New late-emphasis and combination tests based on infimum and supremum logrank statistics with application in oncology trials

Jean Marie Boher, Thomas Filleron, Pierre Bunouf, Richard J Cook

Received: 15 May 2022 | Revised: 20 January 2023 | Accepted: 24 February 2023

DOI: 10.1002/sim.9709

RESEARCH ARTICLE

Statistics in Medicine WILEY

New late-emphasis and combination tests based on infimum and supremum logrank statistics with application in oncology trials

Jean Marie Boher[1,2] | **Thomas Filleron**[3] | **Pierre Bunouf**[4] | **Richard J. Cook**[5]

[1]Biostatistics and Methodology Unit, Institut Paoli-Calmettes, Marseille, France

[2]Aix Marseille Univ, INSERM, IRD, SESSTIM, Marseille, France

[3]Biostatistics Unit, Institut Claudius Regaud-IUCT-O, Toulouse, France

[4]Laboratoires Pierre Fabre, 3 ave Pierre Curie Toulouse, France

[5]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

Correspondence
Jean Marie Boher, Biostatistics and Methodology Unit, Institut Paoli-Calmettes, Marseille, France.
Email: boherjm@ipc.unicancer.fr

Present address
Jean Marie Boher, Département de la Recherche Clinique et de l'Innovation, Institut Paoli-Calmettes, 232 Bd Sainte Marguerite Cedex, 13009, France

Funding information
Canadian Institutes for Health Research, Grant/Award Number: FRN 13887; Ligue Contre le Cancer, Grant/Award Number: PRC2016.LCC/JMB; Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN 155849

Abstract

Immunotherapy cancer clinical trials routinely feature an initial period during which the treatment is given without evident therapeutic benefit, which may be followed by a period during which an effective therapy reduces the hazard for event occurrence. The nature of this treatment effect is incompatible with the proportional hazards assumption, which has prompted much work on the development of alternative effect measures of frameworks for testing. We consider tests based on individual and combination of early- and late-emphasis infimum and supremum logrank statistics, describe how they can be implemented, and evaluate their performance in simulation studies. Through this work and illustrative applications we conclude that this class of test statistics offers a new and powerful framework for assessing treatment effects in cancer clinical trials involving immunotherapies.

KEYWORDS

censored data, delayed treatment effect, immunotherapy cancer trial, supremum statistic

1 | INTRODUCTION

The proportional hazards model and associated logrank test are widely adopted for the analysis of censored data from randomized clinical trials (RCT) with time-to-event endpoints. While the logrank test is the most powerful test for detecting treatment effects within the class of proportional hazards (PH) models, the power can be greatly reduced

Abbreviations: PH, proportional hazards; PWPH, piecewise proportional hazards.

under nonproportional hazards relationship. This is the case in cancer immunotherapy randomized clinical trials (RCTs) where immunological responses can take weeks or months to manifest.[1,2] This observation has resulted in a considerable amount of recent work on methods for designing and analyzing clinical trials with potentially delayed treatment effects. Using evidence from reconstructed individual patient data, Fleming-Harrington (F-H) logrank tests with weight functions emphasizing late differences, along with their supremum versions, have been recommended to increase sensitivity for detecting drug effects on survival in cancer vaccine or immunotherapy oncology trials.[1,3] It has been noticed that late-emphasis logrank tests can lead to a considerable loss of power under non PH hazards other than delayed treatment effects.[4,5] To increase the sensitivity to PH alternatives and general non-PH alternatives, versatile tests based on the maximum of multiple F-H logrank test statistics have been proposed[1,6-8] and recently advocated as primary analysis in oncology trials.[9,10] In this paper we introduce new late-emphasis test statistics based on infimum and supremum logrank statistics obtained by successively deleting the first failure events. We also define combination test procedures based on the simultaneous use of early and late-emphasis infimum and supremum logrank statistics to increase the sensitivity to detect non-PH differences. According to our numerical studies, the proposed tests provide adequate test sizes under null differences and maintain empirical powers close to the power of the standard logrank test under PH assumptions. The new combination test procedure aimed at assessing the superiority of an experimental drug is found to be at least as sensitive to early and late survival differences as the versatile test called *MaxCombo* recently advocated in oncology trials.[10] When there are concerns about delayed treatment effects and/or general non-PH situations, we conclude that late-emphasis and combination test procedures based on infimum and supremum logrank statistics created by successively analyzing or deleting the first failures offer an alternative powerful strategy for assessing treatment effects in cancer clinical trials. The remainder of this paper is organized as follows. In Section 2 we define notation, the class of F-H logrank statistics along with the definition of maximum combination tests using multiple F-H logrank statistics. Section 3 examines the infimum and supremum versions of logrank test statistics, the new combination tests, some elements of large sample theory and computational guidelines on implementation. Numerical studies and illustrative applications in immunotherapy oncology trials are discussed in Section 4 and Section 5, respectively. The paper concludes with general remarks in Section 6.

2 | BACKGROUND

We consider the setting of a two-arm clinical trial in which n individuals are randomized to receive the experimental treatment or standard of care, with the aim of following them over $(0, \tau]$ where τ denotes the administrative censoring time. We let T_i be the time to the clinical endpoint and L_i the loss to follow-up time giving right censoring time $C_i = \min(L_i, \tau)$ for individual $i, i = 1, \ldots, n$. We define group 0 ($Z_i = 0$) as those assigned to receive standard care and group 1 ($Z_i = 1$) as those assigned to receive the experimental treatment. Then if $X_i = \min(T_i, C_i)$ and $\Delta_i = I(X_i \leq C_i)$, data from the full sample are given by $\{(X_i, \Delta_i, Z_i), i = 1, \ldots, n\}$. Let $S_k(t)$ be the survival function in group k, $S_k(t) = P(T_i \succ t | Z_i = k), k = 0, 1$. Our interest lies in two-sample nonparametric tests of the null hypothesis $H_0 : S_1(t) = S_0(t)$ for all t. In counting process notation we let $Y_i(t) = I(X_i \geq t)$ indicate that individual i is at risk (event-free and uncensored) at t, and let $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ indicate that the failure event occurred and was observed by time t; then $\overline{Y}_1(t) = \sum_{i=1}^n Y_i(t)Z_i$ and $\overline{N}_1(t) = \sum_{i=1}^n \overline{N}_i(t)Z_i$ denote the number of subjects at risk and the number of failures observed by time t in group 1, respectively; likewise $\overline{Y}_0(t) = \sum_{i=1}^n Y_i(t)(1 - Z_i)$ and $\overline{N}_0(t) = \sum_{i=1}^n \overline{N}_i(t)(1 - Z_i)$ and we write $\overline{Y}(t) = \sum_{i=1}^n Y_i(t)$ and $\overline{N}(t) = \sum_{i=1}^n \overline{N}_i(t)$. We let $n_0 = \sum_{i=1}^n (1 - Z_i)$ and $n_1 = \sum_{i=1}^n Z_i$ be the numbers of individuals assigned to group k, $k = 0, 1$ and $n = n_0 + n_1$. F-H logrank statistics commonly used when the goal is to detect differences between two survival distributions over a particular period can be expressed generally as,

$$S_{\rho,\gamma} = \int_0^\tau W(t) \left[d\overline{N}_1(t) - \overline{Y}_1(t) \frac{d\overline{N}(t)}{\overline{Y}(t)} \right],$$

