HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2019 December 30.

Published in final edited form as:

Stat Med. 2018 December 30; 37(30): 4589–4609. doi:10.1002/sim.7937.

Designing cancer immunotherapy trials with random treatment time-lag effect

Zhenzhen Xu[a,*], **Yongsoek Park**[#b], **Boguang Zhen**[a], and **Bin Zhu**[c]

[a]Food and Drug Administration, CBER, Silver Spring, MD 20993, U.S.A

[b]University of Pittsburgh, Department of Biostatistics, Pittsburgh, PA 15260, U.S.A.

[c]National Cancer Institute, DCEG, Bethesda, MD 20892, U.S.A.

[#] These authors contributed equally to this work.

Abstract

In some clinical settings such as the cancer immunotherapy trials, a treatment time-lag effect may be present and the lag duration possibly vary from subject to subject. An efficient study design and analysis procedure should not only take into account the time-lag effect but also consider the individual heterogeneity in the lag duration. In this paper, we present a Generalized Piecewise Weighted Logrank (GPW-Logrank) test, designed to account for the random time-lag effect while maximizing the study power with respect to the weights. Based on the proposed test, both analytic and numeric approaches are developed for the sample size and power calculation. Asymptotic properties are derived and finite sample efficiency is evaluated in simulations. Compared to the standard practice ignoring the delayed effect, the proposed design and analysis procedures are substantially more efficient when a random lag is expected; further, compared to the existing methods by Xu et al. (2017) [1] considering the fixed time-lag effect, the proposed approaches are significantly more robust when the lag model is mis-specified. An R package (DelayedEffect.Design) is developed for implementation.

Keywords

Clinical trial; Cancer immunotherapy; Non-proportional hazards assumption; Random time-lag effect; Sample size and power calculation; Treatment time-lag effect

1. Introduction

In clinical trials with time-to-event endpoints, the regular logrank test is most commonly used for primary analysis as well as sample size and power calculation. This test is asymptotically the most powerful non-parametric test when the proportional hazards assumption holds. In some instances, however, a time lag is present before the treatment becomes fully effective and the duration of lag may vary from subject to subject. For example, cancer immunotherapy, one of the most promising advances in cancer therapy

[*]Correspondence to: Zhenzhen Xu, Food and Drug Administration, CBER, Silver Spring, MD 20993, U.S.A Zhenzhen.Xu@fda.hhs.gov.

during recent years, commonly demonstrates a delayed onset of treatment effect in clinical trials [2, 3, 4, 5, 6, 7]. The delay is largely due to the indirect mechanism-of-action of the therapeutic agent, as its underlying basis is to unleash the immune system to fight cancer instead of directly poisoning a tumor with chemotherapy or destroying it with radiation. This indirect mechanism requires the time to mount an effective immune response and the time for that response to be translated into an observable clinical response. As a result, the presence of a time lag in the treatment effect results in a delayed separation of survival curves between the experimental and control groups in clinical trials with time-to-event endpoints and violates the proportional hazards assumption. Furthermore, each patient may respond to the same therapy in a different biological manner and each study may be conducted under heterogeneous clinical settings. Thus, the duration of lag is more suitable to be treated as a random variable rather than a fixed constant. To the best of our knowledge, the existing literature dealing with a time-lag effect always assumes a fixed lag duration [8, 9, 10, 11, 12, 13, 14, 15]. Hence, the question of interests becomes how to properly incorporate the random time-lag effect into the study design and data analysis when the proportional hazards assumption no longer holds.

Xu et al. (2017) [1] introduces a weighted logrank type of statistics, Piecewise Weighted Logrank (PW-Logrank) test, and two methods for sample size and power calculation, to properly and efficiently incorporate the delayed effect into the study design and data analysis. However, this design assumes that the duration of lag time is fixed at a constant t^* and t^* can be properly specified in advance. While this approach has generated enthusiasm, some controversy exists over its practical feasibility of properly pre-specifying t^* and robustness against mis-specifying it, as the true value is unknown in practice. Moreover, the duration of lag may vary heterogeneously from subject to subject rather than fix at a constant; the investigators may feel more comfortable to account for this individual heterogeneity in the study design and data analysis. This motivated our interest and led to this extension. In this paper, we first extend the Piecewise Weighted Logrank (PW-Logrank) test to a Generalized Piecewise Weighted Logrank (GPW-Logrank) test, in order to efficiently account for the delayed effect under the *random lag duration scenario*. In this scenario, we assume that each treated individual takes a specific time t^*_{ind} to develop the full treatment effect, where t^*_{ind} follows a random distribution such as a uniform distribution between T_1 and T_2. $[T_1, T_2]$ defines the support of the general lag time distribution. From a practical perspective, T_1 and T_2 represent the individual patient's shortest and longest possible lag time before the full treatment effect is manifested, respectively. If we measure the treatment effect by comparing the experimental and control arms depicted using a Kaplan-Meier plot, the two survival curves will not separate until T_1, then gradually separate at an increasing hazard ratio till T_2, and remain at a constant hazard ratio afterwards. Figure 1 illustrates the aforementioned delayed pattern where the individual lag time in treatment effect ranges between 3 and 12 months under a hypothetical setting. This figure is generated using a synthetic dataset simulated based on a confidential real pivotal study, where the simulation parameters were adjusted in order to mimic the delayed pattern observed in the real study. Similar patterns can be observed in the literature [5, 6, 7].

In contrast to the existing large family of weighted logrank tests, we show that GPW-Logrank test is the asymptotically most powerful weighted logrank test under the random

lag duration scenario, by optimally allocating respective weights to different subsets of events that occur prior to T_1, between T_1 and T_2, and after T_2. Further, we develop two approaches for the sample size and power calculation based upon GPW-Logrank test:

1. **A**nalytic **p**ower calculation method based on generalized **p**iecewise weighted **l**og-rank t**e**st with random treatment time-lag effect (APPLE+),
2. **S**imulation-based **e**mpirical **p**ower calculation method based on generalized **p**iecewise weighted **l**og-rank t**e**st with random treatment time-lag effect (SEPPLE+).

Under the pre-specified random lag duration scenario, APPLE+ provides a close-form solution to derive the sample size and power relationship, assuming a uniformly distributed lag model, and SEPPLE+ offers a numeric method to achieve the same goal, with the flexibility to incorporate more complex enrollment process, lag models and event time distributions. Finally, we evaluate the efficiency and robustness of the proposed methods theoretically, supported by empirical studies in finite sample situations. An R package (DelayedEffect.Design) is developed to implement the proposed procedures and available on CRAN.

2. GENERALIZED PIECEWISE WEIGHTED LOGRANK TEST

Suppose that N subjects are randomized into the experimental group E and control group C with probability P_E and P_C ($P_E + P_C = 1$), respectively, with the primary aim to compare the survival probabilities between these two groups. Let D, with the size n_D, denote the set of indices of patients who experience the event of interest. At each distinct event time t_j, $j = 1, \ldots, n_D$, let $n_i(t_j)$ ($i \in \{E, C\}$) denote the number of subjects who are still at risk in the group i by time t_j and $X_j \in \{0, 1\}$ indicate whether the j^{th} event belongs to the experimental group. Thus, $p(t_j) = n_E(t_j)/\{n_E(t_j) + n_C(t_j)\}$ refers to the proportion of subjects at risk at time t_j in the experimental group.

In this article, we consider two scenarios of lag duration, the *fixed lag duration scenario* and *random lag duration scenario*. Under the fi*xed lag duration scenario*, the subject-specific lag time t^*_{ind} is assumed to be fixed at t^* for all subjects. Xu et al. (2017) [1] considers this scenario and proposes the PW-Logrank test to test the null hypothesis $H_0 : h_E(t) = h_C(t)$, where $h_E(t)$ and $h_C(t)$ are the underlying hazard functions for the experimental and control groups, respectively. The PW-Logrank test statistic S_w is constructed as follows:

$$S_w = \frac{\sum_{j \in D_1} w_1 \left\{ X_j - p(t_j) \right\} + \sum_{j \in D_2} w_2 \left\{ X_j - p(t_j) \right\}}{\left[\sum_{j \in D_1} w_1^2 p(t_j) \left\{ 1 - p(t_j) \right\} + \sum_{j \in D2} w_2^2 p(t_j) \left\{ 1 - p(t_j) \right\} \right]^{1/2}} \quad (1)$$

where D_1 and D_2 refer to the sets of indices of patients who died before and after t^*. Let $\lambda_1 \geq 1$ and $\lambda_2 > 1$ denote the hazard ratios before and after t^*, respectively. In PW-Logrank test, we allow the earlier and later events that occur before and after t^* to carry respective

weights $w_1 = \log(\lambda_1)/\{\log(\lambda_1) + \log(\lambda_2)\}$ and $w_2 = \log(\lambda_2)/\{\log(\lambda_1) + \log(\lambda_2)\}$. Xu et al. (2017) [1] shows that, for a given sample size, S_w can maximize the study power under the fixed lag duration scenario when the weight at each distinct event time is proportional to the log of the hazard ratio at that time.

