac{n}{n_P n_T} \int_0^\tau W_n^2(s) \frac{Y^P_{n_P}(s) Y^T_{n_T}(s)}{Y_n(s)} \frac{dN_n(s)}{Y_n(s)}.$$

Remark 1
In what follows, Fleming–Harrington's test statistic for late effects (whose weight W_n is given by $W_n^{0,q}$) will be denoted by FH^q_n. Its asymptotic variance estimator $\hat{\sigma}^2_{W_n^{0,q}}$ will be denoted by $\hat{\sigma}^2_q$.

2.2. *Asymptotic relative efficiency and application to weighted logrank tests*

The results earlier imply that the asymptotic distribution of $\mathrm{LR}_{W_n}(t)$ under $\mathcal{H}_1$ is degenerate. The weighted logrank tests are thus consistent and cannot be compared in terms of their power (see [17, 34] for a more detailed exposition). In this setting, an appropriate comparison procedure is to consider sequences of alternatives that converge to the null hypothesis as n tends to infinity. This is the idea of Pitman's asymptotic relative efficiency. An appropriate choice for the alternatives $(\theta^P_{n_P})$ and $(\theta^T_{n_T})$ in (2) is

$$\theta^P_{n_P} = \theta^0 + \gamma\sqrt{\frac{n_T}{n_P(n_P + n_T)}},\quad \theta^T_{n_T} = \theta^0 - \gamma\sqrt{\frac{n_P}{n_T(n_P + n_T)}},$$

Copyright © 2014 John Wiley & Sons, Ltd.

where γ is a constant [17]. Moreover in logrank testing, it is useful to consider the so-called 'shift assumptions up to a change of time' for the alternative hypothesis [34]. Shift assumptions are defined through the following family of distribution functions:

$$F_\theta(t) = \Psi\left(g(t) + \theta\right), \qquad \theta \in \Theta,$$

where $g : [0, \infty[\rightarrow]-\infty, u^+]$ (with $u^+ \in \bar{\mathbb{R}}$) is a non-decreasing differentiable function, and Ψ is a continuous cumulative distribution function with a positive density Ψ' and an almost everywhere continuous second derivative Ψ''. In this setting, it is possible to prove that the logrank is the most powerful test against a proportional hazards alternative and to derive the alternative hypothesis for which the weight (1) is optimal within the meaning of Pitman's asymptotic relative efficiency (see [14] for $p > 0$ and $q = 0$ and [33] for $p \geqslant 0$ and $q \geqslant 0$). Precisely, if $p \geqslant 0$ and $q \geqslant 0$, Fleming–Harrington's statistic with weight $W_n^{p,q}$ has maximum efficiency to test the hypothesis

$$\mathcal{H}_0 : F^T = F^P = F_{\theta^0} \quad \text{against} \quad \mathcal{H}_1 : \left\{ F^T = \Psi^{p,q}\left(g + \theta^T_{n_T}\right) \text{ and } F^P = \Psi^{p,q}\left(g + \theta^P_{n_P}\right) \right\} \tag{4}$$

with

$$\Psi^{p,q}(u) = 1 - \left(\mathcal{L}^{p,q}\right)^{-1}(u + \delta), \tag{5}$$

where $u = g(t) + \theta^0$, δ is a constant, $\mathcal{L}^{p,q}$ is the one-to-one primitive of the function defined from $[0, 1]$ to $\mathbb{R}^-$ by

$$x \rightarrow \frac{1}{xL^{p,q}(x)} \qquad \text{with} \quad L^{p,q}(x) = -B_{inc}(x - 1, q + 1, p)$$

and B_{inc} the incomplete beta function $B_{inc}(x, a, b) = \int_0^x s^{a-1}(1 - s)^{b-1} ds$ [33]. The alternative hypothesis $\mathcal{H}_1$ is thus fully described by the function $\Psi^{p,q}$. This result is very important in practice: it enables us to simulate situations where Fleming–Harrington's test for late effects is optimal and to implement comparative studies with alternative tests. Such studies have been conducted in [33], and their results essentially state that Fleming–Harrington's test with $p = 0$ and $q > 0$ has maximum efficiency to test late effects but is not efficient when the true alternative is proportional hazards. Another interesting result is that the outcome of Fleming–Harrington's test for late effects is rather insensitive to the value of q [33]. This is a nice feature for those clinical trials where: (i) one suspects late effects to occur and (ii) the clinical design requires q to be chosen before the trial begins.

But at the design stage of a trial, one is rarely able to decide firmly between a proportional hazards and a late effects alternative. Thus, it is usually difficult to choose *a priori* the best test to use between the logrank and Fleming–Harrington's tests. In the next section, we provide a solution to this dilemma.

3. The proposed test statistic

In this section, we construct a statistic for testing the hypothesis of equality of two survival distributions. This statistic is designed to have good power against both late effects and proportional hazards alternatives. It is constructed as the maximum of the logrank and Fleming–Harrington statistics. In what follows, we investigate the asymptotic distribution of the proposed statistic, and we assess its performance via simulations. We also propose a sample size formula for this test.

3.1. Maximum weighted logrank statistic: definition and asymptotic distribution

We consider the testing problem

$$\begin{cases} \mathcal{H}_0 : F^T = F^P = F, \\ \mathcal{H}_1 : \cup_{i=1}^m \left\{ F^T = \Psi^{q_i}\left(g + \theta^T(i)\right) \text{ and } F^P = \Psi^{q_i}\left(g + \theta^P(i)\right) \right\}, \end{cases} \tag{6}$$

Copyright © 2014 John Wiley & Sons, Ltd.

where $\Psi^q := \Psi^{0,q}$ is defined by (5), and for a given late effect alternative of the type q_i, the shift $\Delta(q_i)$ is given by $\Delta(q_i) := \theta^T(i) - \theta^P(i)$. For every $i = 1, \dots, m$, let p_i be a known probability that reflects the investigator's degree of certainty that a late effect of type q_i occurs, with $\sum_{i=1}^m p_i = 1$ (note that if $q_i = 0$, the i-th alternative is proportional hazards). Let $\vec{q} = (q_1, \dots, q_m) \in \mathbb{N}^m$ be such that $q_i \neq q_j$ for $i \neq j$, and let $t \geqslant 0$. We define the maximum weighted logrank statistic as

$$\mathrm{MLR}_n^{\vec{q}}(t) = \max_{i=1,\dots,m} \left(\left| \frac{\mathrm{FH}_n^{q_i}(t)}{\hat{\sigma}_{q_i}(t)} \right| \right),$$

where $\mathrm{FH}_n^{q_i}$ and $\hat{\sigma}_{q_i}(t)$ are defined in Remark 1. To construct a decision rule, we need the asymptotic distribution of the process $\mathrm{MLR}_n^{\vec{q}} := \{\mathrm{MLR}_n^{\vec{q}}(t), t \geqslant 0\}$ under $\mathcal{H}_0$. Under $\mathcal{H}_0$, as $n \to \infty$, this process converges weakly to $\max_{i=1,\dots,m}(|\tilde{\mathbb{G}}^{q_i}|)$ where $(\tilde{\mathbb{G}}^{q_1}, \dots, \tilde{\mathbb{G}}^{q_m})$ is a m-variate mean-zero Gaussian process with covariance function defined for any $i, j = 1, \dots, m$ by

$$\left(\tilde{\Sigma}_{i,j}^{\mathcal{H}_0}\right)^2 : (t_1, t_2) \to \mathbb{E}\left[\tilde{\mathbb{G}}^{q_i}(t_1)\tilde{\mathbb{G}}^{q_j}(t_2)\right] = \frac{\left(\Sigma_{i,j}^{\mathcal{H}_0}\right)^2 (t_1 \wedge t_2)}{\Sigma_{i,i}^{\mathcal{H}_0}(t_1)\Sigma_{j,j}^{\mathcal{H}_0}(t_2)},$$

$$\text{with } \left(\Sigma_{i,j}^{\mathcal{H}_0}\right)^2 : t \to \int_0^t w^{q_i}(s) w^{q_j}(s) \frac{\pi^P(s)\pi^T(s)}{\pi(s)} d\Lambda_{\theta^0}(s)$$

and $w^q(s) = (1 - S(s))^q$. A decision rule with asymptotic level $\alpha \in (0, 1)$ rejects the null hypothesis $\mathcal{H}_0$ if the statistic $\mathrm{MLR}_n^{\vec{q}}(\tau)$ exceeds the upper α-quantile of the distribution of $\max_{i=1,\dots,m}(|\tilde{\mathbb{G}}^{q_i}(\tau)|)$.