with $W(t) = \left[\hat{G}(t-)\right]^\rho \left[1 - \hat{G}(t-)\right]^\gamma, \rho \geq 0, \gamma \geq 0$ and $\hat{G}(t)$ denotes the Kaplan-Meier estimate of the survival function in the pooled sample.[11] In order to achieve optimum power, the predetermined weight function $W(t)$ should be proportional to $\beta(t) = \log(\lambda_1(t)) - \log(\lambda_0(t))$, where $\lambda_0(t)$, $\lambda_1(t)$ are the hazard rate functions in each group, $k = 0, 1$.[5,12] Setting $\rho = 0$ and $\gamma = 0$, $S_{0,0}$ reduces to the logrank statistic which has optimum power under proportional hazards alternatives. Versatile test procedures based on the simultaneous use of multiple standardized F-H logrank test statistics $FH^{\gamma,\rho}$ have been proposed to achieve robust performance under various type of non-PH alternatives.[7,8,10] Lee[7] showed that a maximum

combination test based on an early-emphasis logrank test $FH^{1,0}$ and a late-emphasis logrank test $FH^{0,1}$ is nearly as sensitive as its individual components in detecting early and late survival differences. More recently, the *MaxCombo* test based on the combination of the F-H test statistics, $FH^{0,0}$, $FH^{1,0}$, $FH^{0,1}$ and $FH^{1,1}$, have been recommended in oncology trials to provide robust performance under more general alternatives[8,10].[10] It has also been shown that the combination of $FH^{0,0}$ with a sequence of piecewise weighted logrank test statistics created after successively deleting events during an initial period provide robust performance under PH and general delayed treatment alternatives.[13-15] Since all these combination tests follow an asymptotic multivariate normal approximation with independent increments covariance structure, the *P*-values can be obtained using algorithms originally designed for group sequential testing.

3 | LATE-EMPHASIS AND COMBINATION TESTS BASED ON INFIMUM AND SUPREMUM LOGRANK STATISTICS

3.1 | Late-emphasis infimum and supremum logrank statistics

Denote by H_0 the null hypothesis to be tested, $H_0 : S_1(t) = S_0(t)$ for all $t \in (0, \tau]$, $S_{\gamma,\rho}(t)$ the weighted logrank statistics evaluated by time t

$$S_{\gamma,\rho}(t) = \int_0^t W(t) \left[d\overline{N}_1(t) - \overline{Y}_1(t) \frac{d\overline{N}(t)}{\overline{Y}(t)} \right],$$

and $\hat{\sigma}^2_{\gamma,\rho}$ the variance estimator for weighted logrank statistics

$$\hat{\sigma}^2_{\gamma,\rho} = \int_0^\tau W^2(t) \frac{Y_0(t)Y_1(t)}{Y(t)} \left(\frac{Y(t) - dN(t)}{Y - 1} \right) \left(\frac{dN_1(t)}{Y_1(t)} - \frac{dN_0(t)}{Y_0(t)} \right).$$

Gill,[16] Fleming et al[17] have shown the superiority of Renyi-type supremum logrank statistics $\hat{\sigma}^{-1}_{\gamma,\rho} \sup_{t\geq 0} |S_{\gamma,\rho}(t)|$ to the traditional logrank test in a variety of non PH settings, including early differences in survival. Denote by $\widetilde{S}_{0,0}(t)$ the sequence of logrank type statistics obtained by successively deleting the failure events prior to time t,

$$\widetilde{S}_{0,0}(t) = \int_t^\tau \left[d\overline{N}_1(t) - \overline{Y}_1(t) \frac{d\overline{N}(t)}{\overline{Y}(t)} \right], \quad t \geq 0.$$

To give more emphasis to late survival differences, we propose to test for H_0 vs the upper alternative $H_1 : S_1(t) \succ S_0(t)$ using the infimum statistic $\inf_{t\in(0,\tau]} \widetilde{S}_{0,0}(t)$ and to test for H_0 vs the lower alternative $H_1 : S_1(t) \prec S_0(t)$ using the supremum statistic $\sup_{t\in(0,\tau]} \widetilde{S}_{0,0}(t)$. In the following these statistics will be referred to as the late-emphasis infimum logrank statistic (*le-Inf*) and the late-emphasis supremum logrank statistic (*le-Sup*). Given $\widetilde{S}_{0,0}(t) = S_{0,0} - S_{0,0}(t)$, it follows from the standard martingale representation given in (1) that $\widetilde{S}_{0,0}(t)$ is asymptotically equivalent under H_0 to the sum of martingales

$$\sum_{i=1}^n \int_t^\tau \left[Z_i - \frac{y_1(u)}{y(u)} \right] dM_i(u), \quad t \geq 0,$$

where $y_k(t) = \lim_{n_k\to\infty} n_k^{-1}\overline{Y}_k(t)$, $y(t) = y_0(t) + y_1(t)$ and $dM_i(t) = dN_i(t) - Y_i(t)\lambda(t)dt$. As a consequence of the central limit theorems for martingales, the process $n^{-\frac{1}{2}}\widetilde{S}_{0,0}(t)$ converges weakly to a zero-mean Gaussian process when $n \to \infty$. Following Lin et al,[18] we can approximate *P*-values as the sample proportions

$$\frac{1}{M}\sum_{m=1}^M I\left(\inf_{t\in(0,\tau]} \widetilde{S}^m(t) \leq \textit{le-Inf} \right) \text{ and } \frac{1}{M}\sum_{m=1}^M I\left(\sup_{t\in(0,\tau]} \widetilde{S}^m(t) \geq \textit{le-Sup} \right)$$

where $\widetilde{S}^m(t) = \sum_{i=1}^n I(t \leq X_i)\Delta_i U_i^m \left[Z_i - \frac{n_1^{-1}\overline{Y}_1(X_i)}{n^{-1}\overline{Y}(X_i)} \right]$ are independent realizations obtained under the null by sampling random normal deviates U_i^m, $m = 1, \ldots, M, i = 1, \ldots, n$, and replacing martingale increments $dM_i(t)$ and limiting values $y_1(t)$, $y(t)$ by randomly perturbed terms $U_i^m dN_i(t)$ and consistent sample estimates $n_1^{-1}\overline{Y}_1(t)$, $n^{-1}\overline{Y}(t)$, respectively.

Similarly, two-sided *P*-values are derived as sample proportions

$$\frac{1}{M}\sum_{m=1}^{M} I\left(\sup_{t\in(0,\tau]}\left|\widetilde{S}^m(t)\right| \geq |le\text{-}Sup|\right).$$

where $|le\text{-}Sup| = \sup_{t\in(0,\tau]}\left|\overline{S}_{0,0}(t)\right|$.

