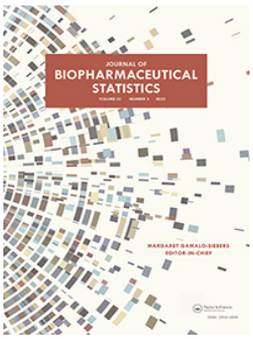


Taylor & Francis
Taylor & Francis Group

Journal of Biopharmaceutical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Determining the late effect parameter in the Fleming-Harrington test using asymptotic relative efficiency in cancer immunotherapy clinical trials

Yuichiro Kaneko & Satoshi Morita

To cite this article: Yuichiro Kaneko & Satoshi Morita (2023): Determining the late effect parameter in the Fleming-Harrington test using asymptotic relative efficiency in cancer immunotherapy clinical trials, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2023.2244055

To link to this article: https://doi.org/10.1080/10543406.2023.2244055

Published online: 10 Aug 2023.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2023.2244055

Taylor & Francis
Taylor & Francis Group

Determining the late effect parameter in the Fleming-Harrington test using asymptotic relative efficiency in cancer immunotherapy clinical trials

Yuichiro Kaneko [a,b] and Satoshi Morita [a]

[a]Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan; [b]Data Science Statistical & RWD Science, Astellas Pharma Global Development Inc, Northbrook, Illinois, USA

ABSTRACT

The delayed treatment effect, which manifests as a separation of survival curves after a change point, has often been observed in immunotherapy clinical trials. A late effect of this kind may violate the proportional hazards assumption, resulting in the non-negligible loss of statistical power of an ordinary log-rank test when comparing survival curves. The Fleming-Harrington (FH) test, a weighted log-rank test, is configured to mitigate the loss of power by incorporating a weight function with two parameters, one each for early and late treatment effects. The two parameters need to be appropriately determined, but no helpful guides have been fully established. Since the late effect is expected in immunotherapy trials, we focus on the late effect parameter in this study. We consider parameterizing the late effect in a readily interpretable fashion and determining the optimal late effect parameter in the FH test to maintain statistical power in reference to the asymptotic relative efficiency (ARE). The optimization is carried out under three lag models (i.e. linear, threshold, and generalized linear lag), where the optimal weights are proportional to the lag functions characterized by the change points. Extensive simulation studies showed that the FH test with the selected late parameter reliably provided sufficient power even when the change points in the lag models were misspecified. This finding suggests that the FH test with the ARE-guided late parameter may be a reasonable and practical choice for the primary analysis in immunotherapy clinical trials.

ARTICLE HISTORY
Received 11 October 2022
Accepted 28 July 2023

KEYWORDS
Asymptotic relative efficiency; delayed effect; Fleming-Harrington test; immunotherapy clinical trial; weighted log-rank test

1. Introduction

Many agents in the field of cancer immunotherapy exhibit a delayed effect, while cytotoxic agents do not (Hoos et al. 2010). A typical delayed effect was observed, for example, in overall survival (OS) curves in a phase III clinical trial in patients with recurrent squamous cell carcinoma of the head and neck (Ferris et al. 2016). In that trial, survival curves overlapped between two treatment groups in the beginning and then started to separate at around 3–4 months after the initiation of the study treatment. Although several papers discuss that such a phenomenon of survival curves may stem from a subgroup of patients who have a non-responsive biomarker profile (Cristescu et al. 2018; Sharma et al. 2017), in this paper we consider the problem under the assumption that this kind of delayed effect is due to a lag before the induction of the immune response, leading to a violation of the proportional hazard assumption when comparing survival curves. In such a situation, applying a conventional log-rank test may lead to a loss of statistical power. Thus far, several weighted log-rank test approaches have been proposed to deal with the delayed effect and circumvent the power loss by incorporating a weight function when comparing treatment groups (Hasegawa 2014; Liu et al. 2018; Xu et al. 2016).

CONTACT Yuichiro Kaneko yuichiro.kaneko@astellas.com Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan

Hasegawa (2014) proposed a log-rank test using the Fleming-Harrington (FH) class of weights for sample size determination in cancer vaccine studies. Xu et al. (2016) proposed the use of a piecewise constant weighted log-rank test (PC test) in cancer vaccine studies. Both methods are implemented under a single change point model, thus being somewhat restrictive. Liu et al. (2018) suggested a piecewise proportional hazards cure rate model with the use of cumulative density function of random delayed time as weight function. We focused on FH test because it is simple and can be easily implemented, and the fact that it is convenient to allocate the weight on early, middle, and late effects by specifying the values of the parameters.

Fleming and Harrington (1991) proposed the FH test as a weighted log-rank test. The FH test uses the weight function $W(t) = \hat{S}(t-)^{\rho}\left(1 - \hat{S}(t-)\right)^{\gamma}$, where $\hat{S}(t-)$ denotes the left-continuous version of the Kaplan–Meier survival function in the pooled sample at time t, and ρ and γ denote the early and late effect parameters, respectively. For example, $W(t)$ with $(\rho,\gamma) = (0,0)$ gives the ordinary (conventional) log-rank test. Prentice (1978) suggested that the combination of $\rho = 1$ and $\gamma = 0$ may be a relevant choice for detecting an early effect. In order to efficiently detect the late effect that is frequently observed in immunotherapy randomized clinical trials (i-RCTs), it may be reasonable to determine $\gamma(>0)$ under $\rho = 0$ so that the required statistical power is maintained as much as possible.

Lin et al. (2020) proposed the Max-combo test, which combines the weight function of the FH test with the four prototypes $(\rho, \gamma) = (0, 0), (0, 1), (1, 0), (1, 1)$. Note that $(\rho, \gamma) = (1, 1)$ is set to evaluate a medium-term effect. The Max-combo test first calculates the four test statistics with the four weights, then selects the weight function with the maximum statistic, and finally derives the adjusted p-value under the multiplicity correction.

In the context of a preventive clinical trial for Alzheimer disease, Garès et al. (2014, 2015, 2017) proposed using a specific form of the survival function in the FH test, namely $\left(1 - \hat{S}(t-)\right)^{\gamma}$, with the late effect γ to be optimized in terms of Pitman's asymptotic relative efficiency (ARE) (Nikitin 2011) under $\rho = 0$. Garès et al. (2014) used ARE to evaluate the relationship between the FH test with a weight function of $\left(1 - \hat{S}(t-)\right)^{\gamma}$ and the piecewise weighted log-rank test with a single change point t^*. They remarked that in the simulations, the parameter γ in the FH test was less sensitive to misspecification than the change point t^*. They proposed to first choose the change point t^*, based on a priori knowledge about the expected late effects, and then to identify the value of γ and use the FH test with selected γ under $\rho = 0$, which matches best with the desired piecewise log-rank test. In Garès et al. (2015), on the other hand, the authors assumed that the delayed effect is not certain to happen in the study design stage, and they proposed the new statistic as maximum of the log-rank statistic and several FH's statistics for late effect. Their simulation showed the proposed test could maintain its statistical power in both proportional hazard and delayed effect situations. Garès et al. (2017) investigated the relationship between the necessary sample size for testing under the FH test with γ under $\rho = 0$ is optimal and investigated the change point achieved by every value of parameter γ in the FH test for various simulation settings. Based on their results, they recommended using a parameter γ value of 3.0 for general choice to detect late effects.

However, a more user-friendly method may be desirable because such a specific form of the survival function may be difficult for physicians to properly interpret and may not necessarily fit an immunotherapy trial well. Therefore, we consider parameterizing the late effect in the framework of a hazard ratio that may be readily interpretable by physicians who design clinical trials and use their results. The objective of this paper is to propose a practically useful and conveniently implementable procedure to determine the value of the late effect parameter γ of the FH test for designing an i-RCT. In order to select γ so that the loss of statistical power is minimized, we use Pitman's ARE as a guide to compare the proposed FH test and the optimal log-rank test under multiple lag models that may be plausibly assumed for an i-RCT. As will be explained in Section 3.3, the ARE evaluates the sample sizes required for two different statistical tests to achieve the same statistical power, assuming the asymptotic normality of the test statistics. Therefore, we first carried out a computer simulation to evaluate the validity of the asymptotic normality assumption under typical sample sizes for real clinical trials. In this simulation, we will also evaluate the robustness of the FH test with selected γ values when the

assumed change point is correctly or incorrectly specified. Then, we will seek the utility of the proposed procedure compared with the optimally weighted log-rank test.

Once the availability of the ARE is established, the computational burdens for designing an i-RCT will be substantially lower than when using the empirical relative efficiency, which requires more massive numerical computations.

This paper is organized as follows. In Section 2, we review the weighted log-rank test. In Section 3, we present the proposed procedure to determine the late effect parameter in the FH test using the ARE. In Section 4, we conduct extensive simulation studies to examine the performance of the proposed method, and to evaluate the performance of the FH test with several γ values in various situations. In Section 5, we apply our proposed method to the real-world example, and we then close with a brief discussion in Section 6.

2. Weighted log-rank tests and relevant notation

A weighted log-rank test may be a reasonable alternative to a conventional log-rank test for comparing survival curves between treatment arms when the proportional hazard assumption is violated, because such a violation may result in a serious loss of statistical power.

We work in the setting of a two-arm randomized clinical trial. The total n subjects are randomly allocated to a control group of n_0 subjects with an allocation ratio p_0, or to a new treatment group of n_1 (with $1-p_0$)subjects. The latent failure time T_{ij}^* is the random variable associated with the survival function, and the latent censoring time C_{ij} is the random variable associated with the censoring function for the jth subject in group i, where $i=0$ denotes the control group and $i=1$ denotes the treatment group. The survival time and censoring time distributions are denoted by $S_i = P\left(T_{ij}^* \geq \mathrm{t}\right)$ and $G_i = P\left(C_{ij} \geq t\right)$, respectively. The latent failure time and latent censoring time are assumed to be independent. The observed failure time is then given by $T_{ij} = \min\left(T_{ij}^*, C_{ij}\right)$, and thus the probability of being at risk is given by

$$\pi_i(t) = P\left(T_{ij}^* \geq t, C_{ij} \geq t\right) = P\left(T_{ij} \geq t\right) = S_i(t-)G_i(t-),$$

where $S_i(t-)$ and $G_i(t-)$ denote the left-continuous versions of the survival function and censoring distribution at time t in group i, respectively. Let $\delta_{ij} = I\left(T_{ij}^* \leq C_{ij}\right)$ denote the observed failure indicator for the j^{th} subject in group i at T_{ij}, with the indicator function $I(\cdot)$.

