Electronic Journal of Statistics
Vol. 8 (2014) 841–860
ISSN: 1935-7524
DOI: 10.1214/14-EJS911

A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests[*]

Valérie Garès

INSERM, UMR 1027, F-31073 Toulouse, France
University of Toulouse III, F-31073, Toulouse, France
e-mail: valerie.gares@inserm.fr

Sandrine Andrieu

University of Toulouse III, F-31073, Toulouse, France
INSERM, UMR 1027, F-31073 Toulouse, France
e-mail: sandrine.andrieu@univ-tlse3.fr

Jean-François Dupuy

IRMAR UMR 6625, CNRS and INSA of Rennes, F-35708 Rennes, France
e-mail: jean-francois.dupuy@insa-rennes.fr

and

Nicolas Savy

University of Toulouse III, F-31073, Toulouse, France
Toulouse Institute of Mathematics, UMR 5219, CNRS, F-31062, Toulouse, France
e-mail: nicolas.savy@math.univ-toulouse.fr

Abstract: The Fleming-Harrington and constant piecewise wei- ghted logrank tests for late effects in clinical trials are considered. Both tests depend on a parameter (q for the Fleming-Harrington and t^* for the constant piecewise weighted logrank) that has to be chosen before the trial analysis. The problem of choosing the most appropriate test and associated parameter value for a given trial is addressed. For this purpose, the tests are compared in terms of their sensitivity to q and t^* and of their asymptotic relative efficiency and necessary sample size. Some guidelines for choosing the most appropriate weight for testing late effects are provided. The methodology is illustrated on a medical dataset.

AMS 2000 subject classifications: Primary 62N03, 62P10; secondary 62M99; third 62E20.
Keywords and phrases: Survival data analysis, Fleming-Harrington test, constant piecewise weighted logrank test, Pitman asymptotic relative efficiency, prevention trial, Alzheimer's disease.

Received April 2013.

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

1. Introduction

Neurodegenerative dementias are a growing public health concern. For example, a recent study has estimated the prevalence of Alzheimer's disease at 115.4 million people in 2050 [20]. There is currently no effective treatment for this pathology, which makes its prevention a priority. Prevention is feasible due to 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 health-care systems (see [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 analysed using the logrank test and concluded that the various treatments do not exhibit a significant effect (*e.g.*, [4, 14, 16, 17]). However, one common and specific feature of these trials is that the treatment effect occurs late. The logrank test assumes that the hazard rates in the treatment and control groups are proportional and thus, it is not appropriate in this setting. One solution to this issue is to use weighted logrank tests.

Weighted logrank tests are constructed by plugging a weight function (usually depending of the sample size n) $(W_n(s),\ s \in \mathbb{R}^+)$ in the logrank statistic. The choice of a particular weight is motivated by the kind of deviation to the null hypothesis (of equality of the survival functions) that we are interested in detecting. A large amount of literature has been devoted to these tests so far and numerous weights have been proposed (see [6] and references therein). We focus here on two weight functions of particular interest:

- the Fleming-Harrington weight for late effects (see [5]) is defined as

$$W_n^q(s) = [1 - \hat{S}_n(s)]^q \tag{1}$$

  where $q \geqslant 0$ and $\hat{S}_n$ is the Kaplan-Meier estimator of the survival function S of the event time under the null hypothesis (thereafter, we refer the resulting weighted test to as the "Fleming-Harrington test" and we denote it by $\mathrm{FH}(q)$). Choosing the most appropriate value of q for a given trial remains a difficult task and an open problem since q is not directly interpretable in terms of late effects. However in [7], the authors have shown that the sensitivity of the Fleming-Harrington test to the value of q is small. The authors also provide some guidelines for using this test in clinical trials.
- the constant piecewise weight (CPW for short) is defined as

$$W^{t^*}(t) = \begin{cases} 0 & \text{if } t < t^* \\ 1 & \text{if } t > t^* \end{cases} \tag{2}$$

  for some t^*. The resulting weighted logrank statistic (subsequently referred to as the "CPWL statistic" and denoted by $\mathrm{CPWL}(t^*)$) has been studied in [21]. One appealing feature of (2) is that the parameter t^* is directly

> interpretable in terms of late effects. In practice, a reasonable value of t^* should therefore be based on the investigator's *a priori* knowledge about the late effects. However, as will be seen later in this paper, the CPWL test suffers from being sensitive to the value of t^*.

In this paper, we aim at providing some clear guidelines for choosing the most appropriate weight for testing late effects in a clinical trial.

We first evaluate the sensitivity of the Fleming-Harrington and CPWL tests to their respective parameters q and t^*. We conclude that the Fleming-Harrington test is less sensitive to the value of q than the CPWL test is to the value of t^*. In view of this result, the Fleming-Harrington weight (1) appears to be more appealing than the CPW (2). However in practice, it is easier to identify a reasonable range of values for t^* than to choose q. By comparing the Fleming-Harrington and CPWL tests (using arguments from asymptotic efficiency theory and some numerical comparisons), we are able to elucidate the relationship between q and t^*. From this, we establish some rules for choosing q from a given t^*. We finally propose a testing procedure, which consists in: 1) choosing t^* based on *a priori* knowledge about the expected late effects, 2) identifying and using the Fleming-Harrington test FH(q) which matches best (in a sense to be specified later) with the desired CPWL(t^*) test.

Applying the Fleming-Harrington and the CPWL tests to clinical trials raises the crucial issue of necessary sample size calculation. Several sample size formulas have been proposed under non-proportional hazards [10, 15, 11, 12, 13]. These formulas are essentially adaptations of the formula for the logrank test. However, it is desirable to obtain a sample size formula which is specific to both the tested alternative hypothesis and the optimal statistic used for testing this hypothesis. Motivated by this idea, we propose a new sample size formula for the Fleming-Harrington and CPWL tests for late effects.

All these issues arise in the GuidAge study (descibed in [19]) which motivates our investigations. GuidAge is a 5-years long prospective prevention study involving patients who spontaneously reported memory complaints. The primary objective was to investigate the effect of a treatment called EGb 761 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 the data to be analysed using the logrank test, which concluded that the treatment is ineffective. A re-analysis using the Fleming-Harrington test with $q = 3$ was conducted after the trial and concluded that the treatment is effective. Motivated by this example, we aim at providing guidelines for conducting the right analysis of clinical trials involving late effects and for choosing the most relevant value of the critical parameter q.