3.2 | Combination test using extreme values of logrank statistics

It is shown that logrank statistics $S_{0,0}(t)$ evaluated at time t are asymptotically equivalent under H_0 to the sums[19]

$$\sum_{i=1}^{n}\int_0^t \left[Z_i - \frac{y_1(u)}{y(u)}\right] dM_i(u), \quad t \geq 0. \tag{1}$$

Let $Inf = \inf_{t\in(0,\tau]} S_{0,0}(t)$ and $Sup = \sup_{t\in(0,\tau]} S_{0,0}(t)$. To achieve robust power performance under general alternatives, including PH, early and late survival differences, we propose to test for H_0 vs one-sided alternatives, $H_1 : S_1(t) \succ S_0(t)$ or $H_1 : S_1(t) \prec S_0(t)$, for some $t \in (0, \tau]$, using the combination test statistics $Combo\text{-}Inf = \min(Inf, le\text{-}Inf)$ and $Combo\text{-}Sup = \max(Sup, le\text{-}Sup)$, respectively. Again, one-sided *P*-values can be approximated using sample proportions

$$M^{-1}\sum_{m=1}^{M} I\left(\min(\inf_{t\in(0,\tau]} S^m(t), \inf_{t\in(0,\tau]}\widetilde{S}^m(t)) \leq Combo\text{-}Inf\right),$$

or

$$M^{-1}\sum_{m=1}^{M} I\left(\max(\sup_{t\in(0,\tau]} S^m(t), \sup_{t\in(0,\tau]}\widetilde{S}^m(t)) \geq Combo\text{-}Sup\right),$$

where $S^m(t) = \sum_{i=1}^{n} I(X_i \leq t)\Delta_i U_i^m \left[Z_i - \frac{n_1^{-1}\overline{Y}_1(X_i)}{n^{-1}\overline{Y}(X_i)}\right], m = 1, \ldots, M$. A two-sided versatile test for H_0 is defined similarly using the maximum combination statistic $Combo\text{-}|Sup| = \max\left(\sup_{t\geq 0}\left|S_{0,0}(t)\right|, \sup_{t\geq 0}\left|\overline{S}_{0,0}(t)\right|\right)$. As before, the probability of obtaining results at least as extreme as the observed results is approximated using the empirical distribution function estimated from a sampling distribution $\left\{\max\left(Sup_{t\in(0,\tau]}\left|S^m(t)\right|, Sup_{t\in(0,\tau]}\left|\widetilde{S}^m(t)\right|\right); m = 1, \ldots, M\right\}$.

3.3 | Covariate-adjusted approach

It is often of interest to adjust the results of comparisons between groups on the individual values of a vector of certain key individual covariates, noted W_i. Kong and Slud[20] developed a robust approach using the covariate-adjusted partial score statistic

$$S_\theta(0, \hat{\beta}) = \frac{\partial}{\partial\theta}\log L(\theta, \beta)_{|\theta=0,\beta=\hat{\beta}},$$

where $L(\theta, \beta)$ denotes the Cox's partial likelihood function derived under a general Cox's proportional hazards model,

$$\lambda(t\,|Z_i, W_i) = \lambda_0(t)\exp(\theta Z_i)h(t, \beta, W_i),$$

and $\hat{\beta}$ the restricted maximum partial likelihood estimator obtained under $H_0 : \theta = 0$. Define for any given $t > 0$, the following partial likelihood score processes,

$$S_\theta(t, \theta, \beta) = \frac{\partial}{\partial\theta}\log L(t, \theta, \beta), \quad \widetilde{S}_\theta(t, \theta, \beta) = \frac{\partial}{\partial\theta}\log \widetilde{L}(t, \theta, \beta),$$

with $L(t,\theta,\beta)=\prod_{i:T_i\le t}L_i(T_i,\theta,\beta)^{\Delta_i}$, $\widetilde{L}(t,\theta,\beta)=\prod_{i:T_i\ge t}L_i(T_i,\theta,\beta)^{\Delta_i}$ and $L_i(t,\theta,\beta)=\frac{\exp(\theta Z_i)h(t,\beta,W_i)}{\sum_{j=1}^{n}Y_j(t)\exp(\theta Z_j)h(t,\beta,W_j)}$. It can be shown using first-order approximation[21] and martingale representations for score process[22] that the covariate-adjusted partial score processes evaluated at $\theta=0$ and $\beta=\hat{\beta}$, $S_\theta(t,0,\hat{\beta})$ and $\widetilde{S}_\theta(t,0,\hat{\beta})$, are asymptotically equivalent under H_0 to the following sums (see Appendix)

$$\sum_{i=1}^{n}\left(\int_0^t\left[Z_i-\frac{z^{(1)}(u,0,\beta_0)}{y(u,0,\beta_0)}\right]dM_i(u,0,\beta_0)-i_{\beta,\theta}(t,0,\beta_0)i_{\beta,\beta}^{-1}(\beta_0)\int_0^\tau\left[W_i-\frac{w^{(1)}(u,0,\beta)}{y(u,0,\beta)}\right]dM_i(u,0,\beta_0)\right), \tag{2}$$

and

$$\sum_{i=1}^{n}\left(\int_t^\tau\left[Z_i-\frac{z^{(1)}(u,0,\beta_0)}{y(u,0,\beta_0)}\right]dM_i(u,0,\beta_0)-\tilde{\zeta}_{\beta,\theta}(t,0,\beta_0)\tilde{\zeta}_{\beta,\beta}^{-1}(\beta_0)\int_0^\tau\left[W_i-\frac{w^{(1)}(u,0,\beta)}{y(u,0,\beta)}\right]dM_i(u,0,\beta_0)\right), \tag{3}$$

where $\beta_0=\lim_{n\to\infty}\hat{\beta}$ and for any $t\ge 0$

$$dM_i(t,\theta,\beta)=dN_i(t)-Y_i(t)h(t,\theta,\beta,W_i)\lambda_0(t)dt,$$

$$z^{(1)}(t,\theta,\beta)=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}Y_i(t)Z_ih(t,\theta,\beta,W_i),\ w^{(1)}(t,\theta,\beta)=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}Y_i(t)W_ih(t,\theta,\beta,W_i),$$

$$y(t,\theta,\beta)=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}Y_i(t)h(t,\theta,\beta,W_i),$$

$$i_{\beta,\theta}(t,\theta,\beta)=-\lim_{n\to\infty}n^{-1}\frac{\partial^2}{\partial\beta\partial\theta}\log L(t,\theta,\beta),\quad \tilde{\zeta}_{\beta,\theta}(t,\beta_0)=-\lim_{n\to\infty}n^{-1}\frac{\partial^2}{\partial\beta\partial\theta}\log\widetilde{L}(t,\theta,\beta).$$

,

$$i_{\beta,\beta}(\beta_0)=-\lim_{n\to\infty}n^{-1}\frac{\partial^2}{\partial^2\beta}\log L(\tau,0,\beta)_{|\beta=\beta_0},\quad \tilde{\zeta}_{\beta,\beta}(\beta_0)=-\lim_{n\to\infty}n^{-1}\frac{\partial^2}{\partial^2\beta}\log\widetilde{L}(\tau,0,\beta)_{|\beta=\beta_0}.$$

New early-emphasis, late-emphasis and versatile combination test statistics adjusted for individual covariates are easily obtained using the supremum or infimum values of covariate-adjusted score statistics $S_\theta(t,0,\beta)$ and $\widetilde{S}_\theta(t,0,\beta)$. Again, the *P*-values are approximated using a sampling distribution of each test statistics obtained by replacing the martingale increments $dM_i(t,\beta_0)$ by randomly perturbed terms $U_{mi}dN_i(t)$ and the unknown limiting values $y(t,\beta)$, $z^{(1)}(t,\beta)$, $w^{(1)}(t,\beta)$, $i_{\beta,\beta}(\beta_0)$ with consistent sample estimates.

4 | SIMULATION STUDIES

Numerical studies were conducted to evaluate the performance characteristics of the proposed early and late individual test statistics and combined test statistics, by repeatedly sampling patients at a uniform rate over 12 months in a control arm ($Z=0$) and an experimental arm ($Z=1$) with equal probability. Let denote by $W_i, i=1,\ldots,n$ an individual binary risk factor. Individual survival data were drawn from exponential distributions with a constant hazard rate $\lambda_0\exp(\beta W_i)$ in the control arm, exponential distributions with a constant hazard rate $\lambda_0\exp(\theta+\beta W_i)$ or two-piece exponential distributions with constant hazard rates $\lambda_1(t)=\lambda_0\exp(\theta_1+\beta W_i)$ if $t\le t_0$ and $\lambda_1(t)=\lambda_0\exp(\theta_2+\beta W_i)$ otherwise in the experimental arm. Survival data were censored to the right at the end of study scheduled 5 years after the start of the study. Parameter values used in simulation studies are presented in Table 1. Four different scenarios illustrated in Figure 1 were investigated: PH differences including null differences, early and delayed differences, and "strong" null hypotheses.[23] Empirical error rates defined as the proportion of times the null hypothesis were rejected at the nominal level of 5% are reported in Tables 2,3. A total of 10 000 trials with moderate sizes were simulated under the null hypothesis. Otherwise, sample sizes were selected to achieve near 90% power with the most efficient test and sampling was repeated 5000 times.