These data can be expressed in terms of counting process notation using $N_{ij}(t) = I\left(T_{ij} \leq t, \delta_{ij} = 1\right), Y_i(t) = \sum_{j=1}^{n} I\left(T_{ij} \geq t\right)$, and $N_i(t) = \sum_{j=1}^{n_i} N_{ij}(t)$, where $N_i(t)$ denotes the number of failures in group i before or at time t, and $Y_i(t)$ denotes the number at risk in group i at time $t-$, and both $N_i(t)$, and $Y_i(t)$ are random variables. Furthermore, $N(t) = N_0(t) + N_1(t)$ and $Y(t) = Y_0(t) + Y_1(t)$. Finally, τ denotes the study duration. Let $\lambda_0(t)$ and $\lambda_1(t)$denote the hazard functions in the control and treatment groups, respectively, and let $\Lambda(t)$ denote the cumulative function of pooled hazard.

For testing the null hypothesis$H_0 : S_0(t) = S_1(t)$ vs. the alternative hypothesis $H_1 : S_0(t) \neq S_1(t)$, the weighted log-rank test statistic with the weight function $W(t)$ is given by

$$T_W = n^{-1} \int_0^\tau W(t)\left\{\frac{Y_0(t)Y_1(t)}{Y(t)}\right\}\left\{\frac{dN_0(t)}{Y_0(t)} - \frac{dN_1(t)}{Y_1(t)}\right\}. \tag{1}$$

The variance of T_W, σ_w^2, is estimated by

$$\hat{\sigma}_w^2 = n^{-1} \int_0^\tau W^2(t) \frac{Y_0(t)Y_1(t)}{Y(t)} d\Lambda(t). \tag{2}$$

Note that the weight function $W(t) = \{\hat{S}(t-)\}^\rho \{1 - \hat{S}(t-)\}^\gamma$ for $\gamma \geq 0, \rho \geq 0$ defines the FH test.

Note also that the test statistic $Z_w = \sqrt{n} T_W / \hat{\sigma}_w$ is asymptotically normal $N\left(\sqrt{n}\mu_w/\sigma_w, 1\right)$ under the local alternative where

$$\mu_w = \int_0^\tau w(t) \frac{p_0\pi_0(t)(1-p_0)\pi_1(t)}{\pi(t)} \{\lambda_0(t) - \lambda_1(t)\} dt, \tag{3}$$

$$\sigma_w^2 = \int_0^\tau w^2(t) \frac{p_0\pi_0(t)(1-p_0)\pi_1(t)}{\pi^2(t)} (p_0\pi_0(t)\lambda_0(t) + (1-p_0)\pi_1(t)\lambda_1(t)) dt, \tag{4}$$

with $\pi_0(t) = \lim_{n\to\infty} \frac{Y_0(t)}{n_0}$, $\pi_1(t) = \lim_{n\to\infty} \frac{Y_1(t)}{n_1}$, $p_0 = \lim_{n\to\infty} \frac{n_0}{n}$, and $\pi(t) = p_0\pi_0(t) + (1-p_0)\pi_1(t)$. Here, $w(t)$ is a deterministic function given by $\lim_{n\to\infty} W(t) = w(t)$ through the convergence in probability.

3. Proposed approach

In order to determine the late effect parameter γ in the FH test, we first assume three plausible lag models that have different hazard functions with different optimal weights. Under each of these models, we calculate the required sample size for the optimally weighted log-rank test as explained in Section 3.2. Then we evaluate the ARE to compare the efficiency of the FH test with that of the optimally weighted log-rank test. Finally, we determine the parameter γ to give the maximum value of the ARE as explained in Section 3.4.

3.1. Three lag models comprising the hazard function

As described in Section 1, the delayed effect has been frequently observed in immunotherapy randomized clinical trials (i-RCTs), and we focus on it here. First, we consider simple prototypical lag models that include change point parameters, as shown below.

Following Zucker and Lakatos (1990), Ye and Yu (2018), and Ding and Wu (2020), we use the hazard model with lag of effect,

$$\lambda_1(t) = \{\theta l(t) + (1 - l(t))\}\lambda_0(t), \tag{5}$$

where $\lambda_0(t)$ and $\lambda_1(t)$ denote the hazard functions for the control and new treatment arms, respectively, and $l(t)$ is a monotonic, non-decreasing function with $0 \leq l(t) \leq 1$. The value of $l(t)$ represents the achieved proportion of the full treatment effect θ at time t. As also discussed in Ye and Yu (2018), θ can be simply interpreted as the hazard ratio when $l(t) = 1$.

We assume the following three lag functions, which may sufficiently cover possible separation patterns between the two survival curves observed in immunotherapy clinical trials:

(a) threshold lag with change point t^* : the treatment has no effect before t^* and achieves full effect after t^*,

$$l(t) = I(t > t^*);$$

(b) linear lag with change point t^*: the treatment effect linearly starts from the beginning and achieves its full effect at change point t^*,

$$l(t) = \left(\frac{t}{t^*}\right) I(t \leq t^*) + I(t > t^*); \text{ and}$$

(c) generalized linear lag with change point t_1^*, t_2^* : the treatment has no effect before t_1^* and then linearly achieves its full effect at t_2^*,

$$l(t) = \frac{t - t_1^*}{t_2^* - t_1^*} I\left(t_1^* \le t < t_2^*\right) + I\left(t \ge t_2^*\right).$$

Note that the lag models are not limited to these three prototypical models but can be extended to other patterns. Also, for generalized linear lag, if the change point t_1^* is set as 0, then it becomes identical to the linear lag with change point t_2^*.

If the lag model is true, the survival function of the new treatment arm is given as

$$S_1(t) = exp\left\{-\left(\int_0^\tau \lambda_1(u)\mathrm{du}\right)\right\} = exp\left\{-\left(\int_0^\tau \{\theta l(t) + (1 - l(t))\}\lambda_0(u)\mathrm{du}\right)\right\}. \tag{6}$$

3.2. Sample size calculation for the optimally weighted log-rank test under each lag model

Given the type I and II error rates, α and β, the required total sample size for the weighted log-rank test is calculated as

$$n = \frac{(z_\alpha + z_\beta)^2 {\sigma_w}^2}{{\mu_w}^2}, \tag{7}$$

where z_α and z_β denote the upper $\alpha/2$ and β percentile points of the standard normal distribution, respectively, and μ_w and σ_w^2 are given above in Equations (3) and (4). Under the alternative hypothesis defined by each of the lag models, we calculate μ_w and σ_w^2 using $\lambda_1(t)$ in Equation (5) and $S_1(t)$ in Equation (6). When using the weight function $W(t) = l(t)$, the sample size takes the minimum value under the lag model, because in this context $W(t) = l(t)$ is the optimal weight in terms of detecting treatment effect over control (Zucker and Lakatos 1990).

3.3. The ARE in weighted log-rank tests

Tests based on the Z_w statistic given in Section 2 have a statistical power converging to unity as $n \to \infty$ for any fixed type of alternative, and hence it is difficult to compare the weighted log-rank tests in terms of statistical power. However, as discussed in Gill (1980), Fleming and Harrington (1991), Garès et al. (2014), Garès et al. (2015), and Garès et al. (2017), under a sequence of alternative hypotheses converging to the null hypothesis as $n \to \infty$, the distribution of a statistic has a finite mean asymptotically, as well as positive variance; hence, comparison between two test statistics is possible. This is the concept of Pitman's ARE. For log-rank type statistics, Gill (1980) discusses the efficiency properties in the framework of the local asymptotic, i.e., in the limit as $n \to \infty$ and as the treatment effect approaches zero at the rate $1/\sqrt{n}$.

ARE is the ratio of asymptotic efficacies (i.e. $\left(\mu_1/\sigma_1\right)^2/\left(\mu_2/\sigma_2\right)^2$) between two test statistics. Note that (μ_1, σ_1) and $\left(\mu_2, \sigma_2\right)$ are sets of asymptotic means and standard deviations for each set of test statistics under the sequence of alternatives. Then, ARE can also be regarded as the reciprocal ratio of sample sizes required for two different statistical tests to achieve the same statistical power. Thus, the ARE allows us to quantitatively compare the efficiencies between two statistical tests under the same statistical hypothesis.

Pitman's ARE comparing Z_{W_1} and Z_{W_2} is given by

$$\rho^2(Z_{W_1}, Z_{W_2}) = \frac{\left\{\int_0^\tau w_1 w_2 \varphi(t) dt\right\}^2}{\int_0^\tau {w_1}^2 \varphi(t) dt \int_0^\tau {w_2}^2 \varphi(t) dt} \tag{8}$$

with

$$\varphi(t) = \frac{p_0(1-p_0)\pi_0(t)\pi_1(t)}{p_0\pi_0(t) + (1-p_0)\pi_1(t)}\lambda_0(t),$$

where π_1 is evaluated under H_0 (Zucker and Lakatos 1990). Note that (8) is equivalent to the asymptotic correlation between Z_{W_1} and Z_{W_2} under H_0 (Gastwirth 1985). The integrals in (8) do not in general have closed forms, and therefore we numerically evaluate (8) as will be explained in Sec. 4.1.

As discussed in Zucker and Lakatos (1990), the statistical power (referred to as the "predicted power") for any test statistic Z_W can be evaluated using

$$\Phi\left\{-z_\alpha + \left(z_\alpha + z_\beta\right)\rho\left(Z_W, Z_{W_{opt}}\right)\right\}, \tag{9}$$

where Φ denotes the standard normal distribution function and $Z_{W_{opt}}$ is the optimal statistic.