Under the *random lag duration scenario*, the subject-specific lag time t^*_{ind} is considered to follow a distribution ranging between T_1 and T_2. Conditional on t^*_{ind}, the null and alternative hypotheses of hazard ratio are specified as

$$H_0: \lambda\left(t \middle| t^*_{ind}\right) = \frac{h_C\left(t \middle| t^*_{ind}\right)}{h_E\left(t \middle| t^*_{ind}\right)} = 1 \quad \text{vs.} \qquad H_1: \lambda\left(t \middle| t^*_{ind}\right) = \frac{h_C\left(t \middle| t^*_{ind}\right)}{h_E\left(t \middle| t^*_{ind}\right)} = \begin{cases} 1, & t \le t^*_{ind} \\ \lambda_2, & t > t^*_{ind} \end{cases}$$

Our goal is to test whether $\lambda_2 = 1$. Since t^*_{ind} is not observed, we cannot directly apply methods proposed in Xu et al. (2017) [1]. Instead, we can integrate out t^*_{ind} and test whether $\lambda_2 = 1$ on the aggregated group level, where the pattern of hazard ratio becomes a three-segments piecewise function under the alternative,

$$H_0: \lambda(t) = \frac{h_C(t)}{h_E(t)} = 1 \quad \text{vs.} \qquad H_1: \lambda(t) = \frac{h_C(t)}{h_E(t)} = \begin{cases} 1, & t \le T_1 \\ \lambda_2^{g(t)}, & T_1 < t < T_2 \\ \lambda_2, & t \ge T_2 \end{cases}$$

where $g(t) = log\{\lambda(t)\}/log(\lambda_2)$is a monotone, increasing continuous function bounded between 0 and 1 with t ranging from T_1 to T_2, whose properties are justified in Appendix. In other words, the individual level hazard ratio function displays a two-phase piecewise-constant pattern with a subject-specific change point, t^*_{ind}, whereas the study level hazard ratio function exhibits a three-phase pattern with two distinct change points, T_1 and T_2, under the alternative. This pattern indicates that the treatment has no detectable effect before T_1 when none of the treated subjects responds to the treatment, then becomes increasingly effective from T_1 to T_2 as treated subjects start to respond one after another, and remains fully effective after T_2 once all treated subjects, still alive, have responded. The increasing trend from T_1 to T_2 occurs due to the individual heterogeneity in patients' immuno-response to treatment.

A wide range of lag models can be specified, depending on how one believes the lag is distributed. In this article, we first consider a general class of lag models to examine the properties of the asymptotically most efficient weighted logrank test and then focus attention on a particular type of lag model, the uniformly distributed lag of length t^*_{ind}, to derive a close-form power calculation method.

To test the alternative hypothesis under the random scenario, PW-Logrank is no longer the most efficient test and thus needs to be generalized. The challenge arises when it comes to estimate $g(t)$, a function impossible to specify exactly in practice because it depends on the underlying survival distribution of patients. In order to circumvent this problem, we show that g(t) converges to the cumulative distribution function (CDF) of the lag time t^*_{ind}, $F_*(t)$

(Equation (9) in the Appendix). This will help us to identify the asymptotically optimal weights in the weighted logrank test in order to construct the fully efficient test under the random lag duration scenario, without knowing the underlying survival distribution. The asymptotic optimality of the resulting weighting scheme is guaranteed by Theorems 2.1 and 2.2, whose justifications are given in Appendix.

Theorem 2.1 Under the condition $\left\{ \lambda (t) \right\} = O\left(n^{-1/2}\right)$ as n $\rightarrow \infty$, the asymptotic power of conventional weighted logrank test is maximized when weights at event times are proportional to the log of the hazard ratios at those times.

Theorem 2.1 is first given by Schoenfeld (1981) [16] then provided by Xu et al. (2017) [1] with an alternative proof. Further, we show that

Theorem 2.2 Suppose that the lag time for each individual follows a distribution with CDF $F_*(t)$, under condition $\log\left\{\lambda_2\right\} = O\left(n^{-1/2}\right)$ as n $\rightarrow \infty$, the asymptotic power of the conventional weighted logrank test is maximized when weights at event times are proportional to $F_*(t)$ at those times.

Remark Throughout this paper we refer to the "maximum" or "optimal" power of weighted logrank test as locally asymptotically maximum or optimal power, since the derivation of the maximum power holds under the local asymptotic condition $\log\left\{\lambda_2\right\} = O\left(n^{-1/2}\right)$ as n $\rightarrow \infty$, i.e. as the sample size goes to infinity, the hazard ratio is close to one. This local asymptotic condition is required to ensure the asymptotic normality distribution of the weighted logrank test statistic as shown by Schoenfeld (1981) [16]. The empirical study in Section 6.1, however, reveals that the weighting scheme derived in Theorem 2.2 can still improve the study power of the conventional logrank test substantially in finite sample situations under moderate treatment effect.

Based on Theorem 2.2, the GPW-Logrank test is therefore proposed, with the general aim to optimize the asymptotic study power with respect to the pre-defined weights when the random delayed effect exists. The corresponding test statistic takes the following form:

$$S_{wr} = \frac{\sum_{k=1}^{3} \sum_{j \in D_k} w_k^*\left(t_j\right)\left\{X_j - p\left(t_j\right)\right\}}{\left[\sum_{k=1}^{3} \sum_{j \in D_k} \left[w_k^*\left(t_j\right)\right]^2 p\left(t_j\right)\left\{1 - p\left(t_j\right)\right\}\right]^{1/2}} \quad (2)$$

where the weighting scheme $w_1^*(t) = 0$, $w_2^*(t) = F_*(t)$ and $w_3^*(t) = 1$ is called asymptotically optimal as it maximizes the asymptotic study power. Here, D_1, D_2 and D_3 refer to the sets of indices of patients who had events before T_1, between T_1 and T_2, and after T_2, respectively. Clearly, the most efficient test statistic under the random lag duration scenario S_{wr} is a special form of the conventional weighted logrank test statistics [16, 17] by assigning piecewise time-dependent weights to three subsets of events whose event times are differentiated chronologically at the change points T_1 and T_2. Furthermore, S_{wr} is also an

generalization of S_w by extending a single, fixed change time t^* to a range of plausible change times distributed between T_1 and T_2 to reflect the randomness of t^* due to the individual heterogeneity.

If the lag t^*_{ind} follows a uniform distribution on $[T_1, T_2]$, then the asymptotically optimal weights $w^*(t)$ can be explicitly written as

$$w^*(t) = F_{*u}(t) = \begin{cases} w_1^*(t) = 0, & t \le T_1 \\ w_2^*(t) = (t - T_1)/(T_2 - T_1), & T_1 < t \le T_2 \\ w_3^*(t) = 1, & t > T_2 \end{cases} \quad (3)$$

Note that $F_{*u}(t)$ is the CDF of a uniform distribution on $[T_1, T_2]$ and the test statistic S_{wr} becomes

$$S_{wru} = \frac{\sum_{j \in D_2} \left(\frac{t - T_1}{T_2 - T_1} \right) \{ X_j - p(t_j) \} + \sum_{j \in D_3} \{ X_j - p(t_j) \}}{\left[\sum_{j \in D_2} \left(\frac{t - T_1}{T_2 - T_1} \right)^2 p(t_j) \{ 1 - p(t_j) \} + \sum_{j \in D_3} p(t_j) \{ 1 - p(t_j) \} \right]^{1/2}} \quad (4)$$

With equal allocation ratio and the censoring distributions being the same across arms, S_{wr} (or S_{wru})can achieve the maximum study power at a two-sided significance level of α as

$$Pow^* = \Phi \left\{ \frac{1}{2} \log(\lambda_2) \sqrt{\bar{d}_{23}} - Z_{1 - \frac{\alpha}{2}} \right\} + \Phi \left\{ -\frac{1}{2} \log(\lambda_2) \sqrt{\bar{d}_{23}} - Z_{1 - \frac{\alpha}{2}} \right\}, \quad (5)$$

where $\bar{d}_{23}$ denotes the expected weighted number of events accumulated after the earliest treatment onset (T_1) and $Z_{1 - \frac{\alpha}{2}}$ refers to the $100 \times (1 - \alpha/2)^{th}$ percentile of the standard normal distribution. The derivation of equation (5) is provided in Appendix. This results have an intuitive, clinical meaningful implication. Under the random lag duration scenario, the weighted logrank test with weights proportional to the CDF of lag time distribution is the most efficient test, which is essentially the regular logrank test taking into account only the events accumulated after the delayed onset. Intuitively thinking, if the treatment effect is not manifested until T_1, then the earlier events before T_1 would neither contribute to the detection of treatment effect nor comply with the proportional hazards assumption so should be ignored. In contrast, the later events after T_1 do contribute, to various extents, and should be included in the analysis exclusively. Among those later events, it also makes sense to assign full weights to the subset of events that occur after T_2 when every treated subject still alive has responded to the treatment, whereas only assign partial weights to the other subset of events that happen between T_1 and T_2 while treated subjects begin to respond one after another.