Under null differences, two-sided supremum logrank statistics |*Sup*| and |*le-Sup*| maintain an experimentwise type I error control (Table 2, scenarios S1 and S2). The same is true for the maximum combination test statistic *Combo*-|*Sup*|

TABLE 1 Model parameters used in simulation experiments.

Scenario	Change point in hazard ratio	Treatment hazard ratio	Covariate hazard ratio	Baseline hazard	Censoring rate (%)	*N*	Test type
Null hypothesis and PH model							
S1	$t_0 = 0$	1.0	1.0	0.030	24	100,150	Two-sided
S2	$t_0 = 0$	1.0	2.0	0.022	24	100,150	Two-sided
S3	$t_0 = 0$	0.5	1.0	0.045	24	150	Two-sided
S4	$t_0 = 0$	0.5	2.0	0.030	26	150	Two-sided
S5	$t_0 = 0$	2.0	1.0	0.020	27	150	Two-sided
S6	$t_0 = 0$	2.0	2.0	0.015	26	150	Two-sided
Delayed treatment effect model							
S7	$t_0 = 6$	(1.0,0.5)	1.0	0.032	32	180	One-sided
S8	$t_0 = 9$	(1.0,0.5)	1.0	0.032	31	205	One-sided
S9	$t_0 = 12$	(1.0,0.5)	1.0	0.032	30	240	One-sided
Early treatment effect model							
S10	$t_0 = 6$	(0.5,1.0)	1.0	0.078	3	300	One-sided
S11	$t_0 = 9$	(0.5,1.0)	1.0	0.078	3	225	One-sided
S12	$t_0 = 12$	(0.5,1.0)	1.0	0.078	4	180	One-sided
"Strong" null hypothesis							
S13	$t_0 = 6$	(3.0,0.75)	1.0	0.04	15	100, 200, 400	One-sided
S14	$t_0 = 9$	(2.5,0.75)	1.0	0.04	14	100, 200, 400	One-sided
S15	$t_0 = 12$	(2.0,0.75)	1.0	0.04	14	100, 200, 400	One-sided

which reported empirical rejection rates similar to those obtained with the standard logrank test statistic $|FH^{0,0}|$. As expected under PH differences,[17] these test statistics are nearly as sensitive as the optimal logrank test statistic (Table 2, scenarios S3-S6). Same conclusions hold for one-sided upper tests under unequal allocation across treatment arms (see Appendix, Table A1). The experimentwise error rates reported in Table 3 illustrate the poor performance of the logrank test statistic under nonproportional hazards situations. Compared to one-sided superiority weighted logrank test statistics, the late-emphasis infimum logrank (*le-Inf*) and infimum logrank (*Inf*) test statistics, respectively, increase the chances to detect late survival differences (Table 3, S7-S9) and early survival differences (Table 3, S10-S12).

The proposed combination superiority test statistic *Combo-Inf* is shown to be nearly as sensitive as the individual optimal component *le-Inf* or *Inf* in detecting late and early survival differences. Compared to the reference *MaxCombo* test, this new combination statistic provides similar and robust power under late and early differences in survival (Table 3, S7-S12), while maintaining experimentwise rejection rates close to error rates obtained with $FH^{0,0}$ under PH (data not shown). Finally, we evaluate the performance of individual test statistics and versatile combination test statistics under strong null hypotheses where survival rates are always below the control arm (Figure 1D). The experimentwise error rates of the *le-Inf* test statistic clearly increase with sample size and reach error rate levels well above the nominal level of statistical significance of 5%. The same is true for the minimum combination test *Combo-Inf*, but to a lesser extent (Table 3, S13-S15). As the logrank test statistic $FH^{0,1}$ putting more weight to late events, these two statistics do not properly account for the existence of early harm and lead to an increased risk of incorrectly concluding that an investigational drug has better efficacy. To limit this risk, a pragmatic approach consists in simultaneously evaluating on the same data the possibility of a beneficial and harmful efficacy of the experimental drug. For example, limiting the investigational drug's efficacy claim to situations where the supremum test statistic (*Sup*) detects no harmful efficacy and the *le-Inf* or the *Combo-Inf* test statistics meet the level of statistical evidence, both at the predefined significance level of 5%, helps to reduce the experimentwise error rates below 0.1% in all scenarios tested, irrespective of sample size. When there are concerns about delayed treatment effects and/or general non-PH situations, our results suggest that the proposed procedures based on individual or combination of infimum and supremum logrank statistics created by successively

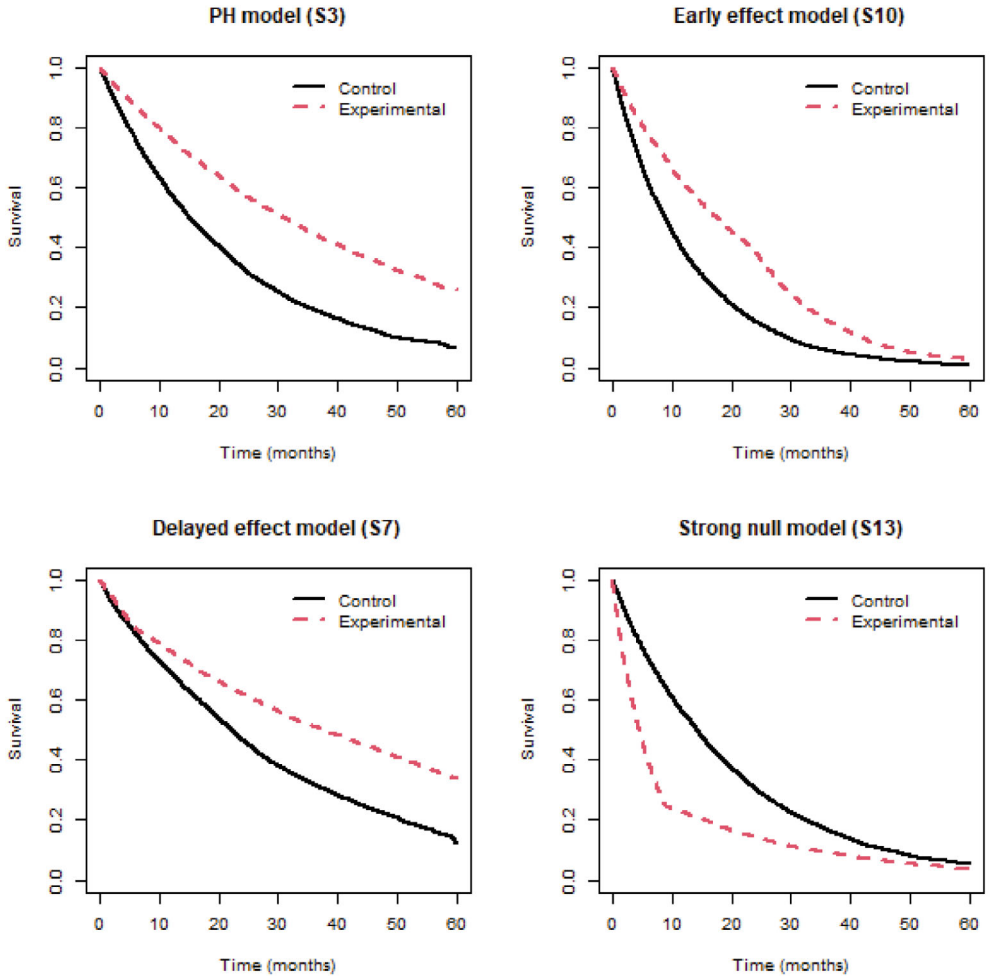

FIGURE 1 Simulation study S3, S7, S10, and S13 scenarios.