Under $\mathcal{H}_1$, the asymptotic distribution of $\mathrm{MLR}_n^{\vec{q}}$ cannot be easily derived. However, we can establish the following weak convergence result, which is sufficient to derive a sample size computation algorithm (Section 3.3). For $k = T, P$, let

$$\Lambda_{\theta^k} : t \to -\ln\left(\sum_{i=1}^m p_i \left(1 - \Psi^{q_i}\left(g(t) + \theta^k(i)\right)\right)\right)$$

and for $i = 1, \dots, m$, let

$$\mu^{q_i} : t \to \frac{1}{2}\int_0^t k^{q_i}(s)\left(d\Lambda_{\theta^P}(s) - d\Lambda_{\theta^T}(s)\right) \quad \text{with} \quad k^{q_i}(s) = (1 - S(s))^{q_i} \frac{\pi^P(s)\pi^T(s)}{\pi(s)}$$

and

$$\left(\Sigma_{i,j}^{\mathcal{H}_1}\right)^2 : t \to \frac{1}{2}\int_0^t w^{q_i}(s) w^{q_j}(s) \left[\pi^P(s)\left(\frac{\pi^T(s)}{\pi(s)}\right)^2 d\Lambda_{\theta^P}(s) + \pi^T(s)\left(\frac{\pi^P(s)}{\pi(s)}\right)^2 d\Lambda_{\theta^T}(s)\right].$$

Then under $\mathcal{H}_1$, as $n \to \infty$, the m-variate process $(\mathrm{FH}_n^{q_1}/\hat{\sigma}_{q_1}, \dots, \mathrm{FH}_n^{q_m}/\hat{\sigma}_{q_m}) - \sqrt{n}(\mu^{q_1}/\Sigma_{1,1}^{\mathcal{H}_1}, \dots, \mu^{q_m}/\Sigma_{m,m}^{\mathcal{H}_1})$ converges weakly to $(\tilde{\mathbb{G}}'^{q_1}, \dots, \tilde{\mathbb{G}}'^{q_m})$, where $(\tilde{\mathbb{G}}'^{q_1}, \dots, \tilde{\mathbb{G}}'^{q_m})$ is a mean-zero m-variate Gaussian process with covariance function

$$\left(\tilde{\Sigma}_{i,j}^{\mathcal{H}_1}\right)^2 : (t_1, t_2) \to \mathbb{E}\left[\tilde{\mathbb{G}}'^{q_i}(t_1)\tilde{\mathbb{G}}'^{q_j}(t_2)\right] = \frac{\left(\Sigma_{i,j}^{\mathcal{H}_1}\right)^2 (t_1 \wedge t_2)}{\Sigma_{i,i}^{\mathcal{H}_1}(t_1)\Sigma_{j,j}^{\mathcal{H}_1}(t_2)}, \quad i, j = 1, \dots, m. \tag{7}$$

When $m = 1$, the problem boils down to proving the convergence of Fleming–Harrington's statistic (see [17], see also [33] for a recent new proof). When $m > 1$, similar arguments allow to prove that under $\mathcal{H}_0$, $(\mathrm{FH}_n^{q_1}/\hat{\sigma}_{q_1}, \dots, \mathrm{FH}_n^{q_m}/\hat{\sigma}_{q_m}) \xrightarrow{\mathcal{L}(\mathbb{D}^m)} (\tilde{\mathbb{G}}^{q_1}, \dots, \tilde{\mathbb{G}}^{q_m})$ where $(\tilde{\mathbb{G}}^{q_1}, \dots, \tilde{\mathbb{G}}^{q_m})$ is as aforementioned. The function $(F_1, \dots, F_m) \to \max_{i=1,\dots m}(|F_i|)$ is continuous from $\mathbb{D}^m$ to $\mathbb{D}$, and the convergence of $\mathrm{MLR}_n^{\vec{q}}$ follows.

Under $\mathcal{H}_1$, similar arguments allow to prove that $(\mathrm{FH}_n^{q_1}/\hat{\sigma}_{q_1}, \dots, \mathrm{FH}_n^{q_m}/\hat{\sigma}_{q_m}) - \sqrt{n}(\mu^{q_1}/\Sigma_{1,1}^{\mathcal{H}_1}, \dots, \mu^{q_m}/\Sigma_{m,m}^{\mathcal{H}_1}) \xrightarrow{\mathcal{L}(\mathbb{D}^m)} (\tilde{\mathbb{G}}'^{q_1}, \dots, \tilde{\mathbb{G}}'^{q_m})$ where $(\tilde{\mathbb{G}}'^{q_1}, \dots, \tilde{\mathbb{G}}'^{q_m})$ is defined earlier. Moreover under $\mathcal{H}_1$, it follows from the Bayes formula that the distribution of T in the group k ($k = T, P$) is given by

Copyright © 2014 John Wiley & Sons, Ltd.

$$\mathbb{P}_{\mathcal{H}_1}\{T<t\}=\sum_{i=1}^{m}\mathbb{P}_{\mathcal{H}_1}\left\{T<t\mid\Omega_i\right\}\ \mathbb{P}\left\{\Omega_i\right\}=\sum_{i=1}^{m}p_i\Psi^{q_i}\left(g(t)+\theta^k(i)\right),$$

where Ω_i is the event 'a late effects of type q_i occurs'. It follows that $\Lambda_{\theta^k}(t)$ is expressed as $-\ln\left(\sum_{i=1}^{m}p_i(1-\Psi^{q_i}(g(t)+\theta^k(i)))\right)$, $k=T,P$.

It is important to note that $\mathrm{MLR}_n^{\vec{q}}$ is not necessarily the optimal statistic for testing the hypothesis (6).

Remark 2

In [33], the authors show that the result of Fleming–Harrington's test is quite insensitive to the value of q provided $q>0$. Thus, one can restrict to $\vec{q}=(0,q)$ (where q can be chosen according to the guidelines given in [33]), and in what follows, we will consider the test $\mathrm{MLR}_n^{\vec{q}}(\tau)$ with $\vec{q}=(0,q)$ (for notational simplicity, we will denote this test by MLR^q).

Remark 3

When $\vec{q}=(0,q)$, the probability p_1 reflects the investigator's degree of certainty that if a difference between groups exist, this difference is of the proportional hazards form. p_1 has to be calibrated based on clinicians *a priori* judgment. Alternatively, one may set $p_1=0.5$, which does not favor any of the proportional hazards or late effects alternatives.