As implied by equation (5), the asymptotically maximum power is driven by the expected number of weighted events accumulated since the first subject develops the treatment effect $\bar{d}_{23}$. To design a study, however, the relationship between the power function and the number of subjects instead of the number of events is of particular interest. The APPLE and SEPPLE methods proposed by Xu et al. (2017) [1] provide analytic and numeric approaches for the sample size and power calculation under the fixed lag duration scenario; in what follows, we extend the analytic approach APPLE to APPLE+, and the numeric method SEPPLE to SEPPLE+, to incorporate the randomness of delayed effect into the power calculation.

3. APPLE+

In this section, we derive a close-form, analytic procedure, APPLE+, for the sample size and power calculation. In the analytic derivation, we assume t^*_{ind} follows a uniform distribution. Suppose that patients arriving a study of total duration τ follow a Poisson process, where the intensity rate of the process is denoted by a. During the enrollment period [0, A] since randomization, the expected number of patients recruited for the study is a $a \times 1$ Further, under equal allocation ratio, the expected number of patients enrolled in either experimental (E) or control (C) group during an infinitesimal period of time $[u, u + du]$ is $a \times du/2$, where u denotes the calendar time when an individual patient arrives the study since the study onset. Among those subjects, the weighted proportions who will experience an event during the calender time intervals $[u, u + T_1]$, $[u + T_1, u + T_2]$ and $[u + T2, \tau]$ are $F_{i1}(T_1)$, $\int_{T_1}^{T_2}\left[\frac{t-T_1}{T_2-T_1}\right]^2 f_{i2}(t)dt$ and $\int_{T_2}^{T-u} f_{i3}(t)dt$ respectively, where $F_{ik}(\cdot)$ and $f_{ik}(\cdot)$, $k \in \{1, 2, 3\}$, denote the cumulative distribution function and probability density function of the exponential distribution. Here the weighted proportions are derived by assuming that the event time T follows an exponential distribution with rate h_{i1} before t^*_{ind} and h_{i2} after t^*_{ind} $\left(h_{C_1} = h_{C_2} = h_C, h_{E_1} = h_C, h_{E_2} = h_E = h_C/\lambda_2\right)$ and by implementing the asymptotically optimal weighting function provided in equation (3) to maximize the study power. Integrating over u, the total expected weighted number of patients who would experience an event before T_1, $\bar{d}_{i1}$ between T_1 and T_2, $\bar{d}_{i2}$ and after T_2, $\bar{d}_{i3}$ are

$$\bar{d}_{i1} = \int_0^A \frac{a}{2} F_{i1}(T_1) du,$$

$$\bar{d}_{i2} = \int_0^A \frac{a}{2} \int_{T_2}^{T_1} \left[\frac{t - T_1}{T_2 - T_1}\right]^2 f_{i2}(t) dt du,$$

$$\bar{d}_{i3} = \int_0^A \frac{a}{2} \int_{T_2}^{\tau - u} f_{i3}(t) dt du.$$

Thus, the total expected number of weighted events accumulated during the post-delay phase after T_1 can be obtained as,

$$\bar{d}_{C23} = \bar{d}_{C2} + \bar{d}_{C3} = \frac{a}{2}\left[DA - \frac{e^{-h_c \tau}}{h_c}\left(e^{h_c A} - 1\right)\right]$$

$$\bar{d}_{E23} = \bar{d}_{E2} + \bar{d}_{E3} = \frac{a}{2}\left\{EA - \frac{C}{h_E^2}\left[e^{-h_E(\tau - A)} - e^{-h_E \tau}\right]\right\}$$

where

$$E = e^{-h_C T_1}\left\{\frac{2}{h_E^2(h_C - h_E)(T_2 - T_1)^3} - \frac{2h_E}{h_C^3(h_C - h_E)(T_2 - T_1)^3} + \frac{2}{h_C^2(T_2 - T_1)^2}\left[1 - \frac{3}{h_C(T_2 - T_1)}\right]\right\} + e^{-h_C T_2}\left\{-\frac{1}{(T_2 - T_1)(h_C - h_E)} + \frac{h_E}{h_C(h_C - h_E)(T_2 - T_1)}\left[1 + \frac{2}{h_C(T_2 - T_1)} + \frac{2}{h_C^2(T_2 - T_1)^2}\right] + \frac{1}{h_C(T_2 - T_1)}\left[1 + \frac{4}{h_C(T_2 - T_1)} + \frac{6}{h_C^2(T_2 - T_1)^2}\right]\right\} + e^{-h_C T_1} e^{-h_E(T_2 - T_1)}\left\{-\frac{2}{h_E(h_C - h_E)(T_2 - T_1)^2} - \frac{2}{h_E^2(h_C - h_E)(T_2 - T_1)^3}\right\}.$$

$$C = \frac{h_E\left[e^{-(h_C - h_E)T_1} - e^{-(h_C - h_E)T_2}\right]}{(T_2 - T_1)(h_C - h_E)},$$

$$D = \frac{2\left[e^{-h_C T_1} - e^{-h_C T_2}\right]}{h_C^2 (T_2 - T_1)^2} - \frac{2e^{-h_C T_2}}{h_C (T_2 - T_1)},$$

The quantities C, D and E do not have meaningful interpretations but they allow us to write the expressions for the two measures $\bar{d}_{C23}$ and $\bar{d}_{E23}$ more succinctly. It follows from $\bar{d}_{23} = \bar{d}_{C23} + \bar{d}_{E23}$ that the relationship between N and $\bar{d}_{23}$ is

$$\bar{d}_{23} = \frac{N}{2}(E + D) - \frac{NC}{2Ah_E^2}\left[e^{-h_E(\tau - A)} - e^{-h_E \tau}\right] - \frac{Ne^{-h_C \tau}}{2Ah_C}\left(e^{-h_C A} - 1\right) \quad (6)$$

Thus, the asymptotically optimal power of the study, Pow^*, given the sample size N is obtained indirectly through the relationship between Pow^* and $\bar{d}_{23}$ and that between $\bar{d}_{23}$ and N. The detailed derivation of APPLE+ is illustrated in Appendix.

4. SEPPLE+

One can also calculate the empirical power for any given sample size using a simulation-based procedure, SEPPLE+, under the random lag duration scenario. Given the sample size N, the enrollment duration A, the total study duration ˝, the domain of lag times $[T_1, T_2]$, the distribution of lag times $F_*(t)$, and the baseline hazard h_C as well as the hazard for the experimental group after the delay h_E, the simulation-based SEPPLE+ algorithm works as follows for each value of the assumed treatment effect $\lambda = h_C/h_E$

Step 1 Draw patients' enrollment times u from a Poisson process with intensity rate $a = N/A$;

Step 2 Randomize patients to the experimental or control groups and draw patients' event times T_{λ_2} from

- $T_{\lambda_2} \sim \text{pexp}(h_C, h_E)$ for subjects in the experimental arm, where pexp(·) denotes the piecewise exponential distribution function with the rate varying at t^*_{ind}. Here the lag time t^*_{ind} is drawn identically and independently from the pre-specified distribution $F_*(t)$ for each treated subject;

- $T_{\lambda_2} \sim \exp(h_C)$ for subjects in the control arm, where exp(·) denotes the regular exponential distribution;

Step 3 Define the observational times $Z = min\left\{T_{\lambda_2}, \tau - u\right\}$ and the event indicators $I = \left\{T_{\lambda_2} \leq \tau - u\right\}$;

Step 4 Apply GPW-Logrank test, with weights determined by maximizing the power function under the pre-specified random lag duration scenario, to compute the p-value p_{λ_2}. Specifically, when t^*_{ind} follows a uniform distribution on $[T_1, T_2]$, the asymptotically optimal weights are provided in formula (3);

Step 5 Repeat step 1 to 4 for a large number of B times and compute the power for the given treatment effect λ_2 as the proportion of iterations whose p_{λ_2} are less than or equal to α. The proposed SEPPLE+ procedure serves to verify the analytic approximation of empirical power by APPLE+ when the lag model is uniformly distributed. Furthermore, SEPPLE+ provides a more flexible approach for power calculation to incorporate any parametric assumption on the enrollment process as well as event time distribution or even relax the parametric assumption, in order to mimic practical scenarios. For example, SEPPLE+ can implement a non-homogeneous Poisson process for enrollment or a more complex event time distribution through simulation studies, which is difficult to accomplish in the analytic derivations.

5. Impact of Mis-specifying Lag Models

Apparently, the proposed APPLE+, SEPPLE+ and the corresponding GPW-Logrank test depend critically upon the proper pre-specification of lag model, particularly the domain of the lag distribution (i.e. $[T_1, T_2]$); but in practice, the true values of T_1, T_2 are unknown in advance. Although it is ideal to properly specify both T_1 and T_2, the mis-specification of either one or both of these bounds is sometimes inevitable. Let us imagine a practical scenario where a study is designed using APPLE+ based on a mis-specified domain $\left[t_1^m, t_2^m\right]$ and the study results are analyzed via GPW-Logrank test also based on $\left[t_1^m, t_2^m\right]$ Then would that selected design still end up being nearly optimal asymptotically? In this section, we examine the impact of mis-specifying the lag model on the study design analytically.