TABLE 2 Empirical sizes and powers of two-sided test statistics based upon supremum logrank statistics under proportional hazard (PH) models.

				Empirical error rates (%)							
	Hazard ratio for			Unadjusted test statistics				Adjusted test statistics			
Scenario	Treatment	Covariate	*N*	$\lvert FH^{0,0}\rvert$	$\lvert Sup\rvert$	$\lvert le\text{-}Sup\rvert$	*Combo-*$\lvert Sup\rvert$	$\lvert FH^{0,0}\rvert$	$\lvert Sup\rvert$	$\lvert le\text{-}Sup\rvert$	*Combo-*$\lvert Sup\rvert$
S1	1.0	1.0	100	5.2	5.0	4.7	4.9	5.1	5.1	4.8	4.8
	1.0	1.0	150	5.2	4.9	5.0	4.9	4.9	4.6	4.9	4.7
S2	1.0	2.0	100	5.0	4.8	4.7	4.8	5.2	4.8	4.7	4.6
	1.0	2.0	150	4.7	4.7	4.7	4.6	5.0	4.8	4.6	4.6
S3	0.5	1.0	150	96.0	94.5	94.7	94.4	96.1	94.9	95.0	94.7
S4	0.5	2.0	150	92.4	91.0	90.4	90.6	95.8	94.4	94.4	94.0
S5	2.0	1.0	150	94.9	93.8	94.0	93.7	95.0	93.7	93.4	93.3
S6	2.0	2.0	150	92.9	91.9	91.3	91.4	94.9	93.8	93.7	93.5

Note: Two-sided test statistics: $|Sup| = \sup_{t\in(0,\tau]} |S_{0,0}(t)|$, $|le\text{-}Sup| = \sup_{t\in(0,\tau]} \left|\hat{S}_{0,0}(t)\right|$ and *Combo-* $|Sup| = \max(|Sup|, |le\text{-}Sup|)$.

TABLE 3 Empirical power of one-sided upper test statistics under late differences, early differences, and strong null hypotheses.

			Empirical type II error rates (%)							
			Individual test statistics						Combination test statistics	
Scenario	**Change point**	***N***	**$FH^{0,0}$**	**$FH^{1,0}$**	**$FH^{0,1}$**	**$FH^{1,1}$**	***Inf***	***le-Inf***	***MaxCombo-Min***	***Combo-Inf***
Delayed effect model										
S7	$t_0 = 6$	180	87.2	74.5	92.4	93.6	83.2	92.2	90.7	91.3
S8	$t_0 = 9$	205	83.1	63.5	92.1	91.8	77.3	90.8	88.8	90.0
S9	$t_0 = 12$	240	76.8	53.0	92.0	88.8	70.0	88.8	85.8	88.3
Early effect model										
S10	$t_0 = 6$	300	54.2	81.1	13.2	30.0	72.8	45.5	67.3	70.8
S11	$t_0 = 9$	225	67.4	88.3	20.1	49.8	83.2	59.5	79.0	80.5
S12	$t_0 = 12$	180	74.9	90.3	28.2	65.1	87.2	68.3	83.8	84.3
Strong null hypotheses										
S13	$t_0 = 6$	100	< 0.1	< 0.1	6.2	1.3	< 0.1	7.5	5.6	3.5
		200	< 0.1	< 0.1	7.6	< 0.1	< 0.1	15.2	11.9	4.4
		400	< 0.1	< 0.1	7.7	< 0.1	< 0.1	30.9	25.3	4.2
S14	$t_0 = 9$	100	< 0.1	< 0.1	3.7	< 0.1	< 0.1	3.9	2.7	2.1
		200	< 0.1	< 0.1	2.7	< 0.1	< 0.1	7.2	4.7	1.1
		400	< 0.1	< 0.1	2.2	< 0.1	< 0.1	17.3	12.6	< 0.1
S15	$t_0 = 12$	100	< 0.1	< 0.1	3.6	< 0.1	< 0.1	3.2	1.9	1.8
		200	< 0.1	< 0.1	3.1	< 0.1	< 0.1	5.5	3.8	1.5
		400	< 0.1	< 0.1	2.0	< 0.1	< 0.1	12.4	8.7	1.0

Note: One-sided upper test statistics: $Inf = \inf_{t\in(0,\tau]} S_{0,0}(t)$, $le\text{-}Inf = \inf_{t\in(0,\tau]} \hat{S}_{0,0}(t)$, $MaxCombo\text{-}Min = \min(FH^{0,0}, FH^{1,0}, FH^{0,1}, FH^{1,1})$ and $Combo\text{-}Inf = \min(Inf, le\text{-}Inf)$.

analyzing or deleting the first failures offer an alternative powerful strategy for assessing treatment effects in cancer clinical trials. Empirical sizes and powers of test statistics based upon supremum logrank statistics under proportional hazards (PH) models and unequal allocation across arms (1:2) are rep.

5 | ILLUSTRATIVE APPLICATIONS

We reanalyzed data from two large RCTs assessing the benefit of combination of an experimental immunotherapy drug with standard chemotherapy vs standard chemotherapy alone. The individual patient data were reconstructed using digitizing software[24] from Kaplan-Meier progression-free survival (PFS) curves originally published for the treatment arms.[25,26] To illustrate the sensitivity of the different methods in settings of moderate sample size, we assess the efficacy of the investigational drug compared to standard chemotherapy regimen using one-tailed tests of superiority and inferiority in three randomly selected sub-samples of each study dataset (see Table 4).

5.1 | Pacific trial (NCT02125461)

A total of 713 patients with stage III nonsmall-cell lung cancer were randomized in this trial to receive either Durvalumab or placebo after standard chemotherapy, including 473 and 237 patients assigned to receive Durvalumab and placebo, respectively. The Kaplan-Meier estimates for treatment arms plotted in Figure 2 show a benefit for the experimental arm in PFS data after a short lag time. We splitted the dataset into five subsamples from a multinomial distribution with probability mass $P = (1/5, \ldots, 1/5)$. One-sided test results assessing the efficacy of Durvalumab vs standard chemotherapy

TABLE 4 *P*-values of test statistics of experimental arm vs control arm in subsamples.