As above, Pitman's ARE asymptotically compares the efficiencies of two statistical tests, assuming the asymptotic normality of their test statistics. Therefore, we empirically evaluate the ARE under a finite number of sample sizes for an actual clinical trial, as shown in Section 4. In other words, we will computationally evaluate the statistical power, that is, the "empirical power," by calculating the percentage of the simulated replications in the trial that reached the prespecified level of statistical significance.

3.4. Procedure to determine the late effect parameter in the FH test using the ARE

To determine the parameter γ in the FH test under the delayed effect situation, we take the following steps. Although there are potentially other non-proportional patterns seen in immunotherapy clinical trials, we only focus on the delayed effect pattern since it occurs commonly (Hoos et al. 2010). First, we assume one of the lag models (threshold lag, linear lag, and generalized linear lag) given in Section 3.1 and give the values of the true change points (t^*, t_1^*, t_2^*). Let $Z_{W_{\mathrm{FH}}}$ be the weighted log-rank statistic as shown in Section 2 with a weight function of $W_{FH}(t) = \left\{\hat{S}(t-)\right\}^\rho\left\{1-\hat{S}(t-)\right\}^\gamma$ for $\gamma \geq 0$ and $\rho \geq 0$, and let Z_l be the weighted log-rank statistic as shown in Section 2 with a weight function of $l(t)$, which is the lag function described in Section 3.1. As discussed by Zucker and Lakatos (1990), the weight function of $l(t)$ is the optimal under the assumed lag models. Then, we numerically evaluate the ARE of $Z_{W_{\mathrm{FH}}}$ in comparison to Z_l by altering the value of γ from 0 to 3.0 by 0.1-unit intervals. Note that because of the asymptotic correlations shown in (8), the ARE for $Z_{W_{\mathrm{FH}}}$ relative to Z_{W_l}, namely $\rho^2(Z_{W_{FH}}, Z_l)$, ranges from 0 to 1 because it indicates the relative efficiency when l is the optimal weight. At the same time, this ARE is also equivalent to that for Z_l relative to $Z_{W_{\mathrm{FH}}}$, namely $\rho^2(Z_l, Z_{W_{FH}})$, when W_{FH} is the optimal weight. In the proposed procedure, we only focus on $\rho^2(Z_{W_{FH}}, Z_l)$ because we determine the parameter γ under the situation where l is the optimal weight and Z_l is optimal statistic under the assumed lag model. Based on the ARE calculation results, we select the parameter γ to give the maximum value of the ARE that is above the threshold value (e.g., 0.8). Then, using Equation (7), we calculate the sample size with $W(t) = W_{FH}$ and the selected γ to achieve the statistical power (e.g., 0.8 or 0.9) under the assumed lag model.

4. Simulation study design and results

To compare the proposed method with the optimally weighted log-rank test and Max-combo test, we first simulated the trial 10,000 times under varying situations as determined by the lag models. We used a balanced design with equal (1:1) allocation to the new treatment and control arms in an i-RCT with a trial period of 36 months. We assumed an exponential survival function for the control arm to achieve a survival probability of 0.30 at 36 months. For the treatment arm, we defined the survival functions using the lag models (see Appendix A for details).

We set the value of the change point t^* from 0 months (no lag) to 18 months in 3-month intervals for the threshold and linear lags. For the generalized linear lag, we set t_2^* similarly as t^* and set t_1^* to be half the value of t_2^*. The hazard ratios (θ) after the change points t^* in threshold lag/linear lag and t_2^* in generalized linear lag are set to be 0.6. Finally, we set accrual period as 10.8 months with uniformly distributed accrual, and set follow-up period as 25.2 (=36 × 0.7) months.

As described in Section 3.3, we first calculated the ARE between the FH test and optimally weighted log-rank test by using Equation (8) with γ values that varied from 0 to 3.0 in 0.1-unit intervals under the three lag models with the true change point ranging from 0 months to 18 months in 3-month intervals. Then, we evaluated the similarity between the predicted power and the empirical power to ensure the availability of the ARE under the assumed circumstances.

First, we compared the empirical statistical power of the FH test, with those of the optimally weighted log-rank test and the Max-combo test (Lin et al. 2020). We predetermined the parameter γ as explained above, following the procedure in Section 3.4.

Second, we thoroughly assessed the impact of misspecification of parameter γ in the FH test, which corresponded to misspecification of the change point under the three lag models. In the second simulation, we likewise simulated the trial 10,000 times using a balanced design with equal (1:1) allocation to the new treatment and control arms under the lag models. The optimally weighted log-rank test under the lag models was set as the benchmark, and the sample size was calculated based on the optimally weighted log-rank test to maintain 80% statistical power (Table S1 of the Supporting Information). As the comparator, we used the Max-combo test. This test, as defined in the first simulation, generally combines the weight function of the FH test with the four prototypes $(\rho, \gamma) = (0, 0), (0, 1), (1, 0), (1, 1)$. In the second simulation, we defined it as Max-combo test 1. In addition, we arranged the late effect in the Max-combo test, not fixing $(\rho, \gamma) = (0, 1)$, but changing γ in $(0, \gamma)$ per change point under each lag model, as in the FH test, and defined it as Max-combo test 2.

Furthermore, to simultaneously evaluate the impact of the mis-specification in lag models and change point, we performed FH test in one lag model (e.g., threshold lag) under the other lag model (e.g., linear lag model). Similarly, the optimally weighted log-rank test under the other lag model (e.g., linear lag model) was set as the benchmark, and the sample size was calculated based on the optimally weighted log-rank test to maintain 80% statistical power. Max-combo test 1 and 2 were also evaluated as comparator.

Third, we sought to evaluate the operational characteristics of the FH test with various γ values under a greater variety of non-proportional hazard situations. These included not only the delayed effect but also other patterns such as the "crossing" survival function and the "diminishing" treatment effect, both of which are seen in several oncology clinical trials (Hellmann et al. 2019; Hussain et al. 2015). For the delayed effect, we created the survival function under threshold lag, linear lag, and generalized linear lag with a change point of 12 (months). In addition, in reference to Garès et al. (2014), we evaluated the survival functions in which the FH test is optimal by altering the γ value from 0.5 to 3.0. To evaluate the performance of the FH test with various γ values, we fixed the total number of events at 200 in all patterns.

To assess the operational characteristics of the FH test with various γ values ($\gamma = 0.5 - 3.0$) in these simulation datasets, we used the log-rank test, Gehan–Breslow test (GB), Max-combo test, and restricted mean survival time (RMST) as comparators. For the RMST, we set the specified times of interest as $t_{RMST} = 25.2$, and $t_{RMST} = 32$. The parameters used to create the survival functions are described in Table S4.

4.1. Evaluation of the proposed approach under the lag models

Figure 1 summarizes the ARE results of the FH test with γ and the optimally weighted log-rank test.

The ARE result under no lag is shown in Figure 1a. As expected, the ARE uniformly decreased when the parameter γ increased. When the survival curves of the two arms separated relatively early,

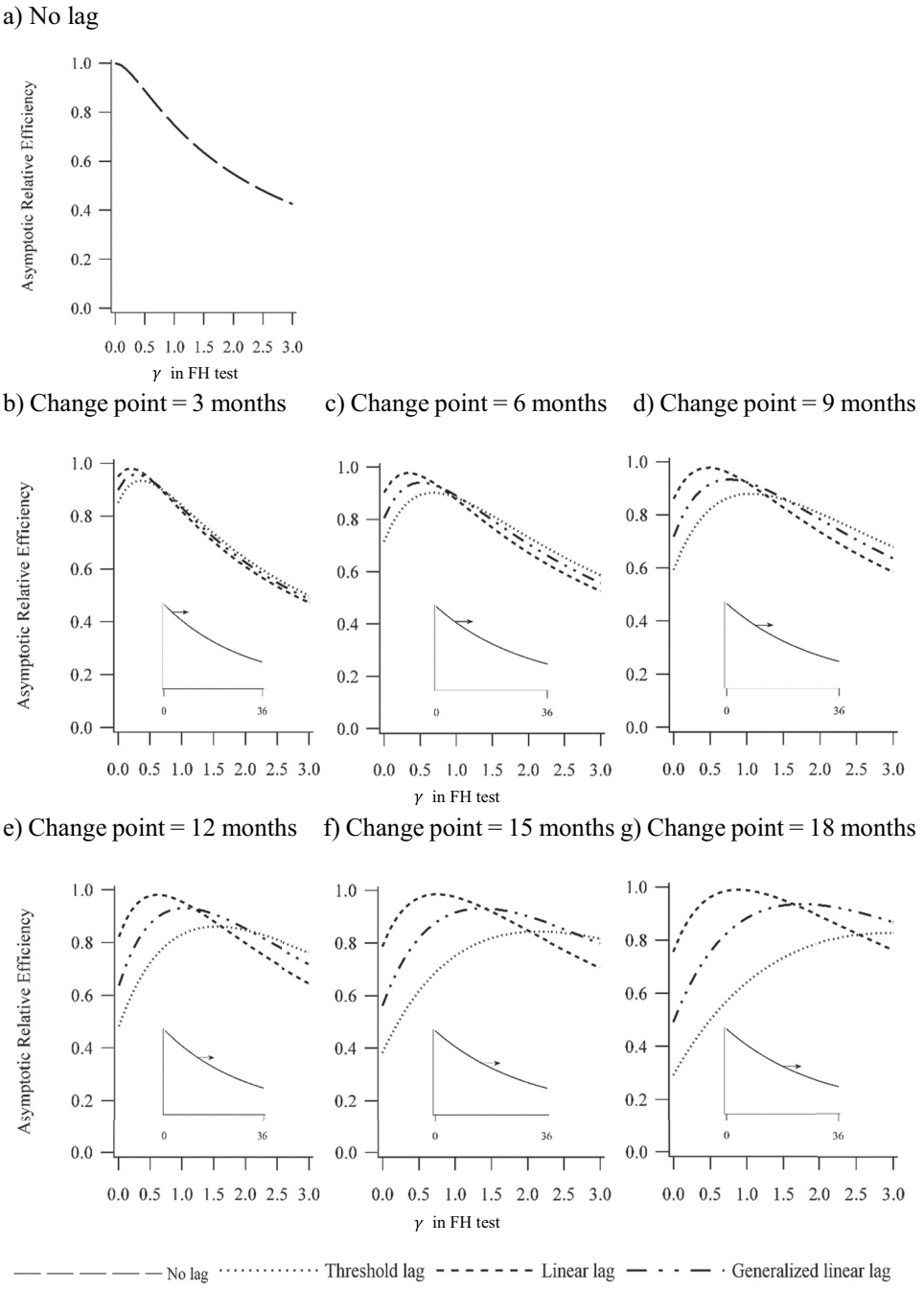


Figure 1. Asymptotic relative efficiency (ARE) results for the FH test with the late effect parameter γ and the optimally weighted log-rank test with the three lag models (threshold, linear, and generalized linear). a) – g) are defined by change point values ranging from 0 month to 18 months. The survival curve in the control group with the assumed change point time is displayed in each panel. The x-axis indicates the parameter values of γ ranging from 0 to 3.0 in 0.1-unit intervals.