3.2. A simulation study

In this section, we assess properties of the test statistic MLR^q via simulations. First, we assess its level and power against proportional hazards and late effects alternatives. Then, we investigate the sensitivity of MLR^q to q. The power of MLR^q against a crossing hazards alternative is assessed in Section 4.2.

Remark 4

In a simulation study described in [30], Kosorok and Lin investigated the behavior of several function-indexed weighted logrank statistics against alternatives for which the optimal weight function is piecewise constant. Here, we work under alternatives for which the logrank and Fleming–Harrington tests are respectively optimal. Thus, the proposed MLR^q is investigated under conditions that are the most favorable for its direct competitors.

Simulation design. For illustrative purpose, we simulate the data in the placebo group from an exponential distribution with parameter a, where a is chosen to yield a censoring proportion $c:=S^P(\tau)$ at τ. We have

$$a=-\frac{\ln\left(S^P(\tau)\right)}{\tau}. \tag{8}$$

We define the magnitude of discrepancy between survival distributions in the treatment and placebo groups through the rate

$$r=\frac{S^T(\tau)-S^P(\tau)}{1-S^P(\tau)}, \tag{9}$$

which is usually fixed by the investigator, in preventive clinical trials. For simulating the data in the treatment group, we consider successively a proportional hazards alternative (case 1) and a late effects alternative (case 2):

Case 1: In this case, the logrank test (or equivalently Fleming–Harrington's test with $p=q=0$) is optimal. A relevant choice for g in (4) yields the following hazard function in the treatment group:

$$\lambda^T(t)=ae^{\Delta(0)}. \tag{10}$$

Case 2: In this case, Fleming–Harrington's test with $p=0$ and $q>0$ is optimal. A relevant choice for g in (4) yields the following hazard function in the treatment group:

$$\lambda^T(t)=a\,\frac{L^{0,q}\left(\left(\mathcal{L}^{0,q}\right)^{-1}\left(\mathcal{L}^{0,q}(e^{-at})+\Delta(q)\right)\right)}{L^{0,q}\left(e^{-at}\right)}. \tag{11}$$

Copyright © 2014 John Wiley & Sons, Ltd.

The shifts $\Delta(0)$ and $\Delta(q)$ in (10) and (11) can be obtained from (8) and (9), as

$$\Delta(i) = \mathcal{L}^{0,i}\left(r\left(1 - S^P(\tau)\right) + S^P(\tau)\right) - \mathcal{L}^{0,i}\left(S^P(\tau)\right), \quad i = 0, q.$$

When $q = 0$, (11) reduces to (10). Therefore in case 1, we simulate the data in the treatment group by using (11) with $q = 0$. In case 2, we use $q = 3$ (in what follows, we denote by q_S the value of q used to simulate the data). In both cases 1 and 2, we consider several simulation scenarios obtained by combining various censoring proportions ($c = 0.2, 0.5, 0.8$), discrepancy rates ($r = 0.05, 0.1, 0.2, 0.3$) and sample sizes ($n = 100, 500, 1000, 2000$ with $n_P = n_T = n/2$). Two thousand data sets are simulated for each combination of c, r and n. The logrank test FH^0 (LR thereafter), Fleming–Harrington's test for late effect FH^3 and the proposed test MLR^3 are applied to the resulting data (the nominal level is set to 0.05), and their respective empirical powers over the 2000 data sets are obtained. The hazard and survival functions used in this simulation study are plotted on Figure 1. For illustrative purpose, we also plot the hazard and survival functions in the treatment group under the alternative hypothesis for which Fleming–Harrington's test with $q = 1, 2, 4$ is optimal. As expected, we note that the larger q is, the later the treatment effect can be detected.

To investigate the level of MLR^q, we let the placebo and treatment groups share the same exponential distribution with parameter a. We consider the same combinations of c and n as aforementioned, and we simulate 2000 data sets. For each of them, we obtain the outcome of MLR^q for q ranging from 1 to 10 (the nominal level is set to 0.05). From this, we obtain the empirical levels of $\text{MLR}^1, \ldots, \text{MLR}^{10}$ over the 2000 data sets.

Using MLR^q requires choosing q, and one may wonder whether the test is sensitive to this value. To elucidate this issue, we generate 2000 data sets for each $q_S = 0, \ldots, 10$, where data in the placebo group are simulated from the exponential distribution with parameter a, and data in the treatment group are simulated from (11) with $q = q_S$. For each q_S, we obtain the empirical power of the logrank test, of Fleming–Harrington's tests $\text{FH}^1, \ldots, \text{FH}^{10}$ and of $\text{MLR}^1, \ldots, \text{MLR}^{10}$. For this simulation study, we consider $c = 0.8$, $n = 2000$ and $r = 0.2$.

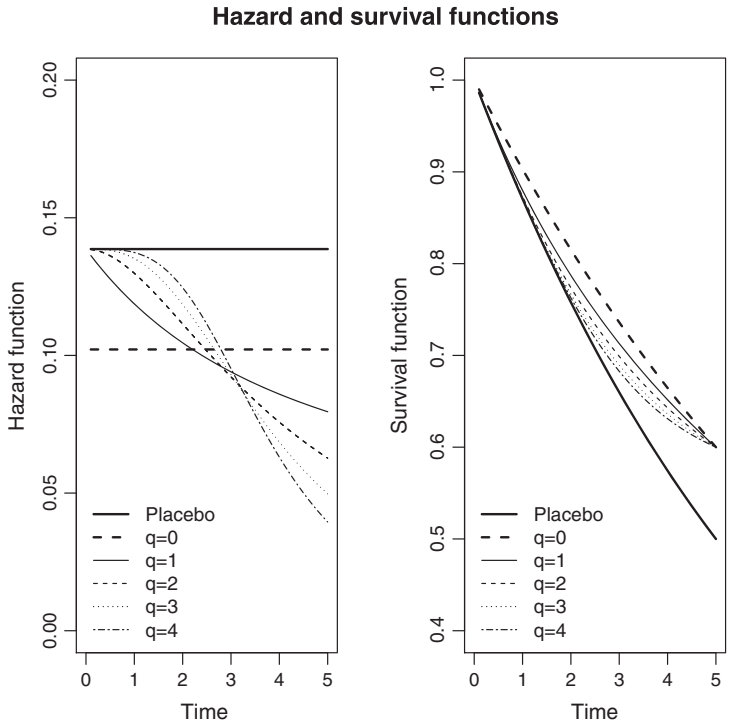

Figure 1. Hazard and survival functions for $q \geqslant 0$. The curves $q = 0, 1, 2, 3, 4$ correspond to the hazard and survival functions in the treatment group in settings where Fleming–Harrington's test FH^q is optimal ($c = 0.5$ and $r = 0.2$).

Copyright © 2014 John Wiley & Sons, Ltd.