	Pacific trial			Bellmunt trial		
	subsample # 1	Subsample # 2	Subsample # 3	Subsample # 1	Subsample # 2	Subsample # 3
Test statistic	$n = 147$	$n = 145$	$n = 142$	$n = 174$	$n = 185$	$n = 182$
F-H's weighted logrank test						
$FH^{0,0}$	0.005(0.995)	0.001(0.999)	0.037(0.963)	0.785(0.215)	0.369(0.631)	0.360(0.640)
$FH^{1,0}$	0.010(0.990)	0.002(0.998)	0.052(0.948)	0.995(0.005)	0.722(0.278)	0.940(0.060)
$FH^{0,1}$	0.004(0.996)	0.005(0.995)	0.041(0.959)	0.025(0.975)	0.079(0.921)	0.001(0.999)
$FH^{1,1}$	0.003(0.997)	0.002(0.998)	0.025(0.975)	0.475(0.529)	0.173(0.827)	0.048(0.952)
Infimum (supremum) logrank test statistics						
Inf(*Sup*)	0.012(0.594)	0.004(0.708)	0.076(0.332)	0.897(0.005)	0.602(0.117)	0.653(0.020)
le-Inf(*le-Sup*)	0.007(0.856)	0.002(0.847)	0.008(0.867)	0.035(0.372)	0.041(0.879)	0.004(0.918)
Versatile tests						
MaxCombo-Min(Max)	0.005(0.997)	0.002(0.999)	0.045(0.981)	0.051(0.047)	0.148(0.430)	0.002(0.110)
Combo-Inf(*Sup*)	0.008(0.755)	0.003(0.873)	0.013(0.434)	0.044(0.006)	0.056(0.138)	0.006(0.024)

Notes: One-sided upper test statistics: $Inf = \inf_{t\in(0,\tau]} S_{0,0}(t)$, $le\text{-}Inf = \inf_{t\in(0,\tau]} \hat{S}_{0,0}(t)$, $MaxCombo\text{-}Min = \min(FH^{0,0}, FH^{1,0}, FH^{0,1}, FH^{1,1})$ and $Combo\text{-}Inf = \min(Inf, le\text{-}Inf)$. One-sided lower test statistics: $Sup = \sup_{t\in(0,\tau]} S_{0,0}(t)$, $le\text{-}Sup = \sup_{t\in(0,\tau]} \hat{S}_{0,0)}(t)$, $MaxCombo\text{-}Max = \max(FH^{0,0}, FH^{1,0}, FH^{0,1}, FH^{1,1})$ and $Combo\text{-}Sup = \max(Sup, le\text{-}Sup)$.

in the first three subsamples are reported in Table 4. The logrank tests $FH^{0,0}$, $FH^{0,1}$and $FH^{1,1}$ and the late emphasis test statistic *le-Inf* evaluating the superiority of the Durvalumab arm over the control arm consistently reject the null hypothesis in favor of the experimental arm in all subsamples analyzed. The *le-Inf* statistic yields the lowest *P*-values, which confirms its good performance observed in simulations under PH or delayed treatment models. Among versatile tests assessing the superiority of Durvalumab, only the *Combo-Inf* test statistic detects significant differences at the 5% nominal level in all subsamples analyzed. All tests assessing the superiority of the control arm over the experimental arm show no significant difference in favor of the standard treatment in any subsample.

5.2 | Bellmunt randomized trial (NCT02256436)

A total of 542 patients with advanced urothelial cancer were equally randomized in this trial to receive Pembrolizumab or the investigator's choice of chemotherapy. The Kaplan-Meier estimates for treatment arms plotted in Figure 3 show a cross in progression-free survival with higher survival rates for the experimental arm after 6 months. The patient data were splitted in three subsamples with equal probability. Among the one-sided weighted logrank tests, only $FH^{1,0}$ and $FH^{0,1}$ detect significant or borderline significant differences in two of three subsamples, respectively, in favor of Pembrolizumab or the standard arm. The proposed late-emphasis test statistic *le-Inf* assessing the superiority of Pembrolizumab show significant differences at the 5% nominal level in all subsamples, while the supremum logrank statistic *Sup* confirms harmful drug efficacy for Pembrolizumab in two subsamples. Versatile test statistics lead to similar conclusions with lower *P* values obtained by combining infimum or supremum versions of logrank statistics. These results reflect the lack of an overall clinical benefit for the experimental drug, as shown in Figure 3, where the harmful effect of the experimental precludes the claim of its efficacy. These results confirms the good performance of our proposed versatile test procedure observed in simulations under various non-PH.

6 | DISCUSSION

In recent years, a considerable attention has been given to the challenge of designing and analyzing clinical trials with potentially delayed treatment effect. This is the case in clinical trials of cancer immunotherapy where time-lag treatment

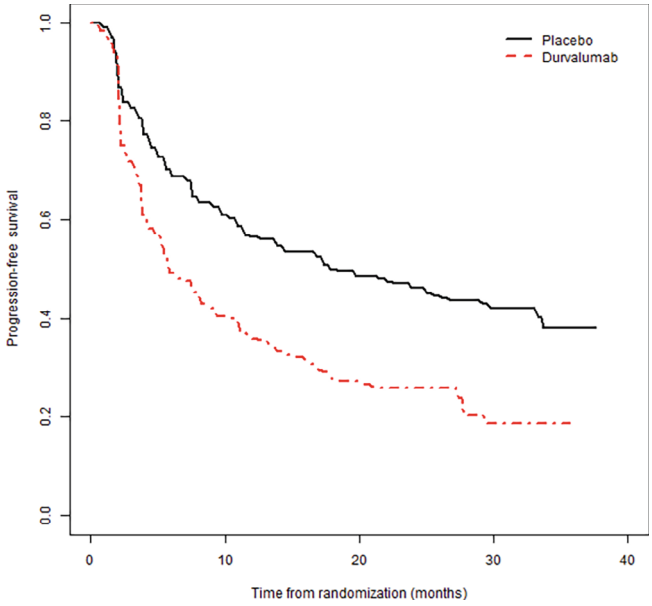

FIGURE 2 Pacific cancer trial—Progression-free survival rates for Durvalumab vs standard chemotherapy reproduced from the original published figure.

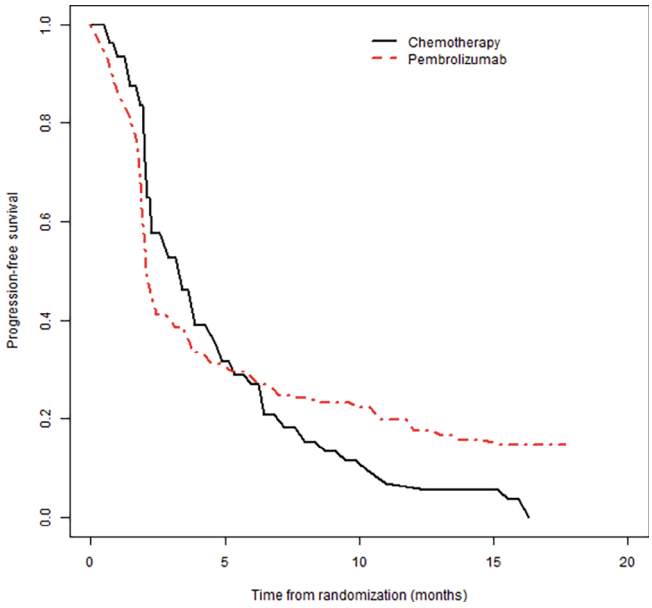

FIGURE 3 Bellmunt cancer trial–Progression-free survival rates for Pembrolizumab vs standard chemotherapy reproduced from the original published figure.