The remainder of the paper is organized as follows. In Section 2, we provide some background on weighted logrank statistics. In particular, we give a brief review of asymptotic relative efficiency and we recall some important results about the optimality of the Fleming-Harrington and CPWL tests for late effects detection. In Section 3, we conduct a simulation study to evaluate the sensitivity of these tests to their respective parameters q and t^*. In Section 4, we compare these tests both theoretically (using arguments of asymptotic efficiency) and numerically. We also investigate the relationship between q and t^*. Finally, we obtain a sample size formula for weighted logrank tests and we compare the sample sizes required by the Fleming-Harrington and CPWL tests. We illustrate our methodology on the GuidAge study in Section 5. A discussion and some perspectives conclude the paper.

2. Preliminaries on weighted logrank tests

2.1. Notations, definition and asymptotic relative efficiency

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 $\Lambda(t) = \int_0^t \lambda(s)ds$. T denotes the duration until the occurrence of some event of interest. In what follows, T is assumed to be right-censored, that is, we only observe the events that occur before some time C. Let T^i and C^i be the latent survival and censoring times respectively for the i-th individual. The observations consist of n independent couples $(X^i, \delta^i)_{i=1\dots n}$, where $X^i = \min(T^i, C^i)$ and $\delta^i = \mathbb{I}_{\{T^i \leqslant C^i\}}$. We assume that T^i and C^i are independent for every $i = 1, \dots, n$. Let G be the distribution function of the $(C^i)_{i=1,\dots,n}$, τ denote the duration of the study and $\tau' = \inf_{t\geqslant 0}\{\pi(t) = 0\}$, where $\pi(t) = (1 - F(t))\,(1 - G(t))$. We assume that $\tau < \tau'$. For every $t \geqslant 0$, we also define the random variables

$$N_n(t) = \sum_{i=1}^{n} \mathbb{I}_{\{X^i \leqslant t, \delta^i = 1\}} \qquad \text{and} \qquad Y_n(t) = \sum_{i=1}^{n} \mathbb{I}_{\{X^i \geqslant t\}}.$$

$N_n(t)$ is the number of failures at t and $Y_n(t)$ is the number of at-risk subjects at time t^-.

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

Consider the following null and alternative hypotheses:

$$\begin{cases} \mathcal{H}_0 & : \ F^T = F^P = F_{\theta^0}, \\ \mathcal{H}_1 & : \ F^T = F_{\theta^T} \quad \text{and} \quad F^P = F_{\theta^P}. \end{cases} \tag{3}$$

To solve this testing problem, one usually relies on the logrank statistic (see [5]) which can be written, at time t, as

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

This statistic is known to be optimal to test $\mathcal{H}_0$ against a proportional hazards alternative. The proportional hazards assumption states that the ratio of the hazards in the treatment and placebo groups is constant over time. When early or late effects are present, this ratio is not constant. In this case, one can extend the logrank to the so-called weighted logrank statistic, which is defined as

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

where (W_n) is a sequence of adapted, bounded, non-negative and predictable weighting processes. Assume that the following two conditions hold:

Condition 2.1. As $n \to \infty$, $n_P/n \to 1/2$ and $n_T/n \to 1/2$.

Condition 2.2. There exists a function $w \in \mathbb{D}$ (where $\mathbb{D}$ is the Skohorod space of càdlàg functions) such that $W_n(s) \xrightarrow{a.s.} w(s)$ as $n \to \infty$.

Under $\mathcal{H}_0$, LR_{W_n} converges weakly to a zero-mean Gaussian process. Under the general alternative $\mathcal{H}_1$, the asymptotic distribution of $\mathrm{LR}_{W_n}(t)$ is degenerate (see [7] and references therein). As a consequence, the weighted logrank tests are consistent: their power converges to 1 as n tends to infinity [5, 9]. In this setting, an appropriate comparison procedure consists in investigating the behaviour of the tests under a sequence of alternatives that converges to $\mathcal{H}_0$ as n tends to infinity. A relevant choice of the alternatives $(\theta^P_{n_P})$ and $(\theta^T_{n_T})$ in (3) is

$$\theta^P_{n_P} = \theta^0 + c \left(\frac{n_T}{n_P(n_P + n_T)} \right)^{1/2} \text{ and } \theta^T_{n_T} = \theta^0 - c \left(\frac{n_P}{n_T(n_P + n_T)} \right)^{1/2} \quad (4)$$

where $c \in \mathbb{R}$ is a constant (see [5]). This is the idea of asymptotic relative efficiency (ARE). There are different ways to define the ARE. In the next paragraph, we briefly review the Pitman ARE (see [18] for a definition and a detailed exposition). Let

$$k(s) = w(s) \frac{\pi^P(s) \pi^T(s)}{\pi(s)}. \quad (5)$$

and assume that the following additional regularity condition holds:

Condition 2.3. The function $\theta \to \lambda_\theta$ is differentiable at θ^0, with $\frac{\partial \lambda_\theta}{\partial \theta}|_{\theta=\theta^0} \neq 0$.

Then two results (due to [9], see Theorem 2.1 below) on the ARE of weighted logrank tests can be stated. These results will be crucial for proving our theorems in Section 4. The first result expresses the ARE of two weighted logrank statistics as the ratio of their respective asymptotic efficiencies (AE for short). The second gives the form of the limiting weight of a weighted logrank statistic with maximal AE. We refer to [9] for the proofs.

Theorem 2.1 ([9])**.** *Let* $\mathrm{LR}_{W^1_n}$ *and* $\mathrm{LR}_{W^2_n}$ *be two weighted logrank statistics satisfying the conditions 2.1, 2.2, 2.3. Consider a sequence of alternatives of the* *form* (3)*, with* $\theta^P_{n_P}$ *and* $\theta^T_{n_T}$ *defined by* (4)*. Then the Pitman ARE of* $\mathrm{LR}_{W^1_n}$ *with respect to* $\mathrm{LR}_{W^2_n}$ *is given by*

$$ARE(\mathrm{LR}_{W^1_n}, \mathrm{LR}_{W^2_n}) = \frac{AE(\mathrm{LR}_{W^1_n})}{AE(\mathrm{LR}_{W^2_n})}, \tag{6}$$

where

$$AE(\mathrm{LR}_{W^j_n}) = \frac{\left(\int_0^\tau \frac{k_j(s)}{\lambda_{\theta^0}(s)} \left.\frac{\partial\lambda_\theta}{\partial\theta}(s)\right|_{\theta=\theta^0} d\Lambda_{\theta^0}(s)\right)^2}{\int_0^\tau (k_j)^2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)}. \tag{7}$$