such as at 3 months and 6 months (Figure 1b,c), the ARE was very similar in the three lag models; its maximum value was just after $\gamma = 0$, and decreased to about 0.5–0.6 at $\gamma = 3.0$. Thus, as expected, when the survival curves are expected to separate early, the ordinary log-rank test should be used in the primary analysis. In the presence of a delayed effect, that is, if the separation of survival curves occurs later, the ARE values demonstrate a greater difference between the lag models.

For threshold lag, the ARE was maximized when γ was specified to be 1.1, 1.5, 2.2, and 3.0 in Figure 1d–g, respectively. Also, the maximized ARE value was smaller for threshold lag than for the other two lag models. For linear lag, an ARE value of nearly 1.0 was achieved when γ was correctly specified around 0.5 to 1.0 in Figure 1b–g. Therefore, the FH test with a γ value of 0.5 to 1.0 is recommended. For generalized linear lag, the maximum ARE was achieved when γ was correctly specified around 0.5 to 2.0 in Figure 1c–g. Therefore, it is recommended that the FH test be used with a γ value of 0.5 to 2.0 for the change point that occurs before halfway through the study.

Figure 2a shows the predicted power as shown in (9), with γ values that varied from 0 to 3.0 in 0.5-unit intervals under the three lag models with the true change point ranging from 0 months to 18 months in 3-month intervals. The predicted power of the optimally weighted log-rank test is set to be 80% under the true change point. In Figure 2b, we confirm that the empirical and predicted powers are sufficiently close (detailed results are available in Table S1 of the Supporting Information). Consequently, we used the ARE to determine the parameter γ; that is, we identified which value of γ, ranging from 0 to 3.0 in 0.1-unit intervals, resulted in the maximum ARE value. Then, for the simulation study, we selected γ from the ARE results. To simplify the process, we selected γ from values ranging from 0 to 3 in 0.5-unit intervals.

Table 1 summarizes the selected γ and the required sample size to maintain the 80% statistical power under type I error = 0.05 in the simulation study. The assumed change points t^* (threshold lag/linear lag) and t_2^* (generalized linear lag) at the sample size calculation took values from 6 months, 12 months, and 18 months, corresponding to early, middle, and late change points.

Table 2 shows the simulation study results. The FH test could almost maintain the planned statistical power of 80% when the assumed change point corresponded to the true change point. In general, the FH test was quite comparable to the optimally weighted log-rank test and the Max-combo test, especially when the true change point was 6 months, although the weighted log-rank test with $W = l(t)$ had higher statistical power when the assumed change point was correctly specified. Under threshold lag, however, the FH test and the Max-combo test had higher statistical power than the optimally weighted log-rank test with $W = l(t)$, when the true change point was earlier than planned. Also, when the true change point was 18 months, the Max-combo test had considerably lower power than the FH test and the optimally weighted log-rank test with $W = l(t)$.

Under linear lag, the FH test was overall similar to the optimally weighted log-rank test with $W = l(t)$ and the Max-combo test.

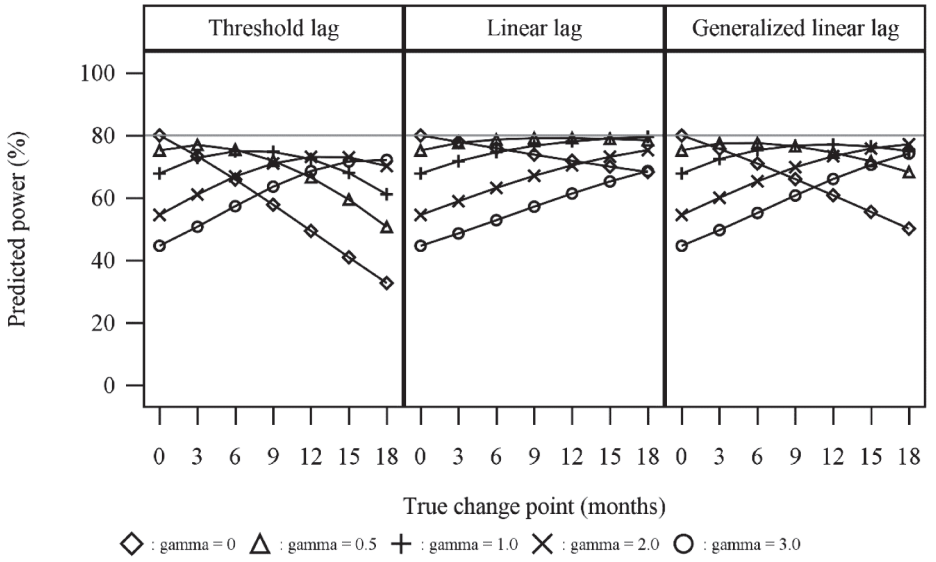
4.2. Evaluation of the impact of parameter misspecification under the lag models

In the second simulation, the parameter γ in the FH test, which corresponded to the change point under each lag model, was selected in accordance with the procedure described in Section 3.4 using ARE (Table S2 of the Supporting Information).

Table 3 shows the second simulation study results when the lag models were correctly specified but the change point was mis-specified. As designed, the optimally weighted log-rank test (the weighted log-rank test with $W = l(t)$) could almost maintain the planned statistical power of 80% when the assumed change point corresponded to the true change point.

For threshold lag, although the weighted log-rank test with $W = l(t)$ had higher statistical power when the assumed change point was correctly specified, the FH test had higher power when the change point was misspecified, except for the assumed change point of 0.

a) Predicted Power (%)



b) The difference between predicted power – empirical power

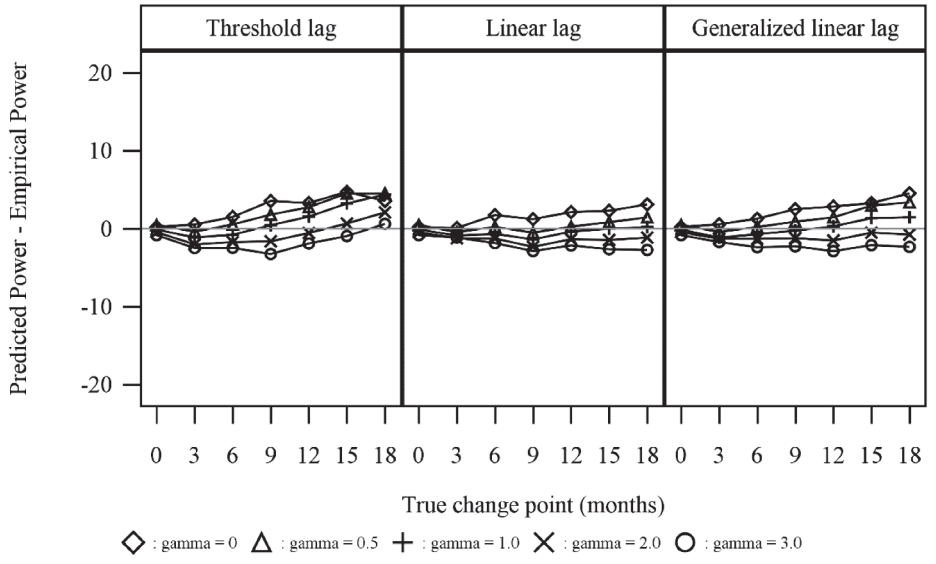


Figure 2. The performance for the FH test with parameter γ under the no-lag or each lag model was true with change point t^* (threshold lag/linear lag) or t_1^*, t_2^* (generalized linear lag). The upper row displays a) predicted power (%), and the lower row displays b) the difference between predicted power and empirical power (i.e., predicted power – empirical power).

Also, compared with Max-combo tests 1 and 2, the FH test had higher power when the change point was correctly specified. On the other hand, in the true change points of 0, 3, and 6, Max-combo tests 1 and 2 maintained statistical power above 70% in all cases, while the FH test had lower statistical power when the change point was assumed much later than true change point. However, if the true change point was relatively later, such as 12, 15, or 18,

Table 1. Total sample sizes (n : used in the simulation study) to achieve 80% statistical power for the FH test with the late effect parameter γ under the threshold, linear, and generalized linear lag models for change points = 6 months, 12 months, and 18 months.

	Threshold lag		Linear lag		Generalized linear lag	
Change point (months)	γ	n	γ	n	γ	n
6	0.5	360	0.5	286	0.5	316
12	1.5	558	0.5	368	1.0	444
18	3.0	982	1.0	474	2.0	646

Table 2. Comparison of the FH test with the optimally weighted log-rank test, and Max-combo test when the true change point was misspecified in each lag model.