Lemma 5.1 Assume that the individual lag time t^*_{ind} follows a uniform distribution within domain $[T_1, T_2]$; if one over-(under-) estimates either one or both bounds of the domain, and calculates the sample size required to achieve the target power using APPLE+ or SEPPLE+ based on the mis-specified domain, then the true power of GPW-Logrank test given the estimated sample size will be greater than or equal to (less than or equal to) the target power.

Lemma 5.1 implies that if investigators are not certain about the exact time frame of treatment time-lag across individuals, then the recommendation is to slightly over-estimate

rather than under-estimate T_1 and/or T_2, in both design and analysis stages, in order to protect against the potential power loss. However, the penalty is that the study cost would be inflated, relative to the asymptotically optimal design, due to the increase in sample size; in some instances, the inflation could even be dramatic if the specified domain is too far away from the truth. Hence, a proper specification of the lag model is still very important in practice. The justification of Lemma 5.1 is provided in Appendix.

6. SIMULATION STUDIES

To assess the characteristics of the proposed design and analysis methods, we conduct extensive simulation studies to compare them with alternatives from various perspectives. For all simulations, the empirical power is computed using 10,000 replications and the significance level is set at two-sided 0.05.

6.1. Evaluation #1: Power Comparison with alternatives

In the first evaluation, we aim to compare the empirical powers of GPW-Logrank test, of PW-Logrank test [1] as well as of the regular logrank test under a simulated scenario where the subject-specific lag varies uniformly between one month and eleven months. To do that, we fix the power of APPLE+ at 80% as the sample size varies and repeat the following steps for each given sample size N between 200 and 1000:

Step 1 Fix the power of APPLE+ with a uniform lag on [1, 11] months at 80% and back calculate the hazard ratio after T_2, λ_2;

Step 2 For a given λ_2 and N, simulate the time-to-event data under the random lag duration scenario where t^*_{ind} ~ Uniform[1, 11] months and compute the empirical powers of the following testing methods:

- GPW-Logrank test assuming a random lag time varying between 1 month and 11 months, with weights $w_1^*(t) = 0,\ \text{if } t \le 1; w_2^*(t) = (t-1)/10,\ \text{if } 1 < t \le 11; w_3^*(t) = 1,$ if $t > 11$. This step is essentially computing the power using SEPPLE+;
- PW-Logrank test assuming a fixed lag time at 6 months, with weights $w_1^*(t) = 0,\ \text{if } t \le 6; w_2^*(t) = 1,\ \text{if } t > 6;$
- Logrank test ignoring the delayed effect.

Figure 2 displays the sample size and power relationship for various methods. The red solid line serves as the reference line representing 80% of power targeted by APPLE+. The simulation-based SEPPLE+ method based on GPW-Logrank test, as shown in the purple curve, appears to achieve a power very close to the target. In contrast, if the lag duration scenario is mistakenly assumed to be fixed and PW-Logrank test (green curve) is applied with the assumed constant lag time located at the median of the true lag domain, then the test would be underpowered by approximate 10%. Further, if the lag effect is ignored, the regular logrank test (light blue curve) is seriously underpowered.

Two conclusions can be drawn from this evaluation. First, the analytic power of APPLE+ approximates the true empirical power of SEPPLE+ really well under the random lag

duration scenario. This implies that the asymptotically optimal weighting scheme derived based on the asymptotic property is well behaved in finite sample situations. Second, when the lag effect is present and the lag time follows a uniform distribution, ignoring the lag effect leads to a serious loss of power under all parameter settings considered; on the other hand, taking into account the lag effect but ignoring the randomness of the lag duration by fixing it at the middle of lag domain, reduces the power loss but still falls short of the target power. In contrast, accounting for the randomness of lag effect in both design and analysis stages meets the exact target power.

Given the significant improvement of PW-Logrank test than the conventional Logrank test [1], in what follows, we evaluate the proposed GPW-Logrank test in comparison with PW-Logrank test only.

6.2. Evaluation #2: Impact of Mis-specification

In the second evaluation, we assess the robustness of proposed methods under various mis-specification scenarios. To do that, we examine the empirical study power when the lag model is mis-specified in either study design using APPLE+ or subsequent analysis using GPW-Logrank test or both. Three aspects of lag model mis-specification are considered: a) the underlying lag duration scenario (fixed vs. random) b) the domain of lag distribution as alluded to in Section 5 and c) the shape of lag distribution.

Specifically, in the following subsection 6.2.1 we evaluate the impact of mis-specifying the lag model in the subsequent analysis only, while the study is designed under the true model. In subsections 6.2.2 and 6.2.3, we explore this impact of mis-specification in both the design and analysis stages, since the ICH E9 guideline (1998) recommends a consistent method to be specified for both design and analysis. From the design perspective, mis-specifying the lag model for APPLE+ will result in an over- or under-estimate of the sample size required to achieve the target power. In addition, mis-specifying GPW-Logrank test, from the analysis perspective, will lead to a loss of power. A study, designed based on a mis-specified APPLE + and analyzed using a mis-specified GPW-Logrank test, could unnecessarily exceed or easily fall short of the target power. Hence, an investigation from the integrated design and analysis perspective is of particular interest.

For ease of notation, we let t^*, t^m be the true or mis-specified constant lag time and $[T_1, T_2], [T_1^m, T_2^m]$ be the true or mis-specified lag domain, respectively.

6.2.1. Impact of mis-specifying lag duration scenario or lag domain or both in analysis—

The proposed GPW-Logrank test accounts for the random lag duration scenario whereas PW-Logrank test [1] assumes the fixed lag duration scenario. When the true underlying scenario is fixed but mis-specified to be random or vice verse, the performances of these approaches become an interesting question to explore. For example, one could assume that the individual lag time varies heterogeneously between 3 months and 9 months and perform GPW-Logrank test whereas, in fact, each subject homogeneously takes 6 months to achieve the full effect so PW-Logrank test should be applied instead. Further, both approaches depend critically upon the pre-specification of lag domain; but in

practice, the true domain is unknown. Then, what if the lag domain of $[T_1, T_2]$ is mis-specified in GPW-Logrank test or the constant lag time t^* is mis-specified in PW-Logrank test? In an even worse scenario, both the lag duration scenario and lag domain could be mis-specified. In this subsection, we use a rainbow colormap (Figure 3) to illustrate the impact of mis-specification, of either the lag duration scenario or the lag domain or both, in the subsequent analysis when the sample size is calculated under the true lag model.

The X-axis and Y-axis in Figure 3 represent T_1 and T_2, respectively, and a rainbow of colors denote the different values of empirical power. The left panel is generated under the fixed lag duration scenario with t^* = 6 months based on 591 subjects and the right panel under the random scenario with T_1 = 3, T_2 = 9 months given 630 subjects. These sample sizes are sufficient to achieve the target power at 80% under the specific true lag model. The points in the off-diagonal area ($T_1 \neq T_2$) correspond to the power of GPW-Logrank test while the points on the diagonal line ($T_1 = T_2$) refer to the power of PW-Logrank test.

Under the fixed lag duration scenario (Left panel of Figure 3), examining the color of diagonal line reveals that PWLogrank test achieves its maximum power at the truth $T_1 = T_2 = t^*$ (black dot) and begins to lose power as $T_1 = T_2$ depart from the truth. If the deviation is more than roughly two months, PW-Logrank test will be subject to a severe power loss. In sharp contrast, a decent off-diagonal area colored in orange to red corresponding to about 5% or less loss of power suggests that GPW-Logrank test is robust under the fixed scenario provided that either T_1 or T_2 is not severely mis-specified.

When the underlying lag duration scenario is random (Right panel of Figure 3), the study power of GPW-Logrank test achieves its maximum at the truth T_1 = 3 and T_2 = 9 (black star) and close-to-maximum in the large oval-shape neighborhood of the star, indicating that the empirical power of the generalized test is close to its optimum as long as the pre-specified domain is not severely off the truth. It is interesting to note that as T_1 becomes smaller and T_2 becomes larger than the underlying truth, the upper left corner of the right panel remains in orange, suggesting that as long as the mis-specified domain covers the truth and is not overly spanned, the proposed generalized test remains very efficient. The most extreme case happens when T_1 = 0, T_2 = 12 months and in this case, this generalized test can still achieve over 75% of power. Interestingly, PW-Logrank test in the random lag duration scenario is apparently underpowered but the degree of power loss seems to be minimal when t^m is specified at the middle of the true domain.