Moreover, the weighted logrank statistic with maximal AE has a limit weight function w *such that* k *in* (5) *is given by*

$$k : s \to \kappa \frac{1}{\lambda_{\theta^0}(s)} \left.\frac{\partial\lambda_\theta}{\partial\theta}\right|_{\theta=\theta^0}(s) \left(\frac{\pi^P(s)\pi^T(s)}{\pi(s)}\right),$$

where κ *is a constant.*

2.2. Optimality of the Fleming-Harrington and CPWL *statistics*

In logrank testing, a useful strategy is to consider the particular pattern of "shift assumptions up to a change of time" for the alternative hypothesis (see [9]). This can be defined through the following family of distribution functions:

$$F_\theta(t) = \Psi(g(t) + \theta), \qquad \theta \in \Theta, \tag{8}$$

where $g : [0, \infty[\to]-\infty, u^+[$ (with $u^+ \in \bar{\mathbb{R}}$) is a differentiable non-decreasing function and Ψ is a continuous cumulative distribution function with positive density Ψ' and an almost everywhere continuous second derivative Ψ'' (see [7, 6] for more details). Under the shift alternative and a relevant choice for g, Theorem 2.1 allows to express the patterns of the hazards in the treatment and placebo groups, for which the Fleming-Harrington and CPWL tests are optimal. These patterns can be given in terms of the shift $\Delta = \theta^P - \theta^T$ (see Theorem 2.2 for the Fleming-Harrington and Theorem 2.3 for CPWL).

Theorem 2.2 ([7])**.** *Given a shift* Δ*, the Fleming-Harrington test with* $q > 0$ *has maximum AE to test*

$$\begin{cases} \mathcal{H}_0 & : \ \lambda^T = \lambda^P, \\ \mathcal{H}_1 & : \ \lambda^T = \lambda^P \ \Gamma^q(., \Delta), \end{cases} \tag{9}$$

where for any $t \in \mathbb{R}^+$*,*

$$\Gamma^q(t, \Delta) = \frac{L^q((\mathcal{L}^q)^{-1}(\mathcal{L}^q(S^P(t)) + \Delta))}{L^q(S^P(t))},$$

and $\mathcal{L}^q]0,1[\to \mathbb{R}^-$ is a one-to-one map defined as the primitive of the function defined from $]0,1[$ to $\mathbb{R}^-$ by:

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

and B_{inc} is the incomplete beta function $B_{inc}(x,a,b) = \int_0^x s^{a-1}(1-s)^{b-1}ds$.

Theorem 2.3 ([21])**.** *Given a shift $\Delta = \theta^P - \theta^T$, the* CPWL *statistic with $0 \leqslant t^* \leqslant \tau$ has maximum efficiency to test*

$$\begin{cases} \mathcal{H}_0 & : \ \lambda^T = \lambda^P, \\ \mathcal{H}_1 & : \ \lambda^T = \lambda^P \ (1 - \Delta \mathbb{I}_{]t^*,\tau[}). \end{cases} \tag{10}$$

Theorems 2.2 and 2.3 are used to construct the data generating processes in the simulation study below. In these simulations, we investigate the sensitivity of the Fleming-Harrington and CPWL statistics to their parameters q and t^* respectively.

Remark 2.1. In fact, the CPWL depends on both t^* and τ. This problem can be overcome by considering $\frac{t^*}{\tau}$. The parameter becomes dimension-free which makes the weight independent of the trial duration. Thereafter, in order to avoid any ambiguity, we consider $\tau = 1$.

3. Sensitivity of the Fleming-Harrington and CPWL statistics

The simulation scenarios are described in Section 3.1 (for the Fleming-Harrington test) and Section 3.2 (for the CPWL test). The results are discussed in Section 3.3.

3.1. Fleming-Harrington test: Sensitivity to q

Data generating process (DGP1). We simulate data according to a generating process under which the Fleming-Harrington test with parameter q_S is optimal (thereafter, q_S will stand for "the q-value used for simulating the data"). Given some $\tau > 0$, $q_S > 0$ and $c = S^P(\tau)$, a rate r is defined as

$$r = \frac{S^T(\tau) - S^P(\tau)}{1 - S^P(\tau)}. \tag{11}$$

The data in the placebo group are simulated from an exponential distribution with parameter $a > 0$, where a is fixed from the desired proportion of censored data:

$$a = -\frac{\ln(S^P(\tau))}{\tau}. \tag{12}$$

Based on Theorem 2.2, the data in the treatment group are simulated from the hazard function

$$\lambda^T(t) = a \ \frac{L^q((\mathcal{L}^q)^{-1}(\mathcal{L}^q(e^{-at}) + \Delta(q)))}{L^q(e^{-at})} \tag{13}$$

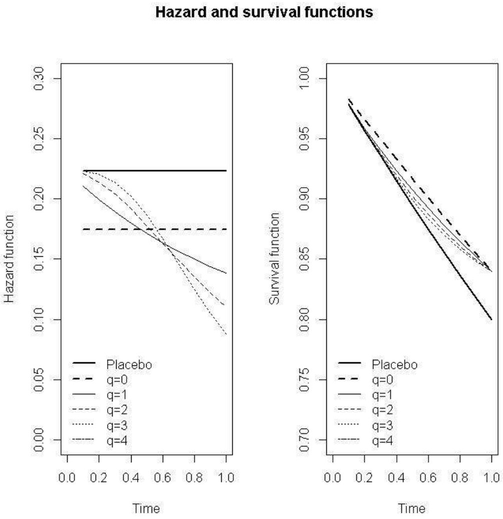

FIG 1. *Hazard and survival functions for the DGP1. The curves for* $q = 0, 1, 2, 3, 4$ *correspond to the hazard and survival functions in the treatment group under optimality of the Fleming-Harrington test.*

with $q = q_S$ and $\Delta(q)$ given by

$$\Delta(q) = \theta^T - \theta^P = \mathcal{L}^q(r(1 - S^P(\tau)) + S^P(\tau)) - \mathcal{L}^q(S^P(\tau)).$$

We consider well-balanced placebo and treatment groups that is, $n_P = n_T = \frac{n}{2}$. A sample simulated from this data generating process is denoted by $\mathcal{S}(q_S, n, r, c)$.

To get an insight into the patterns of λ^T and λ^P for which the Fleming-Harrington test is optimal, we plot the hazard functions (12) and (13) and the corresponding survival functions (see Figure 1). On these graphs, $\tau = 1$ year, $S^P(\tau) = 0.8$, $r = 0.2$ and q varies over $\{0, 1, 2, 3, 4\}$. As expected, we note that the larger q is, the later the treatment effect can be detected.