Results. The results for the empirical level and power are given in Tables I and II, respectively. From Table I, we observe that the proposed MLR^q respects the nominal level. From Table II, the power of the tests LR, FH^3 and MLR^3 increases with n and r and decreases when censoring increases. We also verify that the logrank test (respectively Fleming–Harrington's test FH^3) has maximum power in case 1 (respectively case 2) but performs moderately well when late effects (respectively proportional hazards)

Table I. Empirical level of MLR^q for $q = 1, \ldots, 10$.

n	c	MLR^1	MLR^2	MLR^3	MLR^4	MLR^5	MLR^6	MLR^7	MLR^8	MLR^9	MLR^{10}
100	0.2	0.043	0.046	0.047	0.049	0.052	0.051	0.056	0.050	0.052	0.052
	0.5	0.055	0.054	0.051	0.052	0.051	0.049	0.049	0.051	0.052	0.053
	0.8	0.053	0.056	0.057	0.055	0.050	0.054	0.052	0.053	0.055	0.053
500	0.2	0.041	0.042	0.046	0.045	0.046	0.046	0.045	0.044	0.046	0.042
	0.5	0.058	0.056	0.053	0.054	0.054	0.054	0.055	0.053	0.051	0.051
	0.8	0.047	0.047	0.049	0.047	0.046	0.044	0.044	0.045	0.045	0.045
1000	0.2	0.048	0.044	0.044	0.045	0.048	0.050	0.050	0.051	0.050	0.051
	0.5	0.049	0.048	0.047	0.049	0.046	0.043	0.043	0.044	0.044	0.044
	0.8	0.052	0.048	0.048	0.048	0.049	0.047	0.047	0.051	0.051	0.050
2000	0.2	0.041	0.043	0.043	0.043	0.043	0.043	0.043	0.044	0.044	0.042
	0.5	0.054	0.057	0.058	0.058	0.055	0.056	0.058	0.055	0.054	0.054
	0.8	0.047	0.052	0.056	0.053	0.056	0.056	0.052	0.052	0.052	0.052

c: censoring proportion; n: sample size; MLR: maximum weighted logrank test

Table II. Empirical power (EP) and its relative variation (RV) for the logrank (LR) test, Fleming–Harrington's test FH^3, maximum weighted logrank test MLR^3 and supremum logrank test (SLR).

			$q_S = 0$								$q_S = 3$							
			LR		FH^3		MLR^3		SLR		LR		FH^3		MLR^3		SLR	
n	c	r	EP	RV	EP	RV	EP	RV	EP	RV	EP	RV	EP	RV	EP	RV	EP	RV
500	0.2	0.1	**0.626**	—	0.330	0.47	0.571	0.09	0.594	**0.05**	0.353	0.46	**0.656**	—	0.602	**0.08**	0.275	0.58
		0.2	**0.990**	—	0.803	0.19	0.985	**0.01**	0.987	—	0.849	0.15	**0.996**	—	0.992	—	0.780	0.22
		0.3	**1.000**	—	0.983	0.02	1.000	—	1.000	—	0.996	—	**1.000**	—	1.000	—	0.990	0.01
	0.5	0.1	**0.202**	—	0.121	0.40	0.176	0.13	0.185	**0.08**	0.167	0.45	**0.301**	—	0.261	**0.13**	0.133	0.56
		0.2	**0.626**	—	0.348	0.44	0.571	0.09	0.584	**0.07**	0.497	0.39	**0.817**	—	0.767	**0.06**	0.391	0.52
		0.3	**0.925**	—	0.630	0.32	0.899	0.03	0.911	**0.02**	0.845	0.15	**0.992**	—	0.987	**0.01**	0.776	0.22
	0.8	0.1	**0.097**	—	0.077	0.21	0.094	**0.03**	0.082	0.15	0.095	0.21	**0.121**	—	0.111	**0.08**	0.075	0.38
		0.2	**0.214**	—	0.131	0.39	0.205	**0.04**	0.185	0.14	0.178	0.54	**0.385**	—	0.339	**0.12**	0.139	0.64
		0.3	**0.439**	—	0.244	0.44	0.395	**0.10**	0.389	0.11	0.398	0.45	**0.724**	—	0.660	**0.09**	0.297	0.59
1000	0.2	0.1	**0.906**	—	0.590	0.35	0.870	0.04	0.888	**0.02**	0.599	0.35	**0.916**	—	0.887	**0.03**	0.500	0.45
		0.2	**1.000**	—	0.981	0.02	1.000	—	1.000	—	0.989	0.01	**1.000**	—	1.000	—	0.980	0.02
		0.3	**1.000**	—	1.000	—	1.000	—	1.000	—	1.000	-	**1.000**	—	1.000	-	1.000	—
	0.5	0.1	**0.363**	—	0.194	0.47	0.324	0.11	0.339	**0.07**	0.268	0.48	**0.512**	—	0.463	**0.10**	0.212	0.59
		0.2	**0.909**	—	0.592	0.35	0.873	0.04	0.892	**0.02**	0.763	0.22	**0.983**	—	0.980	—	0.690	0.30
		0.3	**0.999**	—	0.908	0.09	0.997	—	0.998	—	0.987	0.01	**1.000**	—	1.000	—	0.974	0.03
	0.8	0.1	**0.122**	—	0.086	0.30	0.111	0.09	0.113	**0.07**	0.117	0.46	**0.216**	—	0.187	**0.13**	0.094	0.56
		0.2	**0.391**	—	0.231	0.41	0.355	0.09	0.353	**0.10**	0.351	0.48	**0.671**	—	0.607	**0.10**	0.263	0.61
		0.3	**0.701**	—	0.372	0.47	0.647	0.08	0.654	**0.07**	0.675	**0.30**	**0.959**	—	0.658	0.31	0.586	0.39
2000	0.2	0.1	**0.995**	—	0.869	0.13	0.995	—	0.995	—	0.883	**0.12**	**0.998**	—	0.996	—	0.824	0.17
		0.2	**1.000**	—	1.000	—	1.000	—	1.000	—	1.000	—	**1.000**	—	1.000	—	1.000	—
		0.3	**1.000**	—	1.000	—	1.000	—	1.000	—	1.000	—	**1.000**	—	1.000	—	1.000	—
	0.5	0.1	**0.621**	—	0.339	0.45	0.569	0.08	0.589	**0.05**	0.468	0.42	**0.806**	—	0.753	**0.07**	0.385	0.52
		0.2	**0.993**	—	0.855	0.14	0.991	—	0.911	**0.08**	0.964	0.04	**1.000**	—	1.000	—	0.946	0.05
		0.3	**1.000**	—	0.998	—	1.000	—	1.000	—	1.000	—	**1.000**	—	1.000	—	1.000	—
	0.8	0.1	**0.217**	—	0.132	0.39	0.191	0.12	0.193	**0.11**	0.195	0.50	**0.391**	—	0.338	**0.14**	0.149	0.62
		0.2	**0.646**	—	0.360	0.44	0.592	0.08	0.606	**0.06**	0.595	0.35	**0.910**	—	0.885	**0.03**	0.509	0.44
		0.3	**0.945**	—	0.691	0.27	0.930	0.02	0.933	**0.01**	0.925	**0.07**	**0.999**	—	0.999	—	0.881	0.12

The data are simulated according to the procedure described in Section 3.2 with $q_S = 0$ and $q_S = 3$.
c: censoring proportion; n: sample size; r: discrepancy rate (9).

Copyright © 2014 John Wiley & Sons, Ltd.

are present. In contrast, the proposed maximum weighted logrank test MLR^3 performs well in both cases. Its power is close to the maximum power whatever the true alternative is. To see this, we calculate, for each test, the relative variation (RV) of its empirical power p with respect to the maximum achieved power p_{max}. RV is defined as

$$RV = \frac{|p - p_{max}|}{p_{max}}.$$

One clearly observes that the RV of the power of the maximum weighted logrank test is small in every simulation scenario. Moreover, this RV is rather stable with respect to c, n and r, which is not the case for Fleming–Harrington's test (case 1) and the logrank test (case 2).

These findings suggest that the proposed maximum weighted logrank test is an appealing compromise between the logrank and Fleming–Harrington tests when one wishes to test the equality of survival distributions without assuming *a priori* whether the true alternative is proportional hazards or late effects.