To emphasize the robustness gain of GPW-Logrank test over PW-Logrank test, a range of key scenarios in this rainbow colormap are highlighted in Table 1. Under the fixed lag duration scenario where the true lag time t^* = 6 months, correctly specifying t^* in PW-Logrank test achieves 79% power, but under-specifying or over-specifying such single constant of lag time could result in a severe power loss. Particularly, if the mis-specification is as severe as 5 months (i.e., t^m = 1 or t^m = 11 months), the resultant power loss is nearly 17%. This loss reduces to approximate 10% if the mis-specification is less severe by 3 months (i.e., t^m = 3 or t^m = 9 months). In sharp contrast, mis-specifying a domain of lag times at [1, 11] months, [1, 9] months or [3, 11] months in GPW-Logrank test only leads to 1–2% power loss. This evaluation reveals that when the true lag duration scenario is fixed,

PW-Logrank could potentially suffer a severe power loss if t^* is mis-specified, whereas GPW-Logrank is robust to the power loss as long as the pre-specified lag domain covers the truth even though such domain is very wide. The practical implication of this finding is that if the investigator has less certainty about the fixed, constant lag time t^* under the fixed lag duration scenario, specifying a wide range of plausible values for it in GPW-Logrank test is a more sensible approach than picking a single guess in PW-Logrank test. In the same vein, the robustness of GPW-Logrank test is observed under the random lag duration scenario. For example, mis-specifying the lag domain to be wider than the true domain by two months either on one or both sides in GPW-Logrank test only lead to 1% power loss, whereas PW-Logrank could suffer a loss as significant as 14%. In short, GPW-Logrank test does an excellent job of guarding against the mis-specified risk on both lag duration scenario and lag domain relative to PW-Logrank test, where the latter is only robust when the constant lag time is set at the middle of true lag domain. In practice, the true lag domain or lag duration scenario is unknown, and due to the limited knowledge from pilot data, early phase studies or existing literature, it is difficult to determine which lag model is plausible; in addition, from the investigators' viewpoint, picking a single best guess is more difficult than eliciting a range of plausible values for lag parameters. Consequently, PW-Logrank test may not work well but GPW-Logrank test provides a robust approach. This finding re-affirms our motivation to generalize PW-Logrank test to GPW-Logrank test.

6.2.2. Impact of mis-specifying the lag distribution in both design and analysis—Throughout, we only consider a uniformly distributed lag model to derive the analytic power calculation method. In practice, the individual lag time may vary in a non-uniform fashion. In this evaluation, we therefore assess the robustness of the uniformly distributed lag assumption in both APPLE+ and GPW-Logrank test when the underlying truth differs in various Beta- or Gamma-distributed patterns.

Table 2 considers four different non-uniform lag patterns. As the treatment effect varies, APPLE+ always utilizes a uniform pattern to calculate the sample size required to achieve the target power. Given the calculated sample size, the empirical power of GPW-Logrank test, with the weighting scheme determined also on the basis of a uniform pattern, is examined under each scenario of mis-specification. Namely, under Scenarios *S1–S2*, the subject-specific lag follows a *Beta*(2, 3) or *Beta*(3, 2) distribution re-scaled between [2, 10] months, which imply that the majority of patients respond to the therapy either before or after the middle point 6 months, respectively. Scenario *S3* specifies a bell shape by a re-scaled *Beta*(2, 2) distribution, letting the majority to respond around 6 months after the administration of treatment. Finally, Scenario *S4* presumes a Gamma distributed pattern $\Gamma(9, 0.05)$, allowing some patients to react immediately ($T_1 = 0$) but very few others never respond ($T_2 = +\infty$), as evident in many immuno-oncology studies. To compare, we also use APPLE to calculate the sample size and PW-Logrank test for analysis with the constant lag time specified at the middle of lag domain, as subsection #6.2.1 suggests that this approach is most efficient with such lag parameter specification under the random scenario. The results show that GPW-Logrank test, given the weighting scheme and sample size determined based on a mis-specified uniform pattern, can be slightly overpowered under scenario *S1*, slightly underpowered under scenario *S2*, or maintain the target power under

scenarios *S3–S4*. In order to gain insight into the trends observed, we examine the relationship between the true lag pattern and study power analytically in Theorem 6.1.

Theorem 6.1 Consider two treatments 1 and 2 with different lag patterns. Let $F_*^{(1)}(t)$ and $F_*^{(2)}(t)$ denote the respective lag time cumulative distribution functions. Suppose that the treatment effects are the same after the lag time period and $F_*^{(1)}(t) \geq F_*^{(2)}(t)$ for all t, then the conventional weighted logrank test with the same positive weight assignment w(t) have at least as much asymptotic power to detect the effect of treatment 1 than that of treatment 2 given the same sample size and censoring distributions.

Theorem 6.1 provides a theoretical ground to the trends observed in Table 2. Since the CDF of true lag distribution, a re-scaled *Beta*(2, 3) on [2, 10] months, is greater than that of mis-specified uniform lag pattern at all time (Figure 4), GPW-Logrank test based on the sample size calculated to achieve the target power assuming a mis-specified uniform lag would have at least as much asymptotic power as the target (Scenario *S1*). In comparison, the proposed test would be slightly underpowered under a re-scaled *Beta*(3, 2) pattern, as the CDF of this re-scaled Beta lag is smaller than that of assumed uniform lag at all time between [2, 10] months (Scenario *S2*). The fact that the CDFs of a re-scaled *Beta*(2, 2) and Γ(9, 0.05) cross with that of uniform make it difficult to determine whether GPW-Logrank test would be overpowered or underpowered; in the case simulated, the power gain and loss corresponding to the two areas before and after the crossing point of CDF curves offset each other, leaving the overall study power close to the target (Scenarios *S3–S4*) (Table 2). Theorem 6.1 implies that, in practice, if one concerns about the potential power loss due to the mis-specification of underlying lag distribution, then one can slightly over-specify either one or both boundaries of the uniform lag domain, making the uniform CDF cross with or exceed the true lag CDF. For instance, the empirical power loss under scenario *S2* can be salvaged by over-specifying the uniform lag domain from [2, 10] to [2, 12] months (Table 3). In sum, the uniform lag assumption of APPLE+ and GPW-Logrank test is robust against the mis-specification of lag pattern in the cases investigated.

When it comes to the comparison between the proposed APPLE+ & GPW-Logrank test and the existing APPLE & PW-Logrank test, this evaluation suggests that, when the lag pattern is mis-specified but lag domain is correctly specified, the proposed generalized design and analysis approaches are more robust than the existing ones by improving the power by at least 4% − 7%.

6.2.3. Impact of mis-specifying the lag domain in both design and analysis—

This evaluation aims at examining the impact of lag domain mis-specification in both the design and analysis of clinical study when the random delayed effect exists. For comparison, both APPLE+ & GPW-Logrank test and APPLE & PWLogrank test are performed, where the fixed lag time t^m of the latter is taken to be the middle of mis-specified lag domain $\left[T_1^m, T_2^m\right]$ of the former. Given each design & analysis method, the empirical power is computed to verify how much power can be actually achieved through simulations.

Table 4 illustrates this impact of mis-specification empirically. For instance, when the true lag domain t^*_{ind}~Uniform$\left[T_1 = 2, T_2 = 10\right]$ months but is mis-specified at Uniform$\left[T_1^m = 1, T_2^m = 11\right]$ months in both APPLE+ and GPW-Logrank test, to achieve 80% power APPLE+ requires 366 patients; and a simulation study verifies that these many patients can actually achieve the exact target power empirically using GPW-Logrank test. On the contrary, APPLE with t^m = 6 months claims only 329 patients being sufficient to achieve the target power, whereas the empirical power of PW-Logrank test given 329 patients is only 72%. This evaluation suggests that the proposed design and analysis method, APPLE+ & GPW-Logrank test, is more robust than APPLE & PW-Logrank test combination, even when the latter specifies the constant lag time at the middle of lag domain of the former.

7. DISCUSSION

A typical challenge arising in cancer immunotherapy trials is the presence of treatment time-lag effect, where the duration of lag may vary heterogeneously across individuals. An efficient study design and analysis procedure need not only account for the treatment time-lag effect but also take into consideration the individual heterogeneity of the lag duration. To meet the challenge, we propose a weighted logrank type of statistics and two power calculation approaches. Compared to the existing literature, the proposed methods have two advantages. First, GPW-Logrank test is designed to have asymptotically optimal efficiency over a wide range of lag models among the large family of weighted logrank tests. Second, the majority of the literature dealing with a time-lag effect assumes a homogeneous lag duration. This assumption may not be biologically plausible; even though is true, picking the best guess for a single value of such duration is subject to the risk of mis-specification. To mitigate this concern, APPLE+, SEPPLE+ and GPW-Logrank test are proposed which not only take into account the individual heterogeneity but also reduce the risk of mis-specification by permitting a range of plausible values in certain pattern to be specified for the lag time. Although the domain, pattern or delayed scenario of the lag model might still be mis-specified, theoretical and empirical studies demonstrate that these novel methods have good efficiency and robustness performance. Thus, it is reasonable to expect that these methods will advance the frontier of study design and data analysis for cancer immunotherapy trials. The enhanced efficiency is able to significantly reduce the study cost, shorten the drug development process and eventually benefit many cancer patients in a timely fashion for the near future.