		Threshold lag			Linear lag			Generalized linear lag		
True CP (months)	Assumed CP (months)	Optimal	FH	Max-combo	Optimal	FH	Max-combo	Optimal	FH	Max-combo
0	6	86.1	91.8	93.3	86.1	85.1	87.5	-	-	-
3		85.3	88.1	86.2	84.7	83.6	82.9	85.2	86.0	84.1
6		**84.7**	**79.4**	**77.5**	**81.1**	**80.2**	**78.0**	**82.7**	**80.1**	**77.6**
9		67.3	66.8	65.4	75.6	75.6	72.6	73.4	72.0	69.5
12		48.0	51.8	50.4	68.5	69.7	66.6	60.4	62.0	60.1
15		31.5	36.3	35.4	60.4	63.3	60.3	47.2	51.3	50.0
18		20.0	23.4	22.9	52.5	56.5	53.7	35.3	40.5	39.3
0	12	88.7	94.3	99.3	91.6	93.0	94.2	-	-	-
3		87.9	94.3	96.9	90.6	91.7	91.0	88.9	92.5	93.7
6		87.3	92.4	92.6	88.9	89.0	87.5	88.5	90.1	89.6
9		86.5	87.9	84.0	85.8	85.2	83.0	87.4	86.4	83.8
12		**85.8**	**79.6**	**71.0**	**81.3**	**80.5**	**77.6**	**82.9**	**80.1**	**75.2**
15		64.6	66.3	53.0	75.3	74.2	71.6	73.5	71.0	64.3
18		40.9	48.2	34.0	67.5	67.0	64.6	59.8	60.4	51.9
0	18	91.2	97.2	100.0	94.9	94.4	98.0	-	-	-
3		90.6	97.2	99.9	94.5	94.2	96.3	92.7	94.6	98.9
6		89.8	97.3	99.5	93.5	93.2	94.0	92.4	94.7	97.5
9		88.9	96.3	97.8	92.0	91.6	91.1	92.1	93.5	94.5
12		88.7	94.8	92.2	89.6	88.9	87.7	91.2	91.2	89.2
15		87.8	90.2	78.4	86.1	85.3	82.5	88.9	86.8	80.6
18		**87.3**	**79.8**	**56.5**	**80.6**	**80.0**	**76.1**	**82.9**	**80.2**	**68.6**

CP:change point.

then the FH test had higher power than Max-combo test 1 in most of the misspecified change points, and had higher power than Max-combo tests 2, except for the assumed change point of 0.

For linear lag, the four tests were mostly comparable with each other, while the optimally weighted log-rank test with $W = l(t)$ had the highest power, though only marginally, when the change point was correctly specified. In addition, none of the four tests had severely decreased statistical power in any true change point against misspecification under linear lag.

For generalized linear lag, the weighted log-rank test with $W = l(t)$ had the highest power when the change point was correctly specified. However, the FH test was more robust than the weighted log-rank test with $W = l(t)$ when the degree of misspecification was greater. Compared with Max-combo tests 1 and 2, the FH test had higher power when the change point was correctly specified. In terms of robustness against misspecification, however, when the true change points were 3, and 6, Max-combo tests 1 and 2 were more robust than the FH test in case the later change points such as 15 and 18 were assumed for the statistical tests. On the other hand, when the true change points were 15 and 18, the FH test was more robust than Max-combo test 2 in all change points, and had higher power than Max-combo test 1 in most of the change points.

Table 3. Robustness of the optimally weighted log-rank test, FH test, and Max-combo test when CP (months) or FH parameter γ was misspecified and tested under different assumptions in each lag model.

		Threshold lag				Linear Lag				Generalized Linear Lag			
True CP	CP tested	Optimal	FH	Max1	Max2	Optimal	FH	Max1	Max2	Optimal	FH	Max1	Max2
0	**0**	**79.8**	**79.8**	77.5	**78.1**	**79.8**	**79.8**	77.5	**78.1**	-	-		-
	3	74.0	76.4		78.2	78.1	78.5		78.2	-	-		-
	6	66.6	71.8		78.0	76.2	77.6		78.2	-	-		-
	9	59.2	66.4		77.4	74.4	74.7		78.1	-	-	-	-
	12	51.8	61.1		76.7	72.2	73.1		78.1	-	-		-
	15	43.4	52.9		76.0	70.4	71.8		78.0	-	-		-
	18	35.8	45.5		75.5	68.7	69.2		77.7	-	-		-
3	0	72.8	72.8	75.2	75.7	77.9	77.9	77.0	77.5	-	-	76.2	-
	3	**80.3**	**77.6**		**75.8**	**80.4**	**79.4**		**77.8**	**80.3**	**78.7**		**77.1**
	6	73.9	76.3		75.7	79.6	79.2		77.7	77.2	77.9		77.0
	9	66.6	72.9		75.0	78.0	78.1		77.6	72.5	75.6		76.6
	12	58.0	68.6		74.4	76.4	77.2		77.5	68.1	73.6		76.2
	15	49.8	61.1		73.5	74.8	76.4		77.4	63.0	68.7		75.6
	18	40.0	53.3		72.7	73.2	74.1		77.1	57.8	63.8		74.9
6	0	64.4	64.4	72.4	71.8	74.1	74.1	76.3	76.5	-	-	75.0	-
	3	72.1	74.0		71.7	77.9	77.6		76.6	76.5	76.4		74.9
	6	**80.1**	**75.7**		**72.0**	**79.5**	**78.3**		**76.6**	**80.2**	**77.4**		**74.9**
	9	73.0	75.1		72.4	79.1	78.5		76.5	77.8	77.1		75.0
	12	65.0	72.6		72.1	78.2	78.0		76.6	73.3	76.0		75.0
	15	55.4	67.0		71.6	77.0	77.5		76.6	68.1	72.7		74.6
	18	45.9	59.8		70.7	75.6	76.0		76.4	62.8	69.0		74.1
9	0	54.3	54.3	68.5	65.8	72.6	72.6	77.1	77.1	-	-	73.1	-
	3	62.6	68.5		65.8	77.2	77.5		77.2	70.3	72.9		72.1
	6	70.8	72.8		66.9	79.6	78.6		77.2	77.2	75.6		72.2
	9	**80.3**	**74.5**		**69.0**	**80.8**	**79.6**		**77.1**	**80.3**	**77.1**		**72.9**
	12	72.6	74.6		69.7	80.6	79.6		77.2	78.4	77.0		73.1
	15	63.3	71.9		69.4	79.7	79.5		77.2	73.9	75.1		73.1
	18	52.6	66.9		68.3	78.8	78.6		77.1	68.3	72.6		72.8
12	0	46.2	46.2	63.8	57.6	69.8	69.8	76.4	75.6	-	-	71.3	-
	3	52.7	61.4		57.7	74.3	75.5		75.6	64.1	69.0		68.6
	6	61.0	67.4		60.5	77.2	77.2		75.6	70.9	73.1		68.8
	9	69.9	71.8		64.6	79.2	78.9		75.6	77.1	75.9		70.4
	12	**80.1**	**73.7**		**66.3**	**80.0**	**79.2**		**75.8**	**80.2**	**76.8**		**71.3**
	15	71.7	73.3		67.0	80.1	79.3		76.0	79.0	76.7		72.2
	18	60.0	70.5		66.6	79.5	79.0		76.3	75.2	75.7		72.3
15	0	36.3	36.3	56.0	46.7	67.7	67.7	75.5	74.0	-	-	68.7	-
	3	42.2	52.3		47.1	72.0	74.0		74.0	58.1	64.4		63.8
	6	49.5	59.6		51.7	75.4	75.9		74.0	65.0	68.8		64.2
	9	58.3	66.1		57.3	77.8	78.1		74.1	71.8	73.5		66.9
	12	68.0	69.8		60.6	79.1	78.7		74.4	77.3	75.1		68.7
	15	**80.3**	**72.7**		**63.6**	**79.7**	**79.2**		**74.8**	**79.8**	**76.7**		**70.4**
	18	69.4	72.6		64.3	79.6	79.1		75.4	78.4	76.7		70.9
18	0	29.2	29.2	48.1	36.5	65.2	65.2	75.1	72.9	-	-	65.5	-
	3	33.0	43.4		37.6	69.6	72.1		73.0	50.9	58.9		57.9
	6	39.4	51.1		43.2	73.2	74.3		72.9	57.7	64.9		58.8
	9	46.1	58.4		49.6	75.8	77.0		73.0	65.5	70.6		63.2
	12	55.1	63.9		53.9	78.0	77.9		73.3	73.1	73.5		65.5
	15	66.5	69.2		59.1	79.4	78.4		73.9	78.1	76.3		68.4
	18	**80.3**	**71.6**		**61.8**	**79.8**	**79.3**		**74.9**	**80.4**	**77.8**		**69.6**

CP:Change point. FH: FH test with parameter Optimal:Optimally weighted log-rank test under each lag model. The bold values denote situations in which "CP tested" correctly specified "True CP." In the FH test, the parameter γ was selected to achieve the highest ARE under each lag model (Table S2). Max1:Max-combo test with the combinations $(\rho,\gamma) = (0,0),(0,1),(1,0),(1,1)$. Max2:Max-combo test with the combinations $(\rho,\gamma) = (0,0),(0,\gamma),(1,0),(1,1)$. The parameter in is the same value specified in the FH test under each lag model. For CP tested = 0 under threshold lag and linear lag, Optimal and FH reduced to the log-rank test, while Max2 reduced to $(\rho,\gamma) = (0,0),(1,0),(1,1)$.

4.3. Evaluation of the impact of misspecification in lag models and parameters under each lag model

We also evaluated the impact of misspecification when both lag models and change points were incorrectly specified. In Table 4, we show additional second simulation results, focusing especially on assuming the threshold lag or linear lag under generalized linear lag. In Table S3, we show all results of statistical power under misspecification with a) threshold lag, b) linear lag, and c) generalized linear lag.

Table 4. Robustness of the optimally weighted log-rank test, FH test, and Max-combo test when lag models and "CP (months) or FH γ" were misspecified and tested under generalized linear lag.