Simulation design. We simulate $N = 2000$ samples $\mathcal{S}(q_S, n, r, c)$ for each $q_S \in \{0, 1, 2, 3, 4, 5\}$ (and $\tau = 1$). The logrank test and the Fleming-Harrington tests with $q = q_T$ (with q_T successively equal to $1, 2, 3, 4$) are applied to each of the N samples and the empirical powers of all these tests are obtained (in what follows, q_T will stand for "the q-value used for testing the data", as opposed to the value q_S used to simulate them).

Similarly, empirical levels are obtained by simulating $N = 2000$ samples under the hypothesis $\mathcal{H}_0$ of equality of the survival distributions of the treatment and placebo.

We considered several values for n $(100, 500, 1000, 2000)$, c $(0.2, 0.5, 0.8)$, and r $(0.1, 0.2, 0.3)$. Due to space limitations, a part of the results only is provided in the upper part of Table 1 (empirical level) and in Table 3 (empirical powers,

TABLE 1
*Empirical level of the Fleming-Harrington (*FH*) tests (respectively* CPWL *tests) for various values of q_T (respectively t_T^*)*

		FH(q)				
n	c	Logrank	$q_T = 1$	$q_T = 2$	$q_T = 3$	$q_T = 4$
100	0.2	0.052	0.048	0.051	0.049	0.053
	0.5	0.054	0.047	0.047	0.045	0.043
	0.8	0.046	0.045	0.046	0.048	0.047
500	0.2	0.051	0.053	0.050	0.050	0.047
	0.5	0.049	0.052	0.050	0.049	0.040
	0.8	0.046	0.053	0.054	0.052	0.046
1000	0.2	0.049	0.047	0.052	0.054	0.058
	0.5	0.048	0.055	0.047	0.046	0.044
	0.8	0.046	0.048	0.049	0.050	0.050
2000	0.2	0.051	0.047	0.049	0.049	0.051
	0.5	0.050	0.047	0.045	0.048	0.045
	0.8	0.049	0.049	0.053	0.052	0.052

		CPWL(t^*)				
n	c		$t_T^* = 0.2$	$t_T^* = 0.4$	$t_T^* = 0.6$	$t_T^* = 0.8$
100	0.2		0.045	0.052	0.047	0.048
	0.5		0.053	0.047	0.049	0.044
	0.8		0.059	0.057	0.054	0.021
500	0.2		0.041	0.050	0.055	0.047
	0.5		0.044	0.046	0.047	0.047
	0.8		0.052	0.056	0.058	0.052
1000	0.2		0.055	0.054	0.054	0.051
	0.5		0.052	0.049	0.044	0.056
	0.8		0.051	0.050	0.049	0.050
2000	0.2		0.054	0.045	0.046	0.047
	0.5		0.053	0.051	0.048	0.049
	0.8		0.048	0.048	0.059	0.054

$n = 2000$, $c = 0.8$, $r = 0.2$). Additional results can be found in a supplementary document [8] available at the following address: http://www.math.univ-toulouse.fr/~vgares/Supp/CPWFHsupp.pdf.

3.2. CPWL *test: Sensitivity to* t^*

Data generating process (DGP2). We simulate data according to a generating process under which the CPWL(t_S^*) test is optimal (in what follows, t_S^* will stand for "the value of t^* used for simulating the data"). Let $\tau > 0$, $0 \leqslant t_S^* \leqslant \tau$, $c = S^P(\tau)$ and r be defined as in (11). The data in the placebo group are simulated from an exponential distribution with parameter a given by (12). The data in the treatment group are simulated from the hazard function

$$\lambda^T(t) = a(1 - \Delta(t^*)\mathbb{I}_{\{t > t^*\}}) \tag{14}$$

where $t^* = t_S^*$ and

$$\Delta(t^*) = \frac{1}{a} \ln\left(\frac{S^T(\tau)}{S^P(\tau)}\right) \frac{1}{\tau - t^*}.$$

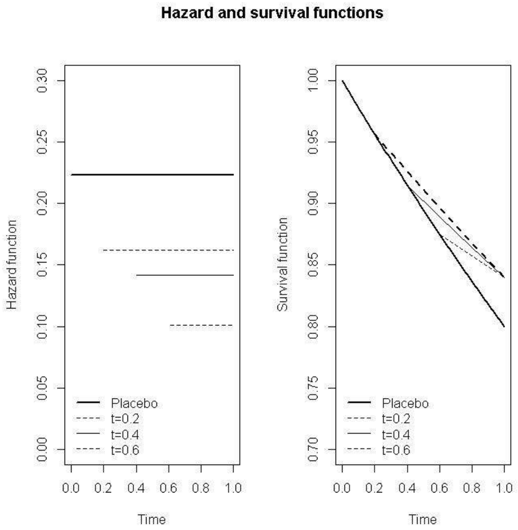

Fig 2. *Hazard and survival functions for the DGP2. The curves for* $t^* = 0.2, 0.4, 0.6, 0.8$ *correspond to the hazard and survival functions in the treatment group under optimality of the* CPWL *test.*

We consider well-balanced placebo and treatment groups. A sample simulated from this data generating process is denoted by $\mathcal{S}(t_S^*, n, r, c)$. Figure 2 plots the hazard and survival functions for this DGP when $\tau = 1$ year, $S^P(\tau) = 0.8$, $r = 0.2$ and t^* varies over $\{0, 1, 2, 3, 4\}$. As expected again, we note that the larger t^* is, the later the treatment effect can be detected.

Simulation design. We simulate $N = 2000$ samples $\mathcal{S}(t_S^*, n, r, c)$, with t_S^* ranging from 0 to 0.6 by 0.2 (and $\tau = 1$). The logrank test and the CPWL tests with $t^* = t_T^*$ (with t_T^* ranging from 0.2 to 0.8 by 0.2) are applied to each of the N samples, and their empirical powers are calculated (in the sequel, t_T^* will stand for "the value of t^* used for testing the data", as opposed to the value t_S^* used for the simulations). Similarly, empirical levels are obtained by simulating $N = 2000$ samples under $\mathcal{H}_0$. We considered the same values of n, r, and c as in the DGP1. The results for the empirical levels (respectively empirical powers) are given in the lower part of Table 1 (respectively in Table 3). Additional results are also provided in the web-based supplementary document [8].