To apply the proposed methods, investigators need to properly pre-specify the lag parameters such as the domain, pattern and delayed scenario of the lag model during the design stage without examining the trial outcome data. Historical studies, pilot data, and a good biological and medical judgment on the mechanism-of-action of the therapeutic agent can all help to specify the lag model. Alternatively, investigators can also estimate the lag model in an indirect manner by observing the pattern of certain immunological surrogate endpoints such as the antibody level from the early studies. In the case when the lower and/or upper bounds of the lag time are difficult to elicit, Lemma 5.1 suggests that slightly over-estimating the bounds offers some protection against the power loss than under-estimating them. Although the asymptotically optimal weighting scheme applies to all lag

models (Theorem 2.2), throughout, we only focus on a uniform lag to derive the analytic power calculation method. Since we have no strong prior information on the lag pattern either from the prior studies or currently available medical literature, we assume the individual lag time distributes uniformly within a certain range as an uniform distribution. In practice, the individual lag time may vary in a non-uniform fashion, for which the robustness of the uniform lag is evaluated and the possible remedy to potential power loss due to the mis-specification of lag distribution is also discussed in Section 6.2. Likewise, we only adopt an exponential distribution to model the event-time distribution in the derivation of the close-form sample size and power calculation method (APPLE+). More complex distributions for the lag model or event time can be assumed but may not lead to a close-form analytic solution as APPLE+. This is warranted for future research. Alternatively, one can always perform numeric SEPPLE+ to accommodate any complex model for lag time or event time.

Acknowledgements

The opinions and information in this article are those of the authors, and do not represent the views and/or policies of the U.S. Food and Drug Administration. This work was partially supported by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, MD, USA. The authors would like to thank Drs. Mitchell Gail and John Scott for helpful discussions. The simulation study used the high-performance computational capabilities of the Scientific Computing Laboratory at the Food and Drug Administration, Center for Devices and Radiological Health; the authors would also like to thank the super-computing support team staff especially Mr Mike Mikailov for providing excellent high-performance computing service support.

APPENDIX

In this Appendix, we first show that the asymptotically optimal weight assignments $w(t)$ at each distinct event time t are proportional to $F_*(t)$ at that time, under the condition $\log(\lambda_2) = O(n^{-1/2})$ as $n \to \infty$ (Theorem 2.2), then derives the close-form sample size and power calculation method, APPLE+, when the lag time follows a uniform distribution under the random lag duration scenario. Finally, we show the impact of mis-specification of the boundaries of t^*_{ind} on the study power.

A.1 Theorem 2.2 proof:

Under the random lag duration scenario, let t^*_{ind} denote the individual level hazard ratio change point or subject-specific lag time and $h_C(t)$ the hazard of the control patients at time t, then the hazard of the treated patients can be written as, where $\lambda = 1/\lambda_2$,

$$h_E\left(t \middle| t^*_{ind}\right) = \begin{cases} h_C(t), & \text{if } t < t^*_{ind} \\ \lambda h_C(t), & \text{if } t \geq t^*_{ind} \end{cases}$$

i.e.

$$h_E\left(t \middle| t^*_{ind}\right) = h_C(t)\left[1 + (\lambda - 1)I\left(t \geq t^*_{ind}\right)\right]$$

and the corresponding cumulative hazard functions are

$$H_E\left(t\middle|t^*_{ind}\right)=\begin{cases}H_C(t), & \text{if } t\le t^*_{ind}\\ \lambda H_C(t)-(\lambda-1)H_C\left(t^*_{ind}\right), & \text{if } t>t^*_{ind}\end{cases}$$

The probability density functions for the patients in the experimental and control groups are

$$\begin{aligned}
f_C(t) &= h_C(t)S_C(t)=h_C(t)e^{-H_C(t)}\\
f_E\left(t\middle|t^*_{ind}\right) &= h_E\left(t\middle|t^*_{ind}\right)S_E\left(t\middle|t^*_{ind}\right)=h_E\left(t\middle|t^*_{ind}\right)e^{-H_E\left(t\middle|t^*_{ind}\right)}\\
f_E(t) &= \int_{t^*_{ind}} f_E\left(t\middle|t^*_{ind}\right)f_*\left(t^*_{ind}\right)dt^*_{ind}=\int_{t^*_{ind}} h_E\left(t\middle|t^*_{ind}\right)e^{-H_E\left(t\middle|t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind}\\
S_E(t) &= \int_{t^*_{ind}} S_E\left(t\middle|t^*_{ind}\right)f_*\left(t^*_{ind}\right)dt^*_{ind}=\int_{t^*_{ind}} e^{-H_E\left(t\middle|t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind}
\end{aligned}$$

So,

$$\begin{aligned}
S_E(t) &= \int_t^\infty S_E\left(t\middle|t^*_{ind}\right)f_*\left(t^*_{ind}\right)dt^*_{ind}+\int_0^t S_E\left(t\middle|t^*_{ind}\right)f_*\left(t^*_{ind}\right)dt^*_{ind}\\
&= e^{-H_C(t)}\int_t^\infty f_*\left(t^*_{ind}\right)dt^*_{ind}+e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind}\\
&= e^{-H_C(t)}\{1-F_*(t)\}+e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind}
\end{aligned}$$

and

$$\begin{aligned}
f_E(t) &= -\frac{dS_E(t)}{dt}\\
&= h_C(t)e^{-H_C(t)}\{1-F_*(t)\}+e^{-H_C(t)}f_*(t)+\lambda H_C(t)e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind}-e^{-\lambda H_C(t)}e^{(\lambda-1)H_C(t)}f_*(t)\\
&= h_C(t)e^{-H_C(t)}\{1-F_*(t)\}+e^{-H_C(t)}f_*(t)+\lambda H_C(t)e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind}-e^{-H_C(t)}f_*(t)\\
&= h_C(t)e^{-H_C(t)}\{1-F_*(t)\}+\lambda H_C(t)e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)}f_*\left(t^*_{ind}\right)dt^*_{ind},
\end{aligned}$$

Thus,

$$\lambda(t) = \frac{h_C(t)}{h_E(t)} = \frac{S_E(t)h_C(t)}{f_E(t)}$$
$$= \frac{h_C(t)e^{-H_C(t)}\{1-F_*(t)\} + h_C(t)e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)} f_*\left(t^*_{ind}\right)dt^*_{ind}}{h_C(t)e^{-H_C(t)}\{1-F_*(t)\} + \lambda h_C(t)e^{-\lambda H_C(t)}\int_0^t e^{(\lambda-1)H_C\left(t^*_{ind}\right)} f_*\left(t^*_{ind}\right)dt^*_{ind}}$$
$$= \frac{1-F_*(t) + \int_0^t e^{(\lambda-1)\left[H_C\left(t^*_{ind}\right) - H_C(t)\right]} f_*\left(t^*_{ind}\right)dt^*_{ind}}{1-F_*(t) + \lambda\int_0^t e^{(\lambda-1)\left\{H_C\left(t^*_{ind}\right) - H_C(t)\right\}} f_*\left(t^*_{ind}\right)dt^*_{ind}}.$$

Let $A = \int_0^t e^{(\lambda-1)\left\{H_C\left(t^*_{ind}\right) - H_C(t)\right\}} f_*\left(t^*_{ind}\right)dt^*_{ind}$ and $S_*(t) = 1 - F_*(t)$. Then $dA/dt = f_*(t) - (\lambda-1)h_C(t)A$. Take derivation of $\lambda(t)$ over t, we have

$$\frac{d\lambda(t)}{dt} = \frac{\left[-(\lambda-1)h_C(t)A\right][S_*(t)+\lambda A] - \left[-f_*(t) + \lambda\left\{f_*(t) - (\lambda-1)h_C(t)A\right\}\right][S_*(t)+A]}{\{S_*(t)+\lambda A\}^2}$$
$$= (1-\lambda)\frac{\left\{f_*(t) - \lambda h_C(t)A\right\}\{S_*(t)+A\} + h_C(t)A\{S_*(t)+\lambda A\}}{\{S_*(t)+\lambda A\}^2} \quad (7)$$
$$= (1-\lambda)\frac{f_*(t)\{S_*(t)+A\} + (1-\lambda)h_C(t)AS_*(t)}{\{S_*(t)+\lambda A\}^2}.$$

When $t < T_1$, we have $F_*(t) = 0$, $S_*(t) = 1 - F_*(t) = 1$, $f_*(t) = 0$, so $\lambda(t) = 1$. When $t \geq T_2$, $F_*(t) = 1$ so $\lambda(t) = 1/\lambda = \lambda_2$. When $T_1 \leq t < T_2$, we let $\lambda(t) = \lambda_2^{g(t)} = (1/\lambda)^{g(t)}$ as defined under the alternative hypothesis of the random delayed duration scenario. In our delayed effect setting, $\lambda < 1$, so we easily see that $d\lambda(t)/dt \geq 0$ for all t, where $d\lambda(t)/dt = 0$ occurs when either $A = 0$ ($A = 0 \Rightarrow f_*(t) = 0$) or $S_*(t) = 0$ ($S_*(t) = 0 \Rightarrow f_*(t) = 0$). This implies that (t) is monotone increasing function over t when $0 < F_*(t) < 1$, so that $g(t)$ is a monotone, increasing continuous function bounded between 0 and 1 with t ranging from T_1 to T_2.