The results of the sensitivity study are given in Table III. In [33], the authors show that the main issue with Fleming–Harrington's test is to choose between a logrank test ($q = 0$) and a 'genuine' Fleming–Harrington's test with $q > 1$. In other words, the main issue is to decide *a priori* whether proportional hazards or late effects occur. Indeed, the authors observe that if the data are simulated under some $q_S > 1$, the power of Fleming–Harrington's test is not very sensitive to a variation of q in the weight $[1 - \hat{S}(s)]^q$ provided that $q > 1$. On the other hand, if the data are simulated under the proportional hazards assumption (i.e., with $q_S = 0$), the power of the logrank test is maximum, while the power of Fleming–Harrington's test decreases drastically as q increases in $[1 - \hat{S}(s)]^q$. From Table III, the proposed test MLR^q again provides a good compromise between the logrank and Fleming–Harrington's test: its power is both close to the maximum achieved power and relatively insensitive to q. In [33], the authors conclude that in most cases, 3 is an appropriate value for q in Fleming–Harrington's test. Similarly, we observe from Table III that using MLR^3 ensures a good power whatever the true alternative is (including the proportional hazards). Moreover, the test MLR^3 is more stable (in terms of power) than FH^3.

Table III. Sensitivity to q of FH^q and MLR^q.

q_S	LR	FH^1	FH^2	FH^3	FH^4	FH^5	FH^6	FH^7	FH^8	FH^9	FH^{10}
0	**0.629**	0.526	0.416	0.334	5 0.289	0.256	0.229	0.211	0.192	0.182	0.167
1	0.625	**0.756**	0.744	0.702	0.655	0.611	0.580	0.545	0.513	0.485	0.456
2	0.609	0.839	**0.864**	0.863	0.850	0.835	0.812	0.781	0.754	0.729	0.700
3	0.623	0.869	0.919	**0.925**	0.922	0.910	0.901	0.890	0.87	0.864	0.852
4	0.626	0.891	0.943	0.959	0.961	**0.963**	0.961	0.960	0.953	0.948	0.939
5	0.608	0.911	0.963	0.976	0.978	**0.982**	0.980	0.979	0.976	0.975	0.973
6	0.597	0.916	0.964	0.982	0.987	0.989	**0.991**	**0.991**	0.990	**0.991**	0.990
7	0.562	0.910	0.962	0.978	0.986	**0.987**	**0.987**	0.986	0.986	0.985	0.983
8	0.564	0.909	0.970	0.988	0.993	0.995	**0.996**	**0.996**	**0.996**	0.995	0.993
9	0.570	0.911	0.972	0.988	0.993	0.996	0.996	0.998	**0.998**	**0.998**	**0.998**
10	0.571	0.912	0.971	0.987	0.995	0.998	**0.999**	**0.999**	0.998	0.998	0.998
q_S		MLR^1	MLR^2	MLR^3	MLR^4	MLR^5	MLR^6	MLR^7	MLR^8	MLR^9	MLR^{10}
0		**0.620**	0.606	0.589	0.584	0.582	0.579	0.571	0.571	0.570	0.563
1		0.729	**0.731**	0.720	0.692	0.679	0.671	0.657	0.650	0.642	0.635
2		0.797	**0.828**	0.826	0.816	0.801	0.784	0.773	0.758	0.747	0.741
3		0.833	0.881	**0.897**	0.896	0.888	0.880	0.875	0.859	0.848	0.841
4		0.864	0.923	0.936	**0.946**	0.945	0.943	0.940	0.934	0.931	0.923
5		0.880	0.947	0.959	0.967	0.968	**0.970**	0.967	0.968	0.965	0.960
6		0.887	0.950	0.971	0.978	0.983	**0.984**	0.983	0.983	0.982	0.982
7		0.874	0.946	0.967	0.974	0.977	0.978	**0.979**	0.978	0.974	0.973
8		0.881	0.954	0.977	0.985	0.989	0.988	0.989	**0.990**	**0.990**	0.989
9		0.887	0.958	0.977	0.986	0.990	**0.994**	0.993	0.993	0.992	0.992
10		0.880	0.956	0.981	0.989	0.993	0.994	**0.996**	**0.996**	**0.996**	**0.996**

On each line, the data are generated using the procedure described in Section 3.2, by using the value q_S ($q_S = 1, \ldots, 10$). The empirical power calculations for LR, FH^q and MLR^q ($q = 1, \ldots, 10$) are based on 2000 samples (with $n = 2000$, $c = 0.8$ and $r = 0.2$).

LR: logrank test; MLR: maximum weighted logrank test; FH: Fleming–Harrington's test.

Copyright © 2014 John Wiley & Sons, Ltd.

Overall, the proposed MLR^3 seems to provide an appealing alternative to both logrank and Fleming–Harrington's tests.

3.3. Sample size calculation

Several sample size formulas have been proposed for weighted logrank tests [29, 35–37]. In this section, we derive a sample size formula for the test MLR^q. This formula is implicit; hence, we describe a numerical algorithm for evaluating the necessary sample size.

Before stating our result, we introduce some further notations. Let (U, V) be a mean-zero bivariate Gaussian vector with variance–covariance matrix $(\tilde{\Sigma}^{\mathcal{H}_1})^2(\tau,\tau)$ given by (7) with $m = 2$ and $(q_1, q_2) = (0, q)$. Let z_α be the upper α-quantile of $\max(|\tilde{\mathbb{G}}^0(\tau)|, |\tilde{\mathbb{G}}^q(\tau)|)$, where the process $(\tilde{\mathbb{G}}^0, \tilde{\mathbb{G}}^q)$ is defined in Section 3.1

$$\alpha = \mathbb{P}_{\mathcal{H}_0}\left\{\max\left(|\tilde{G}^0(\tau)|, |\tilde{G}^q(\tau)|\right) > z_\alpha\right\} = 1 - \mathbb{P}_{\mathcal{H}_0}\left\{\left(\tilde{G}^0(\tau), \tilde{G}^q(\tau)\right) \in \left[-z_\alpha, z_\alpha\right] \times \left[-z_\alpha, z_\alpha\right]\right\}.$$

Under $\mathcal{H}_0$, for a large n, MLR^q is approximately distributed as $\max(|\tilde{\mathbb{G}}^0(\tau)|, |\tilde{\mathbb{G}}^q(\tau)|)$ where $(\tilde{\mathbb{G}}^0(\tau), \tilde{\mathbb{G}}^q(\tau))$ is a Gaussian vector with mean 0 and covariance matrix $(\tilde{\Sigma}^{\mathcal{H}_0})^2(\tau,\tau) := \mathbb{E}[\tilde{\mathbb{G}}^0(\tau)\tilde{\mathbb{G}}^q(\tau)]$. Next,

$$\begin{aligned}
1 - \beta &= \mathbb{P}_{\mathcal{H}_1}\left\{\max\left(\left|\frac{\mathrm{FH}_n^0(\tau)}{\hat{\sigma}_0(\tau)}\right|, \left|\frac{\mathrm{FH}_n^q(\tau)}{\hat{\sigma}_q(\tau)}\right|\right) > z_\alpha\right\} \\
&= 1 - \mathbb{P}_{\mathcal{H}_1}\left\{\left(\frac{\mathrm{FH}_n^0(\tau)}{\hat{\sigma}_0(\tau)}, \frac{\mathrm{FH}_n^q(\tau)}{\hat{\sigma}_q(\tau)}\right) \in \left[-z_\alpha, z_\alpha\right] \times \left[-z_\alpha, z_\alpha\right]\right\} \\
&= 1 - \mathbb{P}_{\mathcal{H}_1}\left\{\left(\frac{\mathrm{FH}_n^0(\tau)}{\hat{\sigma}_0(\tau)} - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}, \frac{\mathrm{FH}_n^q(\tau)}{\hat{\sigma}_q(\tau)} - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}\right) \in E^{0,q}_{\alpha,n}\right\},
\end{aligned}$$