effects have been observed. The use of late-emphasis weighted logrank tests and their supremum versions as primary testing strategy at the design and analysis stages have been suggested in the literature as an alternative to the standard logrank test and/or the hazard ratio estimation.[1,3,27] Friede and Korn have noted that giving more weight to later events can result in considerable power loss if the PH assumptions are satisfied or nearly satisfied.[4] Here we study late emphasis infimum/supremum versions of two-sample logrank statistics obtained by successively deleting the first failure events for evaluating drug differences in survival. Empirical results show that these test statistics increase the chances to detect late survival differences while being nearly as sensitive as the logrank test under PH assumptions. The main drawback of this approach is that down-weighting early events does not properly account for existence of early harm, leading to an increase risk of erroneously concluding in favor of the experimental drug for an inferior drug. Magirr and Burman[23] have

proposed a class of modestly weighted logrank tests that allow down-weighting early events while controlling the type I error risk below the nominal level of significance. Instead we propose a pragmatic approach consisting in simultaneously evaluating on the same data the possibility of a beneficial and harmful efficacy of the experimental drug. For instance, we show that limiting the efficacy claim of the investigational drug when the supremum logrank statistic detects no harmful efficacy and the late-emphasis infimum logrank statistics meet the level of statistical evidence help to considerably reduce the error risk below the nominal level of significance. To provide robust performance under general non PH alternatives, some versatile test procedures that combine multiple weighted logrank tests have been proposed.[7,8,10] In particular, the *MaxCombo* test procedure used as a reference method have been recently recommended in the field of oncology trials as it provides robust power under PH, early and late survival differences or crossing survival curves alternatives. We show how to combine through an efficient Monte Carlo algorithm early and late-emphasis infimum/supremum logrank statistics to yield robust power against different alternatives of interest. Our numerical studies show that the evaluated combination tests provide adequate test sizes under null differences and maintain empirical powers close to the power of the standard logrank test under PH assumptions. Compared to the *MaxCombo* versatile test procedure, the *Combo-Inf* test statistic assessing the superiority of an investigational drug was found to be at least as sensitive under general alternatives, including PH, early and late survival differences. When there are concerns about delayed treatment effects and/or general non PH situations, late-emphasis or combination test procedures based on infimum and supremum versions of logrank statistics created by successively analyzing or deleting the first failure events offers an alternative powerful framework for assessing treatment effects in cancer clinical trials. Eng et al[28] have shown how to increase the sample size for supremum logrank tests to preserve the power under PH assumptions relative to the optimal logrank test. Although sample size formulas have been proposed for the weighted optimal logrank test in delayed processing effect models,[14,27] no existing method for determining sample size allow to preserve the power of the supremum versions of logrank test statistics compared to the optimal weighted logrank test. Following some recommendations, extensive simulation studies have to be performed to determine the sample size needed for the late-emphasis infimum and supremum logrank tests or the more versatile combination test procedure to achieve the desirable power under delayed treatment effect models and more complex non proportional hazards scenarios.[29] The lack of sequential test procedure for the proposed late-emphasis and more versatile combination test statistics allowing for early rejection of the null hypothesis represents a limitation to the implementation of the procedure in future RCTs. A limitation to this work is the lack of appropriate measure to summarize the treatment effect size. The definition of an appropriate summary measure of the effect size under non PH remains a key challenge in RCTs.[30]

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for their valuable comments and suggestions on the manuscript. This work was supported by the *Comité départemental de La Ligue Contre le Cancer de l'Aisne* and the *Comité départemental de La Ligue Contre le Cancer du Doubs Montbélliard*. Jean-Marie Boher and Thomas Filleron were supported by a research grant from the French Ligue Nationale Contre le Cancer (Projet de Recherche Clinique Ligue 2016, project PRC2016.LCC/JMB). Richard Cook was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849), and the Canadian Institutes for Health Research (FRN 13887).

DATA AVAILABILITY STATEMENT

Individual data reconstructed using digitizing software from published Kaplan-Meier curves are available from the corresponding author upon reasonable request. Functions to simulate and analyze two-sample trial data are available in the Data S1.

ORCID

Jean Marie Boher https://orcid.org/0000-0002-5395-839X
Richard J. Cook https://orcid.org/0000-0002-1414-4908

REFERENCES

1. Fine GD. Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Inf J*. 2007;4(4):535-539.
2. Mick R, Chen TT. Statistical challenges in the design of late-stage cancer immunotheray studies. *Cancer Immunol Res*. 2015;3(12):1292-1298.
3. Castañon E, Sanchez-Arraez A, Alvarez-Manceñido F, Jimenez-Fonseca P, Carmona-Bayonas A. Critical reappraisal of phase III trials with immune checkpoint inhibitors in non-proportional hazards settings. *Eur J Cancer*. 2020;136:159-168.

4. Freidlin B, Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *J Clin Oncol*. 2019;37(35):3455-3459.
5. Lin RS, León LF. Estimation of treatment effects in weighted log-rank tests. *Contemp Clin Trials Commun*. 2017;8:147-155.
6. Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*. 1996;52:721-725.
7. Lee SH. On the versatility of the combination of the weighted log-rank statistics. *Comput Stat Data Anal*. 2007;51(12):6557-6564.
8. Karrison T. Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata J*. 2016;16(3):678-690.
9. Lin RS, Lin J, Roychoudhury S, et al. Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. *Stat Biopharm Res*. 2020;12(2):187-198.
10. Roychoudhury S, Anderson KM, Ye J, Mukhopadhyay P. Robust design and analysis of clinical trials with nonproportional hazards: a straw man guidance from a cross-pharma working group. *Stat Biopharm Res*. 2021;1-15. doi:10.1080/19466315.2021.1874507
11. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika*. 1982;69(3):553-566.
12. Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 1981;68(1):316-319.
13. Lee JW, Sather HN. Supremum version of logrank test for detecting late occuring survival differences. *Comput Stat Data Anal*. 1998;26(3):303-311.
14. Xu Z, Park Y, Zhen B, Zhu B. Designing cancer immunotherapy trials with random treatment time-lag effect. *Stat Med*. 2018;37(30):4589-4609.
15. Li B, Su L, Ye Y, Yan F. M&M: a maximum duration design with the Maxcombo test for a group sequential trial of an immunotherapy with a random delayed treatment effect. *Stat Med*. 2022;41(4):815-830.
16. Gill RD. Censoring and stochastic integrals. *Stat Neerl*. 1980;34(2):124.
17. Fleming TR, Harrington DP, O'sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. *J Am Stat Assoc*. 1987;82(397):312-320.
18. Lin DY, Wei LJ, Zing Z. Checking the cox model with cumulative sums of martingale residuals. *Biometrika*. 1993;80(3):557-572.
19. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. Wiley Series in Probability and Statistics. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Ltd; 2005.
20. Kong HF, Slud E. Robust covariate-adjusted logrank tests. *Biometrika*. 1997;84(4):847-862.
21. Gu M, Ying Z. Group sequential methods for survival data using partial likelihood score processes with covariate adjustement. *Stat Sin*. 1995;5:793-804.
22. Lin DY, Wei LJ. The robust inference for the cox proportional hazards model. *J Am Stat Assoc*. 1989;84(408):1074-1078.
23. Magirr D, Burman CF. Modestly weighted logrank tests. *Stat Med*. 2019;38(20):3782-3790.
24. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12(9):13. doi:10.1186/1471-2288-12-9
25. Antonia SJ, Villegas A, Daniel D, et al. Durvalumab after chemoradiotherapy in stage III non–small-cell lung cancer. *N Engl J Med*. 2017;377(20):1919-1929.
26. Bellmunt J, Wit dR, Vaughn DJ, et al. Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *N Engl J Med*. 2017;376(11):1015-1026.
27. Hasegawa T. Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharm Stat*. 2014;13(2):128-135.
28. Eng KH, Kosorok MR. A sample size formula for the supremum log-rank statistic. *Biometrics*. 2005;61(1):86-91. doi:10.1111/j.0006-341X.2005.031206.x
29. Ristl R, Ballarini N, Götte H, Schüler A, Posch M, König F. Delayed treatment effects, treatment switching and heterogeneous patient populations: how to design and analyze RCTs in oncology. *Stat Pharm*. 2021;20(1):129-145.
30. Rufibach K. Treatment effect quantification for time-to-event endpoints - estimands, analysis strategies, and beyond. *Pharm Stat*. 2019;18(2):145-165.