			Tested assuming threshold lag			Tested assuming linear lag		
True CP	Max1	CP Assumed	Optimal	FH	Max2	Optimal	FH	Max2
3	76.2	0	75.1	75.1	76.9	75.1	75.1	76.9
		3	79.4	78.5	77.1	79.8	78.3	77.2
		6	72.2	76.6	76.8	79.7	78.7	77.1
		9	65.2	72.3	76.1	78.2	77.9	77
		12	57.2	67.5	75.4	76.8	77.3	76.9
		15	48.6	59.1	74.6	75.4	76.6	76.8
		18	39.6	51.4	74	74	74.7	76.4
6	75	0	69.7	69.7	74.9	69.7	69.7	74.9
		3	77.8	77.2	75	75.1	75.2	74.9
		6	78	77.2	74.9	78.8	76.4	74.9
		9	70.4	75.5	74.9	79.3	77.4	74.9
		12	62.1	71.8	74.5	78.6	77.4	74.9
		15	53.3	64.5	73.6	77.4	77.2	74.9
		18	43.5	57.6	72.8	76.1	76.6	75
9	73.1	0	63.5	63.5	72.1	63.5	63.5	72.1
		3	71.9	74.4	72.1	68.9	70.7	72.1
		6	79	76.9	72.6	74.1	72.9	72.1
		9	76.2	76.8	73.1	77.8	75.6	72.2
		12	67.9	74.6	73	78.8	76.4	72.4
		15	58.6	69.3	72.2	78.4	76.9	72.6
		18	48.4	63.1	71.3	77.9	77	73
12	71.3	0	58	58	68.6	58	58	68.6
		3	65.2	71.3	68.6	62.7	66.1	68.6
		6	74.1	75.2	69.8	68	69	68.6
		9	78.8	76.9	71.6	73.2	73.1	68.8
		12	75.1	76.4	72.3	76.5	74.1	69.1
		15	65.9	74	71.9	77.8	75.2	69.8
		18	55.2	68.9	71	77.9	76.4	70.8
15	68.7	0	52.2	52.2	63.7	52.2	52.2	63.7
		3	59.3	66.9	63.9	57	61.3	63.8
		6	67.9	72.4	65.9	62.2	64.4	63.8
		9	75.4	75.7	69.2	67	68.8	64.2
		12	77.6	76.9	70.6	72	70.9	64.9
		15	72.7	75.8	71.2	75	72.4	65.9
		18	61.2	72.7	70.3	76.2	74.4	67.8
18	65.5	0	45.6	45.6	57.8	45.6	45.6	57.8
		3	52.5	62.2	58.2	50.1	55.3	57.8
		6	60.6	68.9	61.7	54.8	58.9	57.9
		9	70.5	74.3	66.3	60.8	64.9	58.8
		12	76.7	76.8	68.8	65.8	67.1	60.1
		15	77.3	77.8	70.3	70.5	68.9	61.7
		18	70	76.5	70.1	73.9	72.2	64.4

CP:Change point. FH: FH test with parameter Optimal:Optimally weighted log-rank test under each lag model. In the FH test, the parameter was selected to achieve the highest ARE under each lag model (Table S2). Max1: Max-combo test with the combinations $(\rho, \gamma) = (0,0), (0,1), (1,0), (1,1)$. Max2: Max-combo test with the combinations . The parameter γ in $(0, \gamma)$ is the same value specified in the FH test under each assumed lag model. For CP tested = 0 under threshold lag and linear lag, Optimal and FH reduced to the log-rank test, while Max 2 reduced to $(\rho, \gamma) = (0,0), (1,0), (1,1)$.

Across the misspecified lag models with true change points from 0 to 18, the FH was more beneficial than the optimally weighted log-rank test, Max-combo tests 1 and 2, when assuming threshold lag under generalized lag with a true change point of 12, 15, or 18. In this situation, the FH test was more robust in general than the optimally weighted log-rank test, and it maintained relatively higher (>70%) statistical power for each of the following scenarios: an assumed change point of 3 to 15 under a true change point of 12, an assumed change point of >6 under a true change point of 15, and an assumed change point of >9 under a true change point of 18. In this situation, the FH test also had statistical power equal to or higher than Max-combo tests 1 and 2.

On the other hand, compared with Max-combo tests 1 and 2, the FH test had worse performance when assuming threshold lag under linear lag with a true change point of 0, 3, or 6. In this situation, the FH test had lower statistical power than Max-combo tests 1 and 2, with an assumed change point of >9 under a true change point of 0, or an assumed change point of >12 under a true change point of 3 or 6. There were also several situations where the FH test showed worse performance (<70%) than Max-combo tests 1 and 2, such as the FH test assuming threshold lag under generalized lag with an assumed change point of >12 under a true change point of 3, or an assumed change point of >15 under a true change point of 6.

In these situations, the FH test was commonly less robust than Max-combo tests 1 and 2 when the assumed change point was misspecified relatively later (e.g., 12, 15, 18) under the early true change point (e.g., 0, 3, 6). Compared to optimally weighted log-rank test, the FH test was more robust under any misspecified lag, and generally had similar (<5% difference) or higher statistical power. The primary situations where none of the four tests (FH test, optimally weighted log-rank test, Max-combo tests 1 and 2) had sufficient statistical power were as follows: tests assuming linear lag under threshold lag with a true change point of 15 or 18, and tests assuming generalized linear lag under threshold lag with a true change point of 18, in both cases regardless of assumed change point.

4.4. Evaluation of the FH test in various situations

Figure 3 shows the configurations of survival functions in each pattern for third simulation.

Table 5 summarizes the results of the third simulation.

As expected, the log-rank test had the highest statistical power for the proportional hazard (no lag) pattern. However, the statistical power was more than 20% lower than the test with the highest statistical power for several delayed patterns, such as threshold lag, optimal for FH test withγ: 2.0, γ: 3.0 and other patterns such as the crossing 1, crossing 2, and diminishing 2.

For the FH test with γ values of 0.5–3.0, the statistical power decreased relative to that of the standard log-rank test for the proportional hazard (no lag) pattern as γ increased. On the other hand, the highest statistical powers for the FH test with γ values of 2.0, 0.5, and 1.5 were achieved under threshold lag, linear lag, and generalized linear lag, respectively. For delayed patterns 5–9 based on the survival functions derived by Garès et al. (2014), the corresponding FH test with the γ value could achieve the highest statistical power as designed. For patterns 10–13, the FH test with various γ values retained sufficient statistical power for the crossing patterns, but had the lowest statistical powers in the diminishing patterns, and these powers decreased as the value of γ increased.

For the GB test, the statistical power was slightly lower than that of the standard log-rank test but was comparable to that of the log-rank test. Also, for the diminishing patterns, the statistical power was slightly lower than RMST with $t_{RMST} = 25.2$ and $t_{RMST} = 32$, but was higher than with other statistical tests. However, for the delayed and crossing patterns, the test had considerably lower statistical power.

For the RMST test with $t_{RMST} = 25.2$ and with $t_{RMST} = 32$, the statistical powers were highest with the diminishing 1 and diminishing 2 patterns. However, the statistical power was lower overall than for other tests, especially for delayed patterns 2–9.

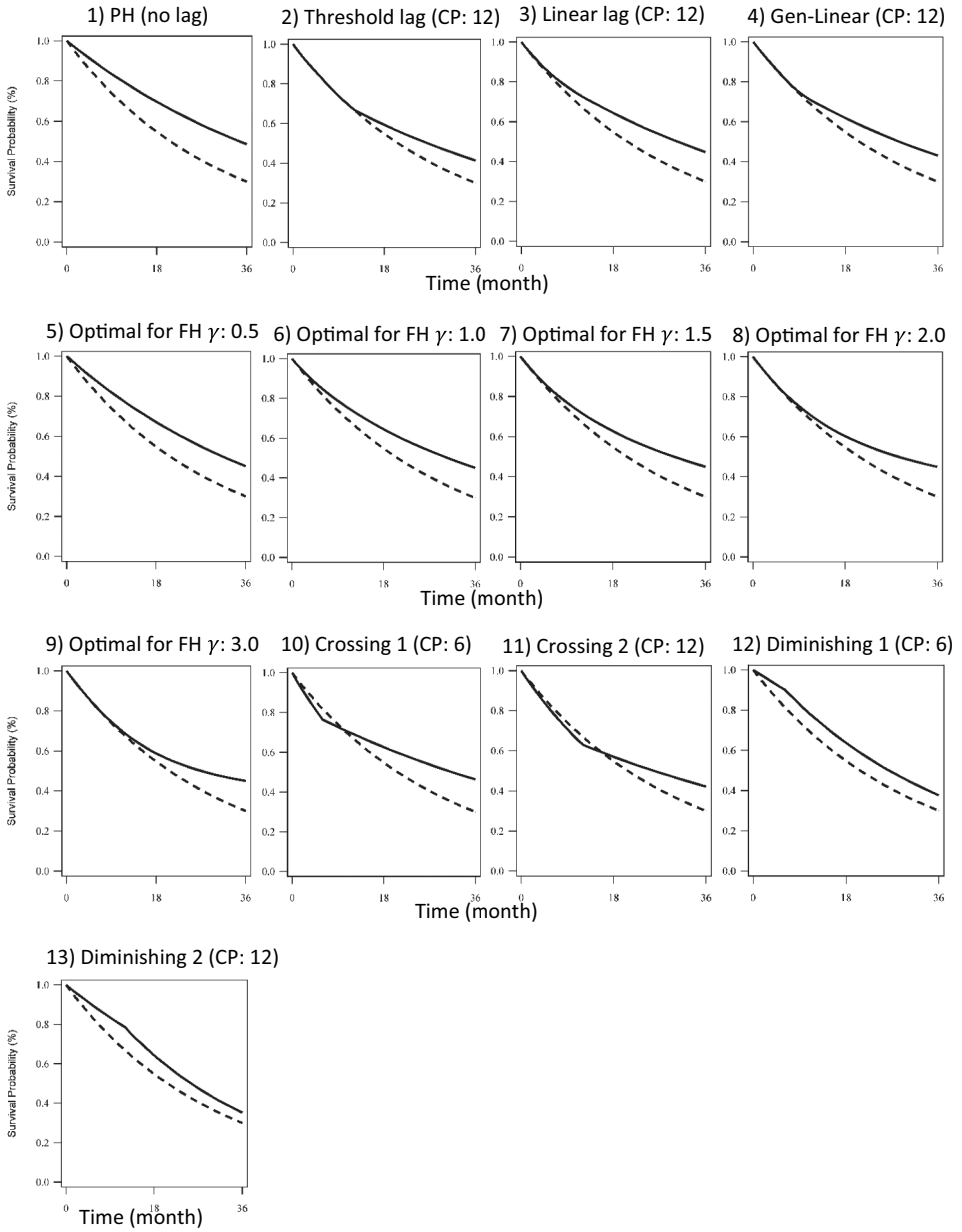


Figure 3. Survival functions for each pattern evaluated in the simulation.