which can be approximated, if n is large enough, by $1 - \mathbb{P}_{\mathcal{H}_1}\{(U, V) \in E^{0,q}_{\alpha,n}\}$ where (U, V) and $E^{0,q}_{\alpha,n}$ are as described earlier. The unicity of the solution comes from the fact that $1 - \beta$ decreases when n decreases. Let r and $S^P(\tau)$ be given. Suppose that we wish to test the hypothesis (6) (with $m = 2$ and $(q_1, q_2) = (0, q)$) by using the test MLR^q with a type-I risk α and a type-II risk β. Then the necessary sample size n is the unique solution of the equation

$$\mathbb{P}_{\mathcal{H}_1}\left\{(U, V) \in E^{0,q}_{\alpha,n}\right\} = \beta \text{ where } E^{0,q}_{\alpha,n} := \left[-z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}, z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}\right] \times \left[-z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}, z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}\right]. \tag{12}$$

Equation 12 provides only an implicit formula for the necessary sample size n. Therefore, we propose to approximate n as follows. From the result earlier, we have that

$$\begin{aligned}
\beta &= \mathbb{P}_{\mathcal{H}_1}\left\{(U, V) \in E^{0,q}_{\alpha,n}\right\} \\
&= \int_{-z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}}^{z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}} \int_{-z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}}^{z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}} f_{(U,V)}(u, v)dudv \\
&= \mathbb{E}\left[\mathbb{I}\left\{U \in \left[-z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}, z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}\right]\right\}\mathbb{I}\left\{V \in \left[-z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}, z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}\right]\right\}\right],
\end{aligned}$$

Copyright © 2014 John Wiley & Sons, Ltd.

which can be approximated by

$$f(n) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}\left\{U_i \in \left[-z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}, z_\alpha - \sqrt{n}\frac{\mu^0}{\Sigma^{\mathcal{H}_1}_{1,1}}\right]\right\} \mathbb{I}\left\{V_i \in \left[-z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}, z_\alpha - \sqrt{n}\frac{\mu^q}{\Sigma^{\mathcal{H}_1}_{2,2}}\right]\right\}$$

if M is large and (U_i, V_i) $(i = 1, \ldots, M)$ are independent copies of (U, V). Thus, we propose to simulate M couples (U_i, V_i) from (U, V) and to take n such that $f(n) < \beta$ and $f(n-1) > \beta$.

The proposed methodology is now evaluated in a numerical study. First, we obtain the necessary sample size for testing the hypothesis (6) with MLR^q under various settings defined by the censoring proportion c ($c = 0.2, 0.5, 0.8$), the discrepancy rate r ($r = 0.1, 0.2, 0.3$) and the probability p_1 ($p_1 = 0.2, 0.5, 0.8$) (p_1 reflects the investigator's degree of certainty that proportional hazards occur). Table IV provides the results for $q = 1, \ldots, 5$, with $\alpha = 0.05$ and $\beta = 0.2$ (the results are obtained for $M = 10^6$). Then, we compare the sample sizes required by the logrank, Fleming–Harrington and maximum weighted logrank tests in a typical setting of a prevention trial, that is, we consider $c = 0.8$ and $r = 0.2$ with $\alpha = 0.05$ and $\beta = 0.2$. p_1 is set equal to 0.5, reflecting the fact that no preference is given to the proportional hazards or late effects alternative. The results are given in Table V(a). In Table V(b), we investigate the sensitivity of the necessary sample size to the probability p_1. We report the results for the maximum weighted logrank test MLR^3 when $c = 0.8$ and $r = 0.2$.

As expected, the necessary sample size for MLR^q increases when the censoring proportion increases and when the discrepancy rate r decreases (smaller late effects require more patients to be detected). We also observe that the necessary sample size for MLR^q does the following: (i) decreases as q increases when p_1 is small and increases with q when p_1 is large and (ii) increases when p_1 increases (as the suspicion of proportional hazards increases, one needs more patients to decide whether proportional hazards or late effects occur). Finally, the necessary sample size for MLR^q is as follows: (i) is larger than for FH^q but the difference stays moderate and (ii) stays close to the necessary sample size for the logrank test.

4. A comparison with the supremum over time logrank test

The supremum over time logrank test (SLR_n thereafter) was proposed as an alternative to weighted logrank tests to detect late effects [28, 29]. In this section, we compare SLR_n and MLR^q via simulations. Precisely, we compare the level, power (against a proportional hazards alternative, a late effects alternative and a crossing hazards alternative) and necessary sample size of the tests.

4.1. A numerical study

The supremum logrank statistic is defined as

$$\text{SLR}_n = \sup_{t\in[0,\tau[}\left(\frac{\text{FH}^0_n(t)}{\hat{\sigma}_0(t)}\right).$$

We consider the same simulation scenarios as in Section 3.2. These scenarios are obtained by combining various censoring proportions ($c = 0.2, 0.5, 0.8$), discrepancy rates ($r = 0.05, 0.1, 0.2, 0.3$) and sample sizes ($n = 100, 500, 1000, 2000$ with $n_P = n_T = n/2$) for both case 1 (proportional hazards alternative) and case 2 (late effects alternative) (the crossing hazards alternative is described in Section 4.2). 2000 data sets are simulated for each scenario. The empirical level and power of SLR_n are added to Table II (the nominal level is set to 0.05).

As expected, the power of SLR_n increases with n and r and decreases when censoring increases. Under a proportional hazards alternative, the RVs of the power of the supremum logrank and maximum weighted logrank tests are similar. Under late effects, the supremum logrank test performs badly compared with the maximum weighted logrank test: in particular, the RV of the power is far larger for the supremum logrank than for the maximum weighted logrank. Finally, the power of SLR_n decreases as q_S increases, that is, as late effects occur later (based on 2000 simulated samples of size $n = 2000$, with $c = 0.8$ and $r = 0.2$, we obtained the following values for the empirical power of SLR_n: 0.603, 0.557, 0.499, 0.512, 0.504, 0.493 when $q_S = 0, 1, 2, 3, 4, 5$, respectively). From these results, the maximum weighted logrank test appears again to offer the best compromise to detect both proportional hazards and late effects.

Copyright © 2014 John Wiley & Sons, Ltd.

Table IV. Sample size calculation for MLR^q (with $\alpha = 0.05$ and $\beta = 0.2$).

				$p_1 = 0.2$		
c	r	MLR^1	MLR^2	MLR^3	MLR^4	MLR^5
0.2	0.1	919	936	925	896	868
	0.2	264	269	267	262	258
	0.3	133	136	137	136	136
0.5	0.1	30,079	2822	2565	2352	2163
	0.2	767	706	650	601	565
	0.3	342	318	297	281	271
0.8	0.1	10,937	9396	8181	7241	6517
	0.2	2624	2269	1996	1797	1644
	0.3	1121	984	881	811	763
				$p_1 = 0.5$		
0.2	0.1	903	959	990	1007	1014
	0.2	261	276	285	291	296
	0.3	132	139	144	149	152
0.5	0.1	3216	3203	3148	3073	2992
	0.2	805	800	791	778	770
	0.3	358	359	357	357	357
0.8	0.1	11,919	11,336	10,765	10,251	9762
	0.2	2852	2734	2622	2517	2439
	0.3	1222	1173	1137	1113	1195
				$p_1 = 0.8$		
0.2	0.1	867	920	954	977	992
	0.2	252	267	276	282	287
	0.3	128	135	139	143	145
0.5	0.1	3290	3407	3480	3519	3554
	0.2	824	854	873	883	894
	0.3	368	382	391	395	400
0.8	0.1	12,660	12,947	13,006	13,054	13,068
	0.2	3042	3107	3141	3158	3167
	0.3	1300	1327	1343	1357	1367

MLR, maximum weighted logrank test.