As shown in Theorem 2.1, under the condition $\log\{\lambda(t)\} = O(n^{-1/2})$ as $n \to \infty$, the asymptotic power of conventional weighted logrank test is maximized when weights at event times are proportional to the log of hazard ratios at those times. Next, we show that, under the same condition $\log\{\lambda(t)\} = O(n^{-1/2})$ as $n \to \infty$, i.e. $\lambda \to 1$ as $n \to \infty$, the log of the hazard ratio at the event time t, $\log\{\lambda(t)\}/\log(\lambda_2)$, converges to $F_*(t)$ as $\log(\lambda_2) \to 0$, so that the asymptotic power of the conventional weighted logrank test is maximized when weights at event times are proportional to the cumulative distributional function of $\left(t^*_{ind}\right)$ evaluated at that time, $F_*(t)$, at those times.

Let $B = dA/d\lambda = \int_0^t \{H_C(t^*_{ind}) - H_C(t)\} e^{(\lambda-1)\{H_C(t^*_{ind})H_C(t)\}} f_*(t^*_{ind}) dt^*_{ind}$, then A $\rightarrow$ F$_*(t)$ and $B \rightarrow \int_0^t \{H_C(t^*_{ind}) - H_C(t)\} f_*(t^*_{ind}) dt^*_{ind} = C$ as $\lambda \rightarrow 1$. It follows that

$$g(t) = \frac{\log\{\lambda(t)\}}{\log\lambda_2} = \frac{\log\{\lambda(t)\}}{-\log\lambda} = \frac{\log\{1 - F_*(t) + A\} - \log\{1 - F_*(t) + \lambda A\}}{-\log\lambda} \quad (8)$$

By taking the derivative of the numerator and denominator with respect to λ, we have

$$\begin{aligned} g(t)\frac{\log\{\lambda(t)\}}{\log\lambda_2} = \frac{\log\{\lambda(t)\}}{-\log\lambda} &\rightarrow -\lambda\left[\frac{B}{1 - F_*(t) + A} - \frac{A + \lambda B}{1 - F_*(t) + \lambda A}\right] \\ &\rightarrow -\left[\frac{C}{1 - F_*(t) + F_*(t)} - \frac{F_*(t) + C}{1 - F_*(t) + F_*(t)}\right] \\ &= F_*(t) \end{aligned} \quad (9)$$

as $\log(\lambda) \rightarrow 0$. *End of the proof of* Theorem 2.2.

The above proof determines the asymptotically optimal weight assignments in the generalized piecewise weighted logrank test. Further, the corresponding asymptotically optimal power can be derived as follows. Given the asymptotic distribution of the weighted logrank statistics under the alternative, the power function can be derived as

$$P_{ow} = \Phi\left(\mu - Z_{1-\frac{\alpha}{2}}\right) + \Phi\left(-\mu - Z_{1-\frac{\alpha}{2}}\right) \quad (10)$$

where μ is the mean of the asymptotic distribution of the weighted logrank statistics under the alternative. Following Schoenfeld (1981) [16], under the alternative, μ converges to

$$\frac{n^{1/2}\int w(t)\log\{\lambda(t)\}\pi(t)(1-\pi(t))V(t)dt}{\left\{\int w^2(t)\pi(t)(1-\pi(t))V(t)dt\right\}^{1/2}}.$$

where function $V(t)$ denotes the probability of observing an event at time t and $\pi(t)$ the probability of observing the event in the experimental group. By substituting $\log\{\lambda(t)\}$ with $(\log\lambda_2)F_*(t)$ and $w(t)$ with $F_*(t)$, and noticing that $\pi(t) \rightarrow P_C$ and $1 - \pi \rightarrow P_E = 1 - P_C$, when censoring distributions are the same in experimental and control groups, the above equation becomes

$$\log(\lambda_2)\left\{nP_CP_E\int F_*^2(t)V(t)dt\right\}^{1/2} \quad (11)$$

Here $n\int F_*^2(t)V(t)dt$ is the expected weighted number of events accumulated after the lag period, and we denote it by $\bar{d}_{23}$.

Thus, with equal allocation ratio and the censoring distribution being the same across arms, the maximum asymptotic power is

$$P_{ow*} = \Phi\left(\frac{1}{2}\log(\lambda_2)\sqrt{\bar{d}_{23}} - Z_{1-\frac{\alpha}{2}}\right) + \Phi\left(-\frac{1}{2}\log(\lambda_2)\sqrt{\bar{d}_{23}} - Z_{1-\frac{\alpha}{2}}\right) \quad (12)$$

This implies that the power of the oncology study is event-driven and when the delayed effect is present, the study power is driven only by the expected weighted number of events accumulated after the earliest treatment onset.

A.2 Derivation of APPLE+ method when t^*_{ind} follows a uniform distribution on (T_1, T_2):

We have established the power function based on the number of events. Next, we will derive a closed form sample size and power calculation formula under the random lag duration scenario. Suppose for each individual, it takes time t^*_{ind} for the delayed treatment effect to kick in after the treatment is administered, and t^*_{ind} varies randomly from individual to individual following a uniform distribution between T_1 and T_2, t^*_{ind}~Uniform$[T_1, T_2]$, where T_1 and T_2 refer to the shortest and longest delayed duration among all individuals, respectively. Suppose u denotes the time when an individual patient arrives the study from the study onset,

- On the individual level, the total course of study can be divided into two phases:
  1. Delayed stage $[u, u + T_1]$ treatment effect is not yet manifested with hazard ratio $\lambda_1 = 1$;
  2. Post-delayed stage $[u + T_1, u + T_2]$ treatment effect is fully manifested with hazard ratio $\lambda_2 = h_C/h_E \geq 1$, where h_E and h_C denote the hazards for the patients in the experimental and control groups, respectively.
- On the aggregated study level, the total study can be divided into three phases:
  - **I.** $[u, u + T_1]$: Treatment effect is not manifested for any individual; average hazard ratio is equal to 1;
  - **II.** $[u + T_1, u + T_2]$: Treatment effect starts to be manifested for individual patients one after another; at time T_1, the first subject achieves the full effect; as of T_2, all subjects reach the full treatment effect; hazard ratio $\lambda(t)$ monotonically increases from 1 to λ_2 in an approximately linear pattern;

III. $[u + T_2, \tau]$: Treatment effect is fully manifested for all individual patients; average hazard ratio becomes constant λ_2.

Assume that patients arriving the trial follow a Poisson process $N(t)$ with intensity a. Then the total patients enrolled during the enrollment period $[0, A]$ has mean $E[N(t)]$ denoted as N with

$$N = a \cdot A$$

where a can be calculated as follows:

1. Under the equal allocation ratio, the number of expected patients in either group $i (i \in \{E, C\})$ arrived during the time interval $[u, u + du]$ is $E[N(u, u + du)] = \frac{\alpha}{2} du$.

2. Among the $\frac{\alpha}{2} du$ patients in the i^{th} group, we calculate the weighted proportion of those who will experience an event during each of the three phases by the end of study τ. In this article, we consider the case where t^*_{ind} follows a uniform distribution, for which, the asymptotically optimal weight assignments can be expressed as:

$$w(t) = F_*(t) = \begin{cases} 0, & \text{if } t < T_1 \\ \dfrac{t - T_1}{T_2 - T_1}, & \text{if } T_1 \le t \le T_2 \\ 1, & \text{if } t > T_2 \end{cases}$$

Suppose that the event time T follows an exponential distribution with rate h_{i1} before t^*_{ind} and h_{i2} after t^*_{ind}, for group $i \in \{E, C\}$ where $h_{C1} = h_{C2} = h_C$ and let the corresponding $F_{ik}(\cdot)$ and $f_{ik}(\cdot)$, $k \in \{1, 2, 3\}$, respectively denote the cumulative distribution function and probability density function of exponential distribution, then

(a) *Conditional density function on the individual level* $f_i\left(t \middle| t^*_{ind}\right)$: where $i \in \{E, C\}$,

$$f_C\left(t \middle| t^*_{ind}\right) = h_C e^{-h_C t}, \quad \text{if } 0 < t \le \tau - u$$

$$f_E\left(t \middle| t^*_{ind}\right) = \begin{cases} h_C e^{-h_C t}, & \text{if } 0 < t \le t^*_{ind} \\ h_E e^{-h_E t} e^{-\left(h_C - h_E\right) t^*_{ind}}, & \text{if } t^*_{ind} < t \le \tau - u \end{cases}$$