3.3. Results

From Table 1, the Fleming-Harrington and CPWL tests appear to respect the nominal level. From Table 3 and the supplementary document [8], the power of both tests increases with n and r and decreases when the censoring increases. In each scenario, we note that the Fleming-Harrington test (respectively the CPWL test) has maximal power when q_T (respectively t_T^*) is taken equal to q_S (respectively t_S^*). We also observe that the empirical power of Fleming-Harrington test only slightly varies when q_T varies, which means that the sensitivity of the Fleming-Harrington test to the value of q_T is very small. Therefore, misspecifying q_T will only have a limited impact on the result of the test. This is a nice feature of the Fleming-Harrington test in view of its application in clinical trials. On the contrary, the CPWL test appears to be sensitive to the value of t^*. Its power decreases markedly when the true value t_S^* is misspecified.

4. A comparison of the CPWL and Fleming-Harrington tests

In this section, we compare the CPWL and Fleming-Harrington tests. We also investigate the relationship between q and t^*. Finally, we obtain a sample size formula for weighted logrank tests and we compare the sample sizes required by the Fleming-Harrington and CPWL tests.

4.1. Asymptotic efficiency comparisons

We first need some additional notations. Let $(\mathrm{LR}_{W_n^1})$ and $(\mathrm{LR}_{W_n^2})$ be two sequences of weighted logrank tests. Then $\widetilde{ARE}(\mathrm{LR}_{W_n^1}, \mathrm{LR}_{W_n^2})$ will denote the ARE of $\mathrm{LR}_{W_n^1}$ with respect to $\mathrm{LR}_{W_n^2}$ under a sequence of alternatives such that $AE(\mathrm{LR}_{W_n^2})$ is maximal. Similarly, $\widetilde{ARE}(\mathrm{LR}_{W_n^2}, \mathrm{LR}_{W_n^1})$ will denote the ARE of $\mathrm{LR}_{W_n^2}$ with respect to $\mathrm{LR}_{W_n^1}$ under alternatives such that $AE(\mathrm{LR}_{W_n^1})$ is maximal.

Theorem 4.1. *Assume that the conditions 2.1, 2.2 and 2.3 hold. Then*

$$\widetilde{ARE}(\mathrm{LR}_{W_n^1}, \mathrm{LR}_{W_n^2}) = \widetilde{ARE}(\mathrm{LR}_{W_n^2}, \mathrm{LR}_{W_n^1}).$$

Proof. The following holds from (6) and (7) in Theorem 2.1:

$$\begin{aligned} &ARE(\mathrm{LR}_{W_n^1}, \mathrm{LR}_{W_n^2}) \\ &\quad = \frac{\left(\int_0^\tau \frac{k_1(s)}{\lambda_{\theta^0}(s)} \left.\frac{\partial \lambda_\theta}{\partial \theta}\right|_{\theta=\theta^0}(s) d\Lambda_{\theta^0}(s)\right)^2}{\int_0^\tau (k_1)^2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)} \cdot \frac{\int_0^\tau (k_2)^2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)}{\left(\int_0^\tau \frac{k_2(s)}{\lambda_{\theta^0}(s)} \left.\frac{\partial \lambda_\theta}{\partial \theta}\right|_{\theta=\theta^0}(s) d\Lambda_{\theta^0}(s)\right)^2}. \end{aligned}$$

Next, if $\mathrm{LR}_{W_n^2}$ has maximal AE, Theorem 2.1 implies that

$$k_2(s) = \frac{1}{\lambda_{\theta^0}(s)} \left.\frac{\partial \lambda_\theta}{\partial \theta}\right|_{\theta=\theta^0}(s) \frac{\pi^P(s)\pi^T(s)}{\pi(s)},$$

which in turn implies that:

$$\begin{aligned} &\widetilde{ARE}(\mathrm{LR}_{W_n^1}, \mathrm{LR}_{W_n^2}) \\ &\quad = \frac{\left(\int_0^\tau k_1(s) k_2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)\right)^2}{\int_0^\tau (k_1)^2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)} \cdot \frac{\int_0^\tau (k_2)^2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)}{\left(\int_0^\tau (k_2)^2(s) \frac{\pi(s)}{\pi^P(s)\pi^T(s)} d\Lambda_{\theta^0}(s)\right)^2}, \end{aligned}$$

$$= \frac{\left(\int_0^\tau k_1(s)k_2(s)\frac{\pi(s)}{\pi^P(s)\pi^T(s)}d\Lambda_{\theta^0}(s)\right)^2}{\int_0^\tau (k_1)^2(s)\frac{\pi(s)}{\pi^P(s)\pi^T(s)}d\Lambda_{\theta^0}(s).\int_0^\tau (k_2)^2(s)\frac{\pi(s)}{\pi^P(s)\pi^T(s)}d\Lambda_{\theta^0}(s)}.$$

It is easily seen that this latter expression is symmetric in (k_1, k_2), which concludes the proof. □

Thereafter, we suppose that T is exponentially distributed under $\mathcal{H}_0$ and that the right-censoring time C is of type I. This means that under $\mathcal{H}_0$, $\pi(t) = S(t) = \exp(-at)$ and $\lambda(t) = a$, for $t \in [0, \tau[$. Then for every $q \in \mathbb{R}^+$ and $t^* \in [0, \tau[$, we define the function

$$f(q, t^*) = \widetilde{ARE}(\mathrm{LR}_{W_n^q}, \mathrm{LR}_{W_n^{t^*}}) = \widetilde{ARE}(\mathrm{LR}_{W_n^{t^*}}, \mathrm{LR}_{W_n^q}).$$

Note that f is well-defined by Theorem 4.1. The next theorem provides some information about the relation between t^* and q.