Table V. Necessary sample size (NSS) calculation (with $c = 0.8$, $r = 0.2$, $\alpha = 0.05$, $\beta = 0.2$).

V(a):

Test	LR	FH^1	FH^2	FH^3	FH^4	FH^5	MLR^1	MLR^2	MLR^3	MLR^4	MLR^5
NSS	2856	2332	1806	1474	1253	1099	2852	2734	2622	2517	2439

V(b):

p_1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
NSS for MLR^3	1645	1816	1996	2202	2404	2622	2821	2993	3141	3244	3314

In Table V(a): $p_1 = 0.5$.

MLR: maximum weighted logrank test; LR: logrank test; FH: Fleming–Harrington's test.

We also compare the supremum logrank and maximum weighted logrank tests in terms of necessary sample size. The sample size calculation for SLR_n can be found in [29]. We calculate the necessary sample size for testing a differential effects alternative having a discrepancy rate r ($r = 0.1, 0.2, 0.3$) at the end of the study, under a censoring proportion c ($c = 0.2, 0.5, 0.8$). Table VI provides the results (for $\alpha = 0.05$ and $\beta = 0.2$). As expected, the necessary sample size for SLR increases with the censoring proportion and decreases when the rate r increases. Moreover, the necessary sample size for SLR_n and the proposed maximum weighted logrank test are of the same order of magnitude.

Copyright © 2014 John Wiley & Sons, Ltd.

Table VI. Sample size calculation for LR, SLR and MLR^3.

c	r	LR	SLR	MLR^3
0.2	0.1	714	755	990
	0.2	191	201	285
	0.3	87	92	144
0.5	0.1	2870	3035	3148
	0.2	675	713	791
	0.3	278	294	357
0.8	0.1	11,415	12,069	10,765
	0.2	2579	2727	2622
	0.3	1024	1082	1137

LR: logrank test; MLR: mean weighted logrank test; SLR: supremum over time logrank test.

4.2. A crossing hazards alternative

The logrank and Fleming–Harrington's test for late effects are appropriate for detecting a proportional hazards alternative and a late effects alternative, respectively. One may thus wonder how the proposed maximum weighted logrank test behaves against a crossing hazards alternative. In this section, this issue is briefly investigated numerically.

We simulate $n = 2000$ censored survival times in two groups ($n_T = n_P = n/2$). The data in placebo and treatment groups are simulated from Weibull distributions $W(10, 1.37)$ and $W(30, 0.97)$, respectively, which ensures a discrepancy rate r equal to 0.2. The censoring fraction c is set to 0.8. The survival functions in the treatment and placebo groups are plotted on Figure 2. The empirical powers of the logrank test, Fleming–Harrington's test FH^3, maximum weighted logrank test MLR^3 and supremum logrank, calculated over 2000 simulated data sets, are, respectively, 0.597, 0.877, 0.849 and 0.524. The maximum weighted logrank test MLR^3 and Fleming–Harrington's test FH^3 outperform both logrank and supremum logrank tests.

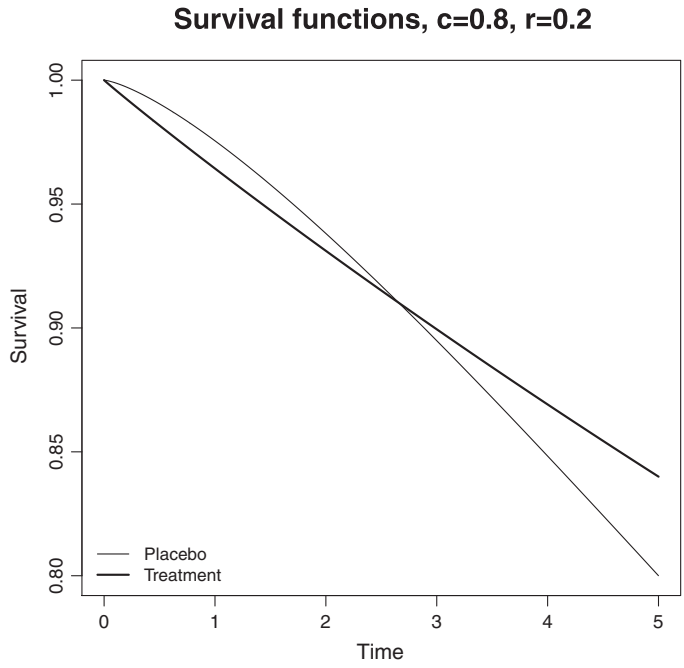

Figure 2. Crossing hazards alternative.

Copyright © 2014 John Wiley & Sons, Ltd.

5. Application to real data: the GuidAge study

GuidAge is a randomized, parallel-group, double-blinded trial registered to ClinicalTrials.gov under the number NCT00276510. Elderly subjects (aged 70 years or older) were enrolled in this trial. These subjects were free of dementia and had expressed a spontaneous memory complaint to their general practitioner in France. The subjects were randomized to either a daily 240 mg dose of standardized Ginkgo biloba extract (EGb761) or a placebo and were followed-up for 5 years by their physician and in expert memory centers. A total of 712 physicians and 25 memory centers participated in the trial. The primary outcome was conversion to probable Alzheimer's disease.

The former analysis of this trial was based on the logrank test. Assuming that under EGb761, the conversion rate from memory complaint to Alzheimer's disease is 25% less than under the placebo, the Alzheimer's disease-free rate after a 5-years long follow-up is equal to 89.63% under EGb761 and to 86.18% under the placebo. The total sample size ($n = 2800$) was calculated by letting $\alpha = 0.05$, $\beta = 0.2$ and by taking account of the dropout rate over the 5 years of follow-up. The p-value of the logrank test is 0.304, yielding the conclusion that there is no significant effect of the treatment.

However, EGb761 is a preventive treatment whose efficacity may require some preliminary exposure before an effect occurs. This is confirmed by the plot of hazard functions in the treatment and placebo groups (Figure 3). In this case, we suggest to use the statistic MLR^3 to test a treatment effect. We set $p_1 = 0.5$, that is, we do not favor any of the proportional hazards or late effects alternatives. Under this setting, the necessary sample size calculated from (12) is $n = 2351$. If we take account of the dropout rate, $n = 3001$. The results of the analysis are given in Table VII. The p-value of the proposed test MLR^3 is 0.008; thus, we conclude to a significant effect of the EGb761. For conciseness, we also report the results for Fleming–Harrington's test FH^q ($q = 2, 3, 4$) and for the maximum weighted logrank test MLR^q with $q = 2, 4$. All these tests conclude to a late effect of EGb761.