For all patterns, the Max-combo test yielded a statistical power that was always 20% less than the test with the highest statistical power.

5. Application

Ferris et al. (2016) conducted an open-label, phase 3 trial of nivolumab for recurrent squamous cell carcinoma of the head and neck in patients whose disease had progressed within 6 months after platinum-based chemotherapy (CheckMate 141 study). The authors

Table 5. Comparison of the performance of the FH test with those of other tests under various patterns.

No.	Pattern	Log-rank	FH 0.5	FH 1.0	FH 1.5	FH 2.0	FH 3.0	GB	RMST 25.2	RMST 32	Max-combo
1	PH (no lag)	**95.3**	92.7	88.3	82.7	76.6	65.6	93.9	87.4	92.5	94.4
2	Threshold lag (CP: 12)	33.2	48.3	55.1	57.7	**57.8**	54.9	19.4	9.7	16.9	47.6
3	Linear lag (CP: 12)	69.7	**79.1**	78.7	75.8	72.3	64.2	55.0	36.0	50.6	76.4
4	Generalized linear lag (Cp: 12)	50.3	64.8	68.7	**68.8**	66.8	60.4	33.4	18.6	29.7	62.9
5	Optimal for FH γ: 0.5	71.8	**77.1**	75.4	72.1	68.4	61.0	60.4	43.1	56.2	74.9
6	Optimal for FH γ: 1.0	63.0	74.2	**75.7**	74.8	72.9	68.0	46.1	27.7	41.8	71.6
7	Optimal for FH γ: 1.5	56.3	70.5	74.9	**76.0**	75.5	72.7	37.3	19.6	32.5	68.8
8	Optimal for FH γ: 2.0	51.3	67.5	73.8	76.2	**76.6**	75.6	30.8	15.1	25.9	66.9
9	Optimal for FH γ: 3.0	43.5	61.9	70.8	75.1	77.2	**78.9**	23.1	10.2	18.8	62.8
10	Crossing 1 (CP: 6)	61.4	86.7	91.8	**92.2**	91.4	87.0	33.0	13.5	28.3	88.2
11	Crossing 2 (CP: 12)	27.4	52.9	66.3	73.0	76.3	**76.9**	10.1	4.9	8.2	57.5
12	Diminishing 1 (CP:6)	46.2	26.5	17.8	14.2	12.3	10.5	54.8	**57.6**	55.6	46.2
13	Diminishing 2 (CP:12)	35.8	17.2	9.8	7.2	6.3	5.5	49.7	**58.1**	53.0	39.9

CP: change point; Log-rank: Log-rank test; FH 0.5: FH test with $\gamma = 0.5$; FH 1.0: FH test with $\gamma = 1.0$.
FH 1.5: FH test with $\gamma = 1.5$; FH 2.0: FH test with $\gamma = 2.0$; FH 3.0: FH test with $\gamma = 3.0$.
GB: Gehan–Breslow test; RMST 25.2: RMST with $t_{RMST} = 25.2$; RMST 32: RMST with $t_{RMST} = 32$.
Max-combo: Max-combo test.

planned to assign 360 patients in a 2:1 ratio to nivolumab or standard therapy. The primary endpoint was overall survival (OS).

In the design stage, the hazard ratio of nivolumab in the control arm was assumed to be 0.667, which was equivalent to 50% improvement in median OS, i.e., 9 months in the nivolumab arm compared to 6 months in the control arm. The planned trial duration was 25 months, with uniform accrual for 14 months during the planning stage. Under these assumptions, it was estimated that 360 patients would ensure 90% statistical power with a two-sided α level of 0.05.

The actual study result was shown to be statistically significant using a conventional (stratified) log-rank test. However, separation of OS curves between the treatment groups was noticed approximately 4 months after the randomization.

We retrospectively considered redesigning the study using all of these study parameters except for the condition of delayed separation, focusing on OS. Since the effects of immune-oncologic agents are often delayed relative to the initial mechanism of action, we assumed a threshold lag with change point t^*. We used the primary endpoint of OS. Then, we planned to conduct hypothesis testing for the primary endpoint with an α level of 0.05.

Table 6 shows the total sample sizes needed for different change points for OS, and different hazard ratios after the change points. As described in Section 3, we selected the γ value in the FH test, which corresponds to the assumed change point according to the ARE, and used equation (7) to calculate the sample size necessary to achieve 90% statistical power to detect the difference between two arms for the primary endpoints of OS when the alternative hypothesis is true. In these ARE calculations, ARE was not impacted by the value of the hazard ratio after the separation, but instead was impacted by the value of the change point. In all patterns, our proposed method succeeded in providing the necessary sample size, reflecting the change point into the value of the late parameter γ.

Table 6. The relationship between change point, hazard ratio after the change point, and total sample size n in the FH test, using γ selected with the ARE.

No	Change point	HRs after change point	Selected γ value (ARE value at selected γ)	Total sample size
1	3	0.5	$\gamma = 0.8$ (ARE = 0.89)	222
2	6	0.5	$\gamma = 1.9$(ARE = 0.85)	375
3	9	0.5	$\gamma = 3.5$(ARE = 0.81)	675
4	3	0.6667	$\gamma = 0.8$ (ARE = 0.89)	612
5	6	0.6667	$\gamma = 1.9$(ARE = 0.85)	1020
6	9	0.6667	$\gamma = 3.5$(ARE = 0.81)	1818

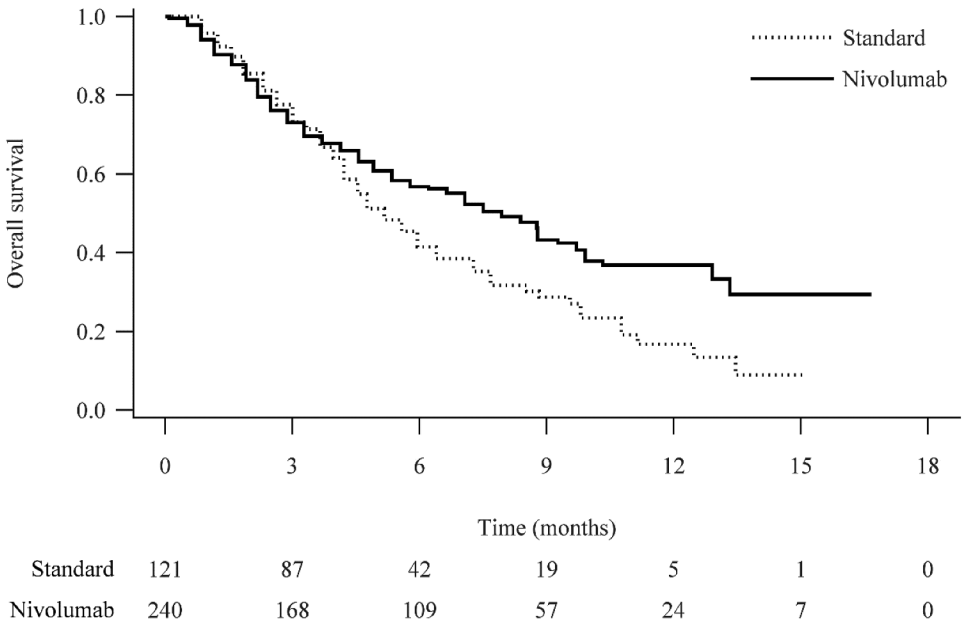


Figure 4. Kaplan–Meier curves for the reproduced OS in the CheckMate 141 study.

We also retrospectively attempted to conduct the CheckMate 141 study using the reproduced data of 361 individual patients. We set the primary endpoint of OS, and tested it with no interim analysis. The procedure to reproduce the patient-level data was conducted in accordance with Guyot et al. (2012).

As shown in Figure 4, the reproduced OS with HR = 0.681 (95% CI 0.517–0.897) was nearly identical to the reported values in the original publication.

To conduct the primary analysis retrospectively, we set the value of late parameter $\gamma = 1.1$ for OS so that the change point was 4 months. The proposed FH test showed a statistically significant between-arm difference for OS ($p < 0.001$), while the original (stratified) log-rank test result was also statistically significant in OS ($p = 0.0055$). This result suggests that our proposed method can be useful in successfully coping with delayed effects thereby increasing the statistical power, compared with the conventional log-rank test.

6. Discussion

There are several limitations to our proposed method to determine the value of γ. First, we fixed the parameter ρ in the FH test as 0, because we only focused on the delayed effect situation in immuno-oncology and only sought to determine the parameter γ ($\gamma > 0$) with $\rho = 0$. There exists another type of survival curve, known as a "crossing" as shown in a phase III study of stage IV or

recurrent NSCLC and a PD-L1 expression level of 1% or more (Hellmann et al. 2019). In a "crossing" curve, the survival curve indicates worse outcomes for treatment than for control in early stages, but demonstrates treatment benefit in later stages. In such a case, no weights or less weight in early stages, and more weights in later stages, as FH test with various γ values, sometimes difficult to interpret (Shen et al. 2023). Moreover, as shown in Magirr (2021) for the FH test with $(\rho, \gamma) = (0, 1)$, when the survival curve in the treatment arm is always lower than that in the control group, our proposed approach and other weighted approaches slightly inflate Type I error. We did not focus on such a situation. Second, within the delayed effect situations, we need to determine the lag pattern and change point before initiating the study. In situations where a priori knowledge about the late effect in the immunotherapy test drug is available and be reliably applied to the study, the lag pattern and change point can be derived based on when and how to achieve the full effect, and the corresponding parameter γ in the FH test can be determined by our proposed approach. In addition, if it is certain that the delayed effect will occur but the change point is uncertain before the study, to avoid severe loss of statistical power due to misspecification it may be preferable to conservatively define the change point as a later value (e.g., a higher value in the historical data) rather than an earlier value, as shown in Table 2.