Theorem 4.2. *Let $t^* \in [0, \tau[$. Then there exists a unique $q(t^*) \in \mathbb{R}^+$ such that*

$$\max_{q\in\mathbb{R}^+} f(q, t^*) = f(q(t^*), t^*).$$

Let $q \in \mathbb{R}^+$. Then there exists a unique $t^*(q) \in [0, \tau[$ such that

$$\max_{t^*\in[0,\tau[} f(q, t^*) = f(q, t^*(q)).$$

Proof. If T is exponentially distributed under $\mathcal{H}_0$, f can be written explicitly:

$$f(q, t^*) = \frac{2q+1}{(q+1)^2}\frac{1-\exp(-a\tau)}{\exp(-at^*) - \exp(-a\tau)}\left(1 - \left(\frac{1-\exp(-at^*)}{1-\exp(-a\tau)}\right)^{q+1}\right)^2.$$

Letting

$$x = \frac{1-\exp(-at^*)}{1-\exp(-a\tau)},$$

f can be reparameterized as

$$f(q, x) = \frac{2q+1}{(q+1)^2}\frac{1}{1-x}\left(1 - x^{q+1}\right)^2.$$

Consider the partial maps $q \to f(q, x)$ and $x \to f(q, x)$. Using some standard analysis arguments, it is tedious but straightforward to show that both functions admit a unique maximum, which concludes the proof. □

This theorem proves the existence and uniqueness of the maximum of the partial maps $q \to f(q, x)$ and $x \to f(q, x)$. However, it is important to note that it provides no information about the shape of the relations $q \to t^*(q)$ and $t^* \to q(t^*)$. Numerical methods can be used to obtain some information about these relations. Figure 3 provides a 3D-plot of $\widetilde{ARE}(\mathrm{LR}_{W_n^{t^*}}, \mathrm{LR}_{W_n^q})$ as a function of t^* and q. One clearly observes that for every q (respectively t^*), there is a unique t^* (respectively q) such that the ARE is maximal. An additional figure is provided in the web-based supplementary document [8].

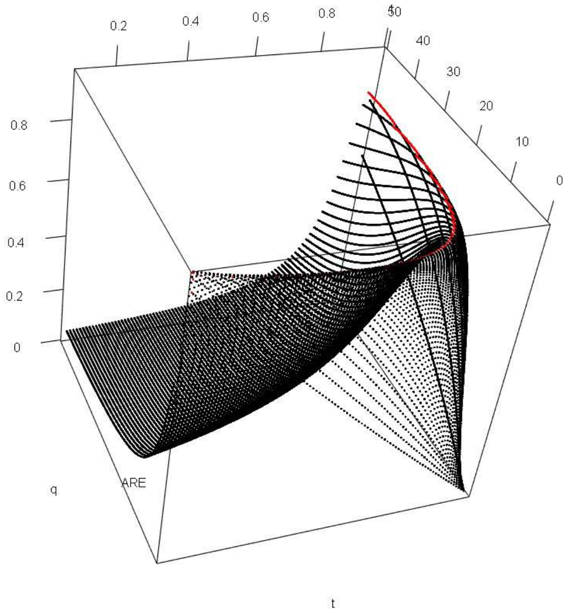

Fig 3. *3D-plot of ARE as a function of t^* and q.*

From [7], the Fleming-Harrington test is not very sensitive to the value of q, which is a desirable property in view of applications. But choosing the most relevant q for testing a given pattern of late effects is difficult since q is not directly interpretable in terms of late effects. On the contrary, t^* is easily interpretable in terms of late effects but the CPWL test is highly sensitive to a variation of t^*. Theorems 4.1, 4.2 and Figure 3 precisely give us a way out of this dilemma. Based on these results, we propose the following testing strategy for a given clinical trial: 1) choose t^* based on *a priori* knowledge about the expected late effects, 2) identify and use the test FH(q) which is the closest from CPWL(t^*) in terms of asymptotic efficiency (using Theorem 4.2).
This procedure should be relevant only if the map $t^* \to q(t^*)$ is not too sensitive to the value of t^*. Note on Figure 3 that this map is not a straight line, thus its sensitivity to a variation of t^* depends on t^*. But one can observe that the range of t^* where $t^* \to q(t^*)$ is sensitive is limited to a relatively extreme domain, which ensures a good stability of the choice of q for most of the t^* values.

As an illustration, Table 2 provides the correspondence between t^* and q when $c = 80\%$ and $r = 20\%$.

4.2. Simulation-based comparisons

In this simulation study, we investigate the behaviour of the CPWL test (respectively Fleming-Harrington test) when the data are simulated under optimal

Table 2

Correspondence between q and t^* to give $f(q,t^*)$ maximal, $r = 0.2$, $c = 0.8$

FH	$q =$	1	2	3	4
CPW	$t^*(q) =$	0.3	0.5	0.6	0.7

CPW	$t^* =$	0.2	0.4	0.6	0.8
FH	$q(t^*) =$	0.5	1.2	2.4	5.9

Table 3

Empirical power of FH *tests (respectively* CPWL *tests) for various q_T (respectively t^*_T) when the data are generated under the optimal hypothesis for* FH(q_S) *(respectively* CPWL(t^*_S)*). $c = 0.8$, $r = 0.2$, $n = 2000$*

	FH(q)				
q_S	Logrank	$q_T = 1$	$q_T = 2$	$q_T = 3$	$q_T = 4$
0	**0.640**	0.534	0.420	0.349	0.294
1	0.620	**0.743**	0.713	0.670	0.632
2	0.609	0.845	**0.877**	0.871	0.853
3	0.593	0.873	0.912	**0.914**	**0.914**
4	0.587	0.887	0.940	0.957	**0.961**
5	0.588	0.910	0.962	0.974	**0.980**
	CPWL(t^*)				
t^*_S/τ		$t^*_T = 0.2$	$t^*_T = 0.4$	$t^*_T = 0.6$	$t^*_T = 0.8$
0		0.543	0.420	0.294	0.167
0.2		**0.737**	0.615	0.439	0.261
0.4		0.745	**0.873**	0.704	0.402
0.6		0.722	0.861	**0.978**	0.782

alternatives for the Fleming-Harrington test (respectively CPWL test). We consider two sets of scenarios for late differences (letting $\tau = 1$ in both scenarios):

- For each $q_S \in \{0, 1, 2, 3, 4\}$, we simulate $N = 2000$ samples $\mathcal{S}(q_S, 2000, 0.2, 0.8)$. The CPWL test with $t^* = t^*_T$ (for t^*_T ranging from 0.2 to 0.8 by 0.2) and the logrank test are applied to the N samples, and their empirical powers are calculated.
- We simulate N samples $\mathcal{S}(t^*_S, 2000, 0.2, 0.8)$, with t^*_S ranging from 0 to 0.6 by 0.2. The logrank and Fleming-Harrington tests with $q = q_T$ (with q_T ranging from 1 to 4 by 1) are calculated on each sample and their empirical powers are obtained.