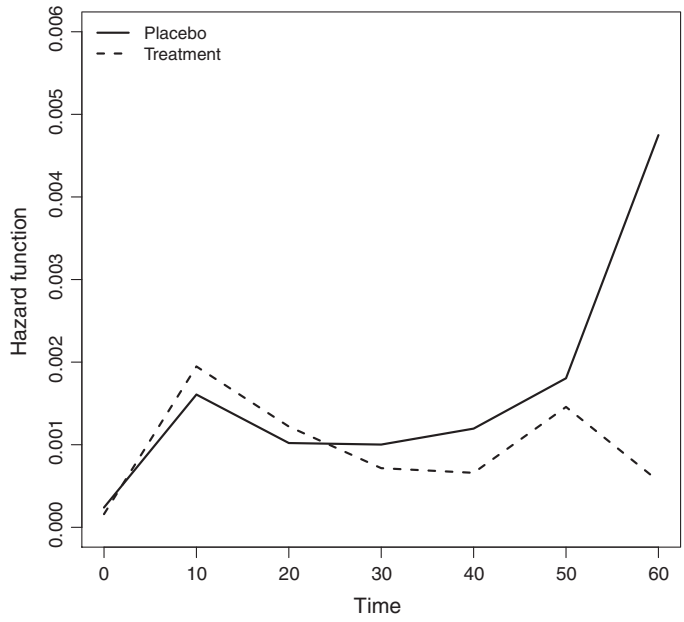

Figure 3. Hazard functions for the treatment and placebo groups in GuidAge study.

Table VII. GuidAge study. Various test statistics and their p-value.

	Logrank	FH^2	FH^3	FH^4	MLR^2	MLR^3	MLR^4	SLR
Test statistic	1.027	2.562	2.814	2.882	2.562	2.814	2.882	1.023
p-value	0.304	**0.010**	**0.004**	**0.003**	**0.018**	**0.008**	**0.006**	0.6084

In bold are the significant results at the 5% level.
FH: Fleming–Harrington's test; MLR: maximum weighted logrank test; SLR: supremum over time logrank test.

Copyright © 2014 John Wiley & Sons, Ltd.

6. Summary and conclusion

We propose a new statistic for testing equality of two survival distributions when late differences between groups may exist. Generally, at the design stage of a clinical trial, one cannot assure that late effects do really exist or not, and it is thus difficult to choose *a priori* between a logrank test and Fleming–Harrington's test for late effect. The proposed test statistic provides a solution to this dilemma. It has good power against both alternatives of proportional hazards and late effects. Moreover, it outperforms both logrank test (when late effects occur) and Fleming–Harrington's test (when proportional hazards hold). The proposed test also performs better than the supremum logrank test. We also provide a sample size formula for the proposed test statistic, and we illustrate our methodology on a real data set in the field of Alzheimer's disease.

We consider here the detection of late effects. Letting $p > 0$ and $q = 0$ in (1) allow detection of early effects and a maximum weighted logrank test for early effects can be constructed along the same lines as MLR^q. This test will enjoy similar properties as MLR^q.

Finally, in this article, we consider the maximum of the logrank and Fleming–Harrington's tests. Several other functions may be used to combine these statistics. The investigation of such functions and their relative merits is the topic for future research.

Acknowledgements

The research was supported by a grant from France Alzheimer/Fondation de France. The sponsor had no role in this study. The authors thank IPSEN for the access of the Guidage data.

References

1. Wimo A, Prince M. *World Alzheimer Report 2010: The Global Economic Impact of Dementia*. Alzheimer's Disease International: London, 2010.
2. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *American Journal of Public Health* 1998; **88**(9):1337–1342.
3. Brookmeyer R. Forecasting the global burden of Alzheimer's disease. *Alzheimer's and Dementia* 2007; **3**(3):186–191.
4. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer* 1959; **22**:719–748.
5. DeKosky ST. Ginkgo biloba for prevention of dementia: a randomized controlled trial. *Journal of the American Medical Association* 2008; **300**(19):2253–2262.
6. Lyketsos CG. Naproxen and celecoxib do not prevent Alzheimer's disease in early results from a randomized controlled trial. *Neurology* 2007; **68**(21):1800–1808.
7. Shumaker SA. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *Journal of the American Medical Association* 2003; **289**(20):2651–2662.
8. Shumaker SA. Conjugated equine estrogens and incidence of probable dementia and mild cognitive impairment in postmenopausal women: Women's Health Initiative Memory Study. *Journal of the American Medical Association* 2004; **291**(24):2947–2958.
9. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B* 1972; **34**:187–220.
10. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965; **52**:203–223.
11. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977; **64**(1):156–160.
12. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society: Series A* 1972; **135**:185–206.
13. Prentice RL. Linear rank tests with right censored data. *Biometrika* 1978; **65**(1):167–179.
14. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**(3):553–566.
15. Buyske S, Fagerstrom R, Ying Z. A class of weighted log-rank tests for survival data when the event is rare. *Journal of the American Statistical Association* 2000; **95**(449):249–258.
16. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 1990; **77**(4):853–864.
17. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc: New York, 1991.
18. Wu L, Gilbert PB. Flexible weighted log-rank tests optimal for detecting early and/or late survival differences. *Biometrics* 2002; **58**(4):997–1004.
19. Wallenstein S, Berger A. Weighted logrank tests to detect a transient improvement in survivorship. *Biometrics* 1997; **53**(2):736–744.
20. Gastwirth JL. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of The American Statistical Association* 1985; **80**(390):381–384.
21. Zucker DM. The efficiency of a weighted log-rank test under a percent error misspecification model for the log hazard ratio. *Biometrics* 1992; **48**(3):893–899.

Copyright © 2014 John Wiley & Sons, Ltd.

22. Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* 1996; **52**(2): 721–725.
23. Breslow NE, Edler L, Berger J. A two-sample censored-data rank test for acceleration. *Biometrics* 1984; **40**(4):1049–1062.
24. Self SG. An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics* 1991; **47**(3):975–986.
25. Lai TL, Ying Z. Rank regression methods for left-truncated and right-censored data. *The Annals of Statistics* 1991; **19**(2):531–556.
26. Pecková M, Fleming TR. Adaptive test for testing the difference in survival distributions. *Lifetime Data Analysis* 2003; **9**(3):223–238.
27. Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. *Biometrics* 2010; **66**(1):30–38.
28. Fleming TR, Harrington DP, O'Sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association* 1987; **82**(397):312–320.
29. Eng KH, Kosorok MR. A sample size formula for the supremum log-rank statistic. *Biometrics* 2005; **61**(1):86–91.
30. Kosorok MR, Lin CY. The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association* 1999; **94**(445):320–332.
31. Feng W, Wahed AS. Supremum weighted log-rank test and sample size for comparing two-stage adaptive treatment strategies. *Biometrika* 2008; **95**(3):695–707.
32. Vellas B, Coley N, Ousset PJ, Berrut G, Dartigues JF, Dubois B, Grandjean H, Pasquier F, Piette F, Robert P, Touchon J, Garnier P, Mathiex-Fortunet H, Andrieu S. Long-term use of standardised Ginkgo biloba extract for the prevention of Alzheimer's disease (GuidAge): a randomised placebo-controlled trial. *Lancet Neurology* 2012; **11**(10):851–859.
33. Garès V, Andrieu S, Dupuy JF, Savy N. About the parameter of Fleming–Harrington's test in prevention randomized controlled trials 10 2014. Submitted.
34. Gill RD. *Censoring and Stochastic Integrals. Mathematical Centre Tracts 124*. Mathematisch Centrum: Amsterdam, 1980.
35. Halperin M, Rogot E, Gurian J, Ederer F. Sample sizes for medical trials with special reference to long-term therapy. *Biometrics* 1967; **21**:13–24.
36. Schork MA, Remington RD. The determination of sample size in treatment-control comparisons for chronic disease studies in which noncompliance on nonadherence is a problem. *Journal of Chronic Diseases* 1967; **20**:233–239.
37. Lakatos E, Lan KG. A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* 1992; **11**: 179–191.

Copyright © 2014 John Wiley & Sons, Ltd.