How to cite this article: Boher JM, Filleron T, Bunouf P, Cook RJ. New late-emphasis and combination tests based on infimum and supremum logrank statistics with application in oncology trials. *Statistics in Medicine*. 2023;42(12):1981-1994. doi: 10.1002/sim.9709

APPENDIX

Elements of large sample theory

Using first-order approximation and standard probabilistic arguments,[21] it follows that the observed covariate-adjusted partial score processes $n^{-1/2}S_\theta(t,0,\hat{\beta})$ and $n^{-1/2}\widetilde{S}_\theta(t,0,\hat{\beta})$ can be decomposed as follows,

$$n^{-1/2}S_\theta(t,0,\hat{\beta}) = n^{-1/2}S_\theta(t,0,\beta_0) - i_{\theta,\beta}(t,\beta_0)i^{-1}_{\beta,\beta}(\beta_0)n^{-1/2}S_\beta(\tau,0,\beta_0) + O_P(1),$$

and

$$n^{-1/2}\widetilde{S}_\theta(t,0,\hat{\beta}) = n^{-1/2}\widetilde{S}_\theta(t,0,\beta_0) - \widetilde{i}_{\theta,\beta}(t,\beta_0)i^{-1}_{\beta,\beta}(\beta_0)n^{-1/2}\widetilde{S}_\beta(\tau,0,\beta_0) + O_P(1),$$

where $O_P(1)$ designates a negligible term. Given the following martingale decomposition of the Cox partial likelihood score processes,[22]

$$n^{-1/2}S_\theta(t,\theta,\beta) = n^{-1/2}\sum_{i=1}^{n}\int_0^t \left[Z_i - \frac{z^{(1)}(u,\theta,\beta)}{y(u,\theta,\beta)}\right] dM_i(u,\theta,\beta) + O_P(1),$$

and

$$n^{-1/2}S_\beta(t,\theta,\beta) = n^{-1/2}\sum_{i=1}^{n}\int_0^t \left[Z_i - \frac{w^{(1)}(u,\theta,\beta)}{y(u,\theta,\beta)}\right] dM_i(u,\theta,\beta) + O_P(1).$$

it then follows that the observed processes $n^{-1/2}S_\theta(t,0,\hat{\beta})$ and $n^{-1/2}\widetilde{S}_\theta(t,0,\hat{\beta})$ are asymptotically equivalent under $H_0 : \theta = 0$ to the sums of independent random variables given in (2) and (3), respectively. Following Lin et al[18] an empirical approximation to the null distribution of the different test statistics based on the extreme values of statistics $S_\theta(t,0,\hat{\beta})$ and $\widetilde{S}_\theta(t,0,\hat{\beta})$ are derived by sampling independent realizations,

$$S^m(t,\hat{\beta}) = \sum_{i=1}^{n} I(X_i \le t)\Delta_i U_i^m \left[Z_i - \frac{\overline{Z}^{(1)}(X_i,\hat{\beta})}{Y(X_i,\hat{\beta})}\right] - I(t,\hat{\beta})I^{-1}(\hat{\beta})\sum_{i=1}^{n}\Delta_i U_i^m \left[W_i - \frac{\overline{W}^{(1)}(X_i,\hat{\beta})}{\overline{Y}(X_i,\hat{\beta})}\right],$$

TABLE A1 Empirical sizes and powers of test statistics based upon supremum logrank statistics under proportional hazards (PH) models and unequal allocation across arms (1:2).

				Empirical error rates (%)—*One-tailed upper tests* - $\alpha = 0.025$							
	Hazard ratio for			**Unadjusted test statistics**				**Adjusted test statistics**			
Scenario	**treatment**	**Covariate**	***N***	**$FH^{0,0}$**	***Inf***	***le-Inf***	***Combo-Inf***	**$FH^{0,0}$**	***Inf***	***le-Inf***	***Combo-Inf***
S1	1.0	1.0	100	3.0	1.7	1.6	1.6	2.8	1.7	1.7	1.7
	1.0	1.0	150	2.9	2.0	2.0	1.9	3.1	2.0	2.1	2.0
S2	1.0	2.0	100	2.9	1.6	1.7	1.7	3.5	1.9	2.1	1.9
	1.0	2.0	150	2.8	1.8	1.9	1.7	3.0	2.1	2.3	2.1
S3	0.5	1.0	150	94.2	89.9	90.1	89.3	93.7	89.2	89.8	89.2
S4	0.5	2.0	150	89.8	84.8	84.1	83.6	93.1	88.5	88.7	88.5
				Empirical error rates (%) - *Two-tailed tests* - $\alpha = 0.050$							
	Hazard ratio for			**Unadjusted test statistics**				**Adjusted test statistics**			
Scenario	**treatment**	**Covariate**	***N***	**$\lvert FH^{0,0}\rvert$**	***|Sup|***	***|le-Sup|***	***Combo-|Sup|***	**$\lvert FH^{0,0}\rvert$**	***|Sup|***	***|le-Sup|***	***Combo-|Sup|***
S1	1.0	1.0	100	5.2	5.0	5.0	5.5	5.1	4.7	4.7	5.3
	1.0	1.0	150	5.6	5.3	5.1	5.5	5.2	5.0	5.0	5.3
S2	1.0	2.0	100	5.9	5.1	5.1	5.6	5.6	5.0	4.9	5.4
	1.0	2.0	150	4.9	4.8	4.8	5.2	5.2	4.9	4.8	5.3
S5	2.0	1.0	150	92.2	92.8	92.8	92.7	91.4	92.0	92.0	92.1
S6	2.0	2.0	150	88.5	89.8	89.1	89.7	91.2	91.9	91.6	91.6

One-sided upper test statistics: $Inf = \inf_{t\in(0,\tau]} S_{0,0}(t)$, $le\text{-}Inf = \inf_{t\in(0,\tau]} \hat{S}_{0,0}(t)$ and $Combo\text{-}Inf = \min(Inf, le\text{-}Inf)$. Two-sided test statistics: $|Sup| = \sup_{t\in(0,\tau]} |S_{0,0}(t)|$, $|le\text{-}Sup| = \sup_{t\in(0,\tau]} |\hat{S}_{0,0}(t)|$ and $Combo\text{-}|Sup| = \max(|Sup|, |le\text{-}Sup|)$.

and

$$\widetilde{S}^m(t,\hat{\beta}) = \sum_{i=1}^{n} I(X_i \geq t)\Delta_i U_i^m \left[Z_i - \frac{\overline{Z}^{(1)}(X_i,\hat{\beta})}{\overline{Y}(X_i,\hat{\beta})} \right] - I(t,\hat{\beta})I^{-1}(\hat{\beta}) \sum_{i=1}^{n} \Delta_i U_i^m \left[W_i - \frac{\overline{W}^{(1)}(X_i,\hat{\beta})}{\overline{Y}(X_i,\hat{\beta})} \right],$$

obtained by drawing independent random normal deviates $\left\{U_i^m; m = 1, \ldots, M, i = 1, \ldots, n\right\}$, and replacing unknown martingale increments $dM_i(t, 0, \beta_0)$ and limiting values in equations (2) and (3) by randomly perturbed terms $U_i^m dN_i(t)$ and consistent sample estimates.