Table 4 gives the empirical power of the CPWL test for the various combinations of q_S and t^*_T (that is, for data generated under optimal alternatives for the

Table 4

Empirical power of the CPWL(t^*_T) *test when the data are generated under the optimal hypothesis for* FH(q_S)*. $c = 0.8$, $r = 0.2$, $n = 2000$*

q_S	Logrank	$t^*_T = 0.2$	$t^*_T = 0.4$	$t^*_T = 0.6$	$t^*_T = 0.8$
0	**0.644**	0.543	0.420	0.294	0.167
1	0.650	0.715	**0.719**	0.624	0.425
2	0.605	0.723	**0.790**	0.773	0.630
3	0.578	0.707	0.831	**0.873**	0.783
4	0.601	0.715	0.856	**0.918**	0.882

TABLE 5
Empirical power of $\mathrm{FH}(q_T)$ *when the data are generated under the optimal hypothesis for* $\mathrm{CPWL}(t_S^*)$. $c = 0.8$, $r = 0.2$, $n = 2000$

t_S^*	$q_T = 0$	$q_T = 1$	$q_T = 2$	$q_T = 3$	$q_T = 4$
0	**0.635**	0.512	0.402	0.329	0.276
0.2	0.620	**0.694**	0.608	0.515	0.452
0.4	0.623	**0.822**	0.814	0.766	0.707
0.6	0.594	0.896	0.948	**0.957**	0.953

Fleming-Harrington test). Similarly, Table 5 gives the empirical power of the Fleming-Harrington test when the data are generated under optimal alternatives for the CPWL. We provide results for $r = 0.2$, $c = 0.8$ and $n = 2000$. Some additional results are provided in the supplementary document [8].

As expected, we observe from Table 4 that as q_S increases, the value of t_T^* which ensures the largest power for a $\mathrm{CPWL}(t_T^*)$ test increases (a similar remark holds from Table 5 when t_S^* increases). We also note that the power of the Fleming-Harrington test is less sensitive to q_T (for a given t_S^*) than the power of $\mathrm{CPWL}(t^*)$ is to t^* for a given q_S. This confirms our previous finding that the Fleming-Harrington test is less sensitive to q than the CPWL test is to t^*. In this sense, the Fleming-Harrington test should be preferred in practice.

4.3. Sample size calculations

Before launching a clinical trial, one needs to know how much resource is needed to ensure that the study has enough power to detect the difference of interest. In this section, we compare the sample sizes required by the Fleming-Harrington and CPWL tests for testing late effects. Assuming a type I censoring scheme, we provide a sample size formula for testing the hypotheses (3) using a weighted logrank test.

Theorem 4.3. *Let r and $S^P(\tau)$ be given and assume that $n_T = n_P = \frac{n}{2}$. The sample size needed to achieve a power $1-\beta$ with a type I error α, when testing the hypotheses (3) using a weighted logrank test, is given by:*

$$n = \frac{\sigma_1^2}{\mu^2}(z_{1-\alpha/2} + z_{1-\beta})^2, \tag{15}$$

where z_γ denotes the quantile of order γ of a standard normal distribution and

$$\sigma_1^2 = \int_0^\tau w(s)\left(\frac{\pi^P(s)(\pi^T(s))^2}{(\pi(s))^2}d\Lambda_{\theta^P}(s) + \frac{(\pi^P(s))^2\pi^T(s)}{(\pi(s))^2}d\Lambda_{\theta^T}(s)\right),$$

$$\mu = \int_0^\tau w(s)\frac{\pi^P(s)\pi^T(s)}{\pi(s)}(d\Lambda_{\theta^P}(s) - d\Lambda_{\theta^T}(s)),$$

with $S(s) = \frac{S_{\theta^P}(s)+S_{\theta^T}(s)}{2}$.

The proof is straightforward and is therefore omitted.

Table 6

Sample size computation for Fleming-Harrington test with different values of q (upper table) and the corresponding CPWL(t^*) *test with $t^* = t^*(q)$ (lower table)*

		FH(q)				
c	r	Logrank	$q = 1$	$q = 2$	$q = 3$	$q = 4$
0.2	0.1	838	875	820	755	697
	0.2	251	252	237	220	206
	0.3	129	128	122	11	11
0.5	0.1	3181	2795	2310	1953	1691
	0.2	811	699	581	496	436
	0.3	366	315	264	230	208
0.8	0.1	12387	9692	7454	6018	5056
	0.2	2992	2332	1806	1474	1253
	0.3	1281	1001	788	655	572

		CPWL(t^*)				
c	r		$t^* = 0.2$	$t^* = 0.4$	$t^* = 0.6$	$t^* = 0.8$
0.2	0.1		722	623	449	183
	0.2		206	171	112	24
	0.3		103	83	50	1
0.5	0.1		2571	2010	1353	580
	0.2		634	481	299	87
	0.3		278	204	116	13
0.8	0.1		9735	7272	4718	2027
	0.2		2300	1670	1016	327
	0.3		964	677	379	66

We illustrate this theorem in a short numerical study. Letting $\alpha = 0.05$ and $\beta = 0.2$, we calculate the sample size needed for testing the hypotheses (9) (respectively (14)) using the Fleming-Harrington test (respectively the CPWL test). We consider various settings, defined by the censoring fraction: 0.2, 0.5, 0.8 and the rate value: 0.1, 0.2, 0.3. Table 6 gives the sample size needed for the Fleming-Harrington test (for different q) and for the corresponding CPWL(t^*) test with $t^* = t^*(q)$.

For both tests, the sample size needed to achieve the prescribed power and level increases as the censoring increases, and decreases when the rate r increases. Also, for the Fleming-Harrington test, the sample size decreases when q increases from 1 (the sample size is sometimes larger for $q = 1$ than for $q = 0$). For the CPWL test, the sample size decreases when t^* increases from 0. Finally, we observe that the sample size needed for the Fleming-Harrington test is generally larger than for the CPWL test. However, the difference stays moderate in most of the cases.

5. Application to real data: GuidAge Study

Setting of the trial. GuidAge is a randomized, parallel-group, double-blinded trial. Elderly subjects (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 standardised ginkgo biloba extract (EGb761) or a placebo, and were followed-up for 5 years by their physician and in expert memory centres. A total of 712 physicians and 25 memory centres participated in the trial. The primary outcome was conversion to probable Alzheimer's disease. This study is registered to ClinicalTrials.gov under the number NCT00276510.

The former statistical analysis. The former analysis of this trial was based on the logrank test. Assuming that under EGb 761, 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 EGb 761 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 proposed statistical analysis. The clinical trial considered here is a prevention trial. Thus we can assume that a late effect exists and we suggest to use the Fleming-Harrington test with $p = 0$. We need to choose q. Based on [1], we assume that an effect occurs between the second and third years of the trial. When t^* ranges from 2 to 3, q lies between 1.2 and 2.4. One usually wishes to use an integer value for q. Thus, in order to minimize the necessary sample size, we suggest to use $q = 3$. Under this value, the necessary sample size is 1778.