Moreover, in terms of practically determining γ in the FH test even if the change point is uncertain, it is better to specify γ in an approximate range of 0.5–1.5, because as shown in Table S1, this will prevent severe loss of statistical power (i.e., over 60%) except in extreme cases where the change point is defined as 15 or 18 months of threshold lag. However, despite these considerations, incorrectly specifying the assumed lag pattern or change point may result in loss of statistical power. Moreover, we assumed that the event time in the control group was exponentially distributed, since this is usually the case in trials of cytotoxic agents. Finally, we assumed that censoring was only uniform during the recruiting period. If other censoring assumptions are expected, such as random censoring, our approach may not be applicable.

The calculation of the ARE between optimally weighted log-rank test and the FH test with the parameter γ ($\gamma > 0$) with $\rho = 0$ was implemented in R, and the R code is available in the Supporting Information. Sample size calculations for FH tests under each lag model were also implemented in R, with the R code available in the Supporting Information.

Acknowledgements

The authors thank the associate editor and the referees for their constructive comments and suggestions, which helped improve this article.

The authors thank Dr. Valérie Garès for kindly sharing the computation code that we used in the simulation study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research of Satoshi Morita was partially supported by the Japan Agency for Medical Research and Development (AMED) under Grant Number [JP20lk0201061].

ORCID

Yuichiro Kaneko http://orcid.org/0000-0002-6369-1115

References

Cristescu, R., R. Mogg, M. Ayers, A. Albright, E. Murphy, J. Yearley, X. Sher, X. Q. Liu, H. Lu, M. Nebozhyn, et al. 2018. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade–based immunotherapy. *Science: Advanced Materials and Devices*. 362(6411):eaar3593. doi:10.1126/science.aar3593.

Ding, X., and J. Wu. 2020. Designing cancer immunotherapy trials with delayed treatment effect using maximin efficiency robust statistics. *Pharmaceutical Statistics* 19 (4):424–435. doi:10.1002/pst.2003.

Ferris, R. L., G. Blumenschein Jr, J. Fayette, J. Guigary, A. D. Colevas, L. Licitra, K. Harrington, S. Kasper, E. E. Vokes, C. Even, et al. 2016. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *The New England Journal of Medicine*. 375(19):1856–1867. doi:10.1056/NEJMoa1602252.

Fleming, T. R., and D. P. Harrington. 1991. *Counting processes and survival analysis*. New York: John Wiley and Sons.

Garès, V., S. Andrieu, J. F. Dupuy, and N. Savy. 2014. A comparison of constant piecewise weighted test and Fleming Harrington's test — Application in clinical trials. *Electronic Journal of Statistics* 8 (1):841–860. doi:10.1214/14-EJS911.

Garès, V., S. Andrieu, J. F. Dupuy, and N. Savy. 2015. An omnibus test for several hazard alternatives in prevention randomized controlled clinical trials. *Statistics in Medicine* 34 (4):541–557. doi:10.1002/sim.6366.

Garès, V., S. Andrieu, J. F. Dupuy, and N. Savy. 2017. On the Fleming–Harrington test for late effects in prevention randomized controlled trials. *Journal of Statistical Theory and Practice* 11 (3):418–435. doi:10.1080/15598608.2017.1295889.

Gastwirth, J. L. 1985. The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association* 80 (390):380–384. doi:10.1080/01621459.1985.10478127.

Gill, R. D. 1980. Censoring and stochastic integrals. In *of Mathematical Centre Tracts*, Vol. 124, Amsterdam: Mathematisch Centrum Amsterdam.

Guyot, P., A. E. Ades, M. J. Ouwens, and N. J. Welton. 2012. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology* 12 (1):9. doi:10.1186/1471-2288-12-9.

Hasegawa, T. 2014. Sample size determination for the weighted log-rank test with the Fleming-Harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics* 13 (2):128–135. doi:10.1002/pst.1609.

Hellmann, M. D., L. Paz-Ares, R. B. Caro, B. Zurawski, S. W. Kim, E. C. Costa, K. Park, A. Alexandru, L. Lupinacci, E. Jimenez, et al. 2019. Nivolumab plus ipilimumab in advanced non–small-cell lung cancer. *The New England Journal of Medicine*. 381(21):2020–2031. doi:10.1056/NEJMoa1910231.

Hoos, A., A. M. Eggermont, S. Janetzki, F. S. Hodi, R. Ibrahim, A. Anderson, R. Humphrey, B. Blumenstein, L. Old, and J. Wolchok. 2010. Improved end points for cancer immunotherapy trials. *Journal of the National Cancer Institute* 102 (18):1388–1397. doi:10.1093/jnci/djq310.

Hussain, M., D. Rathkopf, G. Liu, A. Armstrong, W. Kevin Kelly, A. Ferrari, J. Hainsworth, A. Joshi, R. R. Hozak, L. Yang et al. 2015. A randomised non-comparative phase ii trial of cixutumumab (imc-a12) or ramucirumab (imc-1121b) plus mitoxantrone and prednisone in men with metastatic docetaxel-pretreated castration-resistant prostate cancer. *European Journal of Cancer (Oxford, England: 1990)*. 51(13):1714–1724. doi:10.1016/j.ejca.2015.05.019.

Lin, R. S., J. Lin, S. Roychoudhury, K. M. Anderson, T. Hu, B. Huang, L. F. Leon, J. J. Z. Liao, R. Liu, X. Luo et al. 2020. Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. *Statistics in Biopharmaceutical Research*. 12(2):187–198. doi:10.1080/19466315.2019.1697738.

Liu, S., C. Chu, and A. Rong. 2018. Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. *Pharmaceutical Statistics* 17:541–554. doi:10.1002/pst.1878.

Magirr, D. 2021. Non-proportional hazards in immuno-oncology: Is an old perspective needed? *Pharmaceutical Statistics* 20 (3):512–527. doi:10.1002/pst.2091.

Nikitin, Y. 2011. Asymptotic relative efficiency in testing. In *International encyclopedia of statistical science*, ed. M. Lovric, 72–75. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-04898-2_127.

Prentice, R. L. 1978. Linear rank tests with right censored data. *Biometrika* 65 (1):167–179. doi:10.1093/biomet/65.1.167.

Sharma, P., S. Hu Lieskovan, J. A. Wargo, and A. Ribas. 2017. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell* 168 (4):707–723. doi:10.1016/j.cell.2017.01.017.

Shen, Y., X. Wang, M. Sirisha, F. Mulkey, J. Zhou, X. Gao, L. Zhang, T. Gwise, S. Tang, M. Theoret, et al. 2023. Nonproportional hazards - an evaluation of the MaxCombo test in cancer clinical trials. *Statistics in Biopharmaceutical Research* 15 (2):300–309. doi:10.1080/19466315.2021.20084.

Xu, Z., B. Zhen, Y. Park, and B. Zhu. 2016. Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine* 36 (4):592–605. doi:10.1002/sim.7157.

Ye, T., and M. Yu. 2018. A Robust approach to sample size calculation in cancer immunotherapy trials with delayed treatment effect. *Biometrics* 74 (4):1292–1300. doi:10.1111/biom.12916.

Zucker, D. M., and E. Lakatos. 1990. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 77 (4):853–864. doi:10.1093/biomet/77.4.853.

Appendix.A Hazard and survival functions defined by the three lag models

- Hazard and survival functions in the control group are given as

$$\lambda_0(t) = \lambda_0 \text{ and } S_0(t) = \exp(-\lambda_0 t), \text{ respectively.}$$

- Hazard and survival functions in the new treatment arm are defined as follows:

1. No lag (proportional hazard assumption):

$$\lambda_1(t) = \theta\lambda_0,\ S_1(t) = \exp(-\lambda_0\theta t)$$

2. Threshold lag:

$0 \le t < t^*$, $\lambda_1(t) = \lambda_0,\ S_1(t) = \exp(-\lambda_0 t)$

$t^* \le t$, $\lambda_1(t) = \theta\lambda_0,\ S_1(t) = \exp(-\lambda_0 t^* - \theta\lambda_0(t - t^*))$

3. Linear lag:

$0 \le t < t^*$, $\lambda_1(t) = \lambda_0\left(\frac{t}{t^*}(\theta - 1) + 1\right)$

$$S_1(t) = \exp\left(-\frac{1}{2}\lambda_0\frac{1}{t^*}(\theta - 1)t^2 - \lambda_0 t\right)$$

$t^* \le t$, $\lambda_1(t) = \theta\lambda_0$,

$$S_1(t) = exp\left(\frac{1}{2}\lambda_0\theta t^* - \frac{1}{2}\lambda_0 t^* - \lambda_0\theta t\right)$$

4. Generalized linear lag:

$0 \le t < t_1^*$, $\lambda_1(t) = \lambda_0,\ S_1(t) = \exp(-\lambda_0 t)$

$t_1^* \le t < t_2^*$, $\lambda_1(t) = \lambda_0\left\{\frac{t - t_1^*}{t_2^* - t_1^*}(\theta - 1) + 1\right\}$

$$S_1(t) = exp\left(-\lambda_0 t_1^* - \tfrac{1}{2}\lambda_0\tfrac{t - t_1^*}{t_2^* - t_1^*}\left(\theta t - \theta t_1^* - t - t_1^* + 2t_2^*\right)\right)$$

$t_2^* \le t$, $\lambda_1(t) = \theta\lambda_0$

$$S_1(t) = exp\left(-\tfrac{1}{2}\lambda_0\left(t_1^* - \theta t_1^* - \theta t_2^* + t_2^* + 2\theta t\right)\right).$$