Results and discussion. The p-value of the logrank test is 0.3044, yielding the conclusion that there is no significant effect of the treatment. Under the former and planned statistical analysis design, the study is thus declared negative. But the proposed Fleming-Harrington test with $p = 0$ and $q = 3$ has p-value 0.0041. From this result, there is a very significant treatment effect and the study would be declared positive (note that there is no contradiction between the two conclusions, since the corresponding statistics do not test the same kind of difference between the groups).

In order to illustrate the sensitivity of the Fleming-Harrington and CPWL tests to q and t^*, we tested the data with q ranging from 1 to 5 and t^* ranging from 1 to 5. The results are given in Tables 7 and 8. The tests are significant

TABLE 7
Fleming-Harrington test in the GuidAge study. NS: non significant, S: significant, VS: very significant difference

$q =$	0	1	2	3	4	5
Statistic	1.030	1.964	2.562	2.814	2.882	2.858
p-value	0.304	0.049	0.010	0.004	0.003	0.002
	NS	*S*	*S*	*VS*	*VS*	*VS*

TABLE 8
CPWL(t^*)*'s test in the GuidAge study. NS: non significant, S: significant, VS: very significant difference*

$t^* =$	0.5	1	1.5	2	2.5	3	3.5	4	4.5
Statistic	-0.947	-0.848	-1.68	-1.690	-2.142	-2.351	-2.096	-2.122	-2.360
p-value	0.344	0.397	0.092	0.091	0.032	0.019	0.036	0.034	0.018
	NS	*NS*	*NS*	*NS*	*S*	*S*	*S*	*S*	*S*

for every $q \geqslant 1$ and $t^* \geqslant 2.5$. The results are less sensitive for the Fleming-Harrington test than for the CPWL test.

6. Conclusion and discussion

We have shown that the Fleming-Harrington test statistic possesses the nice feature of a small sensitivity to q. On the contrary, the choice of t^* influences substantially the outcome of the CPWL test. But t^* can be directly interpreted in terms of late effects, which is not the case for q. In this paper, we proposed to combine the advantages of both statistics to define a testing strategy for late effects in a clinical trial.

Using the asymptotic relative efficiency, we have described the relation between q and t^*. From this, we have proposed a two-step strategy for testing late effects in a clinical trial: one may first choose t^* (based on *a priori* knowledge about the expected late effects) and then identify and use the Fleming-Harrington test FH(q) which matches best with the desired CPWL(t^*) test. Moreover, we have shown that such a procedure does not result in an unreasonable increase of the necessary sample size. The proposed strategy therefore retains the nice features of both tests (namely the interpretation of t^* in terms of late effects and the robustness of FH(q) to the value of q).

We considered here the Fleming-Harrington and constant piecewise weights. Several other weight functions might be used to detect late effects. Investigating their relative merits and generalizing our testing strategy to incorporate their respective strength constitute a topic for future research.

Supplementary Material

Supplement to "A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests"
(doi: 10.1214/14-EJS911SUPP; .pdf).

References

[1] Andrieu, S., Gillette, S., Amouyal, K., Nourhashemi, F., Reynish, E., Ousset, P. J., Albarede, J. L., Vellas, B. and Grandjean, H. (2003). Association of Alzheimer's disease onset with ginkgo biloba and other symptomatic cognitive treatments in a population of women aged 75 years and older from the EPIDOS study. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* **58(4)** 372–377.

[2] Brookmeyer, R. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's and Dementia* **3(3)** 186–191.

[3] Brookmeyer, R., Gray, S. and C., K. (1998). Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *American Journal of Public Health* **88(9)** 1337–1342.

[4] DeKosky, S. T. (2008). Ginkgo biloba for prevention of dementia: A randomized controlled trial. *Journal of the American Medical Association* **300(19)** 2253–2262.

[5] Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons Inc, New York. MR1100924 (92i:60091)

[6] Garès, V., Andrieu, S., Dupuy, J. F. and Savy, N. (2014). A omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. In revision for Statistics in Medicine.

[7] Garès, V., Andrieu, S., Dupuy, J. F. and Savy, N. (2014). About the parameter of Fleming-Harrington's test in prevention randomized controlled trials. Submitted to Journal of Royal Statistical Society – Serie C.

[8] Garès, V., Andrieu, S., Dupuy, J. F. and Savy, N. (2014). Supplement to "A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests". DOI:10.1214/14-EJS911SUPP.

[9] Gill, R. (1980). *Censoring and stochastic integrals*. Mathematisch Centrum. MR0596815

[10] Halperin, M., Rogot, E., Gurian, J. and Ederer, F. (1967). Sample sizes for medical tirials with special reference to long-term therapy. *Biometrics* **21** 13–24.

[11] Lakatos, E. (1986). Sample sizes determination in clinical trials with time-dependant rates of losses and noncompliance. *Controlled Clinical Trials* **7** 189–199. MR931637

[12] Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* **44** 229–241. MR931637

[13] Lakatos, E. and Lan, K. G. (1992). A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* **11** 179–191.

[14] Lyketsos, C. G. (2007). Naproxen and celecoxib do not prevent Alzheimer's disease in early results from a randomized controlled trial. *Neurology* **68(21)** 1800–1808.

[15] Schork, M. A. and Remington, R. D. (1967). The determination of sample size in treatment-control comparisons for chronic disease studies in which noncompliance on nonaddherence is a problem. *Journal of Chronic Diseases* **20** 233–239.

[16] Shumaker, S. A. (2003). 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* **289(20)** 2651–2662.

[17] Shumaker, S. A. (2004). 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* **291(24)** 2947–2958.

[18] Van der Vaart, A. W. (1998). *Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge. MR1652247 (2000c:62003)

[19] Vellas, B., Coley, N., Ousset, P. J., Berrut, G., Dartigues, J. F., Dubois, B., Grandjean, H., Pasquier, F., Piette, F., Robert, P., Touchon, J., Garnier, P., Mathiex-Fortunet, H. and Andrieu, S. f. (2012). Long-term use of standardised ginkgo biloba extract for the prevention of Alzheimer's disease (GuidAge): A randomised placebo-controlled trial. *Lancet Neurology*.

[20] Wimo, A. and Prince, M. (2012). World Alzheimer report 2012 Technical Report.

[21] Zucker, D. M. and Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77** 853–864. MR1086695 (92c:62053)