(b) *Marginal density function on the study level* $f_{ik}(t)$: where $i \in \{E, C\}$ and $k \in \{1, 2, 3\}$, marginalized over the random distribution of t^*_{ind},

$$f_C(t) = f_{C1}(t) = f_{C2}(t) = f_{C3}(t) = h_C e^{-h_C t}, \text{ if } 0 < t \le \tau - u$$

$$f_E(t) = \begin{cases} f_{E1}(t) = h_C e^{-h_C t}, & \text{if } 0 < t \le T_1 \\ f_{E2}(t) = A' e^{-h_E t} - B e^{-h_C t} + h_C e^{-h_C t} \dfrac{T_2 - t}{T_2 - T_1}, & \text{if } T_1 < t \le T_2 \\ f_{E3}(t) = C h_E e^{-h_E t}, & \text{if } T_2 \le \tau - u \end{cases}$$

where

$$A' = \frac{h_E e^{-(h_C - h_E)T_1}}{(T_2 - T_1)(h_C - h_E)},$$

$$B = \frac{h_E}{(T_2 - T_1)(h_C - h_E)},$$

$$C = \frac{h_E\left[e^{-(h_C - h_E)T_1} - e^{-(h_C - h_E)T_2}\right]}{(T_2 - T_1)(h_C - h_E)}.$$

(c) The weighted proportion of subjects who will experience an event during each of the three phases $[u, u + T_1], [u + T_1, u + T_2], [u + T_2, \tau]$ are

- *Subset 1*: A fraction $F_{i1}(T_1)$ will die during the delayed phase $[u, u + T_1]$;
- *Subset 2*: A fraction $\int_{T_1}^{T_2} \left(\frac{t - T_1}{T_2 - T_1}\right)^2 \cdot f_{i2}(t)dt$ will die during $[u + T_1, u + T_2]$;
- *Subset 3*: A fraction $\int_{T_2}^{T - u} 1 \cdot f_{i3}(t)dt$ will die during $[u + T_2, \tau]$ where *Subset 2* and *Subset 3* can be used to test treatment effect and therefore contributes to the power of the test.

3. Hence, among the patients who arrive during $[u, u + du]$, the expected weighted number of patients who will die in each phase from either experimental or control group can be obtained by multiplying the respective weighted proportion with $\frac{a}{2}du$. Integrating over u, the total expected value of weighted number of events during $[u, u + T_1], \bar{d}_{i1}$, during $[u + T_1, u + T_2], \bar{d}_{i2}$, and during $[u + T_2, \tau], \bar{d}_{i3}$, can be derived, respectively,

- Control Group:

$$\bar{d}_{C23} = \bar{d}_{C2} + \bar{d}_{C3} = \frac{a}{2}[DA - \frac{e^{-h_C\tau}}{h_C}(e^{h_CA} - 1)]$$

where

$$D = \frac{2[e^{-h_CT_1} - e^{-h_CT_2}]}{h_C^2(T_2 - T_1)^2} - \frac{2e^{-h_CT_2}}{h_C(T_2 - T_1)}$$

- Experimental Group:

$$\bar{d}_{E23} = \bar{d}_{E2} + \bar{d}_{E3} = \frac{a}{2}[EA - \frac{C}{h_E^2}(e^{-h_E(\tau - A)} - e^{-h_E\tau})]$$

where

$$E = e^{-h_CT_1}\left\{\frac{2}{h_E^2(h_C - h_E)(T_2 - T_1)^3} - \frac{2h_E}{h_C^3(h_C - h_E)(T_2 - T_1)^3} + \frac{2}{h_C^2(T_2 - T_1)^2}[1 - \frac{3}{h_C(T_2 - T_1)}]\right\} + e^{-h_CT_2}\left\{-\frac{1}{(T_2 - T_1)(h_C - h_E)} + \frac{h_E}{h_C(h_C - h_E)(T_2 - T_1)}[1 + \frac{2}{h_C(T_2 - T_1)} + \frac{2}{h_C^2(T_2 - T_1)^2}] + \frac{1}{h_C(T_2 - T_1)}[1 + \frac{4}{h_C(T_2 - T_1)} + \frac{6}{h_C^2(T_2 - T_1)^2}]\right\} + e^{-h_CT_1}e^{-h_E(T_2 - T_1)}\left\{-\frac{2}{h_E(h_C - h_E)(T_2 - T_1)^2} - \frac{2}{h_E^2(h_C - h_E)(T_2 - T_1)^3}\right\}$$

- That is, the total expected weighted number of patients who would experience events after the delayed onset T_1 is

$$\bar{d}_{23} = \bar{d}_{C23} + \bar{d}_{E23} = \frac{N}{2}(E + D) - \frac{NC}{2Ah_E^2}[e^{-h_E(\tau - A)} - e^{-h_E\tau}] - \frac{Ne^{-h_C\tau}}{2Ah_C}(e^{h_CA)} - 1) \quad (13)$$

Thus, the asymptotically maximum power of the study, Pow^*, given the sample size N is obtained indirectly through the relationship between Pow^* and $\bar{d}_{23}$ (formula (12)), and that between $\bar{d}_{23}$ and N (formula (13)).

A.3 Proof of Lemma 5.1:

To show Lemma 5.1, we first show the following Theorem 6.1.

Theorem 6.1 proof:

Proof: In Theorem 2.2, we have shown that $\log\{\lambda^{(i)}(t)\}$ converges to $\log(\lambda_2)F_*^{(i)}(t)$ as $\log(\lambda_2) \to 0$, where $i \in \{1, 2\}$. Under the condition $\log(\lambda_2) = O(n^{-1/2})$, the weighted logrank test statistics follows the standard normal distribution with means

$$\mu^{(i)} = \frac{n^{1/2}\int w(t)\log(\lambda_2)F_*^{(i)}(t)\pi^{(i)}(t)\left\{1-\pi^{(i)}\right\}V^{(i)}(t)dt}{\left[\int \{w(t)\}^2\pi^{(i)}(t)\left\{1-\pi^{(i)}(t)\right\}V^{(i)}(t)dt\right]^{1/2}}$$

As n is large and $\log(\lambda_2)$ is close to zero, $\pi^{(1)}(t) \approx \pi^{(2)}(t)$ and $V^{(1)}(t) \approx V^{(2)}(t)$. Since $F_*^{(1)}(t) \geq F_*^{(2)}(t)$ and $w(t)$'s are positive, $|\mu^{(1)}|$ is asymptotically larger than or equal to $|\mu^{(2)}|$. As indicated in formula (10), the power becomes bigger as μ gets larger. This implies that the weighted logrank test will provide larger power to detect the effect of treatment 1 than to detect the effect of treatment 2.

Theorem 6.1 can lead to Lemma 5.1.

Lemma 5.1 proof:

Proof: Let $F_*(t)$ denote the true lag time cumulative distribution function (CDF) based on the true domain $[T_1, T_2]$ and $F_*^{(m)}(t)$ the assumed CDF based on the mis-specified domain $[t_1^m, t_2^m]$, then the true effect corresponds to $F_*(t)$ and the assumed effect refers to $F_*^{(m)}(t)$. The true and assumed marginal hazard ratio functions would differ due to the misspecified bound(s) of lag pattern. Under the random lag duration scenario where t_{ind}^* follows a uniform distribution with either one or both of bounds being over-(under-)estimated, we can easily see that $F_*(t) \geq F_*^{(m)}(t)$ for all t. It follows from Theorem 6.1 that the generalized piecewise weighted logrank test with the weights $w(t)\alpha F_*^{(m)}(t)$ have at least (at most) as much asymptotic power to detect the true effect than the assumed effect given the same sample size. Since the sample size estimated to achieve the target power is calculated using APPLE + or SEPPLE+ based on the mis-specified lag time distribution $F_*^{(m)}(t)$ on domain $[t_1^m, t_2^m]$, it follows that the true power of the generalized piecewise weighted logrank test given the estimated sample size is no less (more) than the target power.

References

1. Xu Z, Zhen B, Park Y, Zhu B. Designing therapeutic cancer vaccine trials with delayed treatment effect. Statistics in Medicine 2017; 36(4):592–605. [PubMed: 27807870]
2. Hoos A, Parmiani G, Hege K, Sznol M, Loibner H, Eggermont A, Urba W, Blumenstein B, Sacks N, Keilholz U, et al. A clinical development paradigm for cancer vaccines and related biologics. Journal of Immunotherapy 2007; 30(1):1–15. [PubMed: 17198079]

3. Sliwkowski MX, Mellman I. Antibody therapeutics in cancer. Science 2013; 341(6151):1192–1198. [PubMed: 24031011]
4. Melero I, Gaudernack G, Gerritsen W, Huber C, Parmiani G, Scholl S, Thatcher N, Wagstaff J, Zielinski C, Faulkner I, et al. Therapeutic vaccines for cancer: an overview of clinical trials. Nature Reviews Clinical oncology 2014; 11(9):509–524.
5. Kantoff PW, Higano CS, Shore ND, Berger ER, Small EJ, Penson DF, Redfern CH, Ferrari AC, Dreicer R, Sims RB, et al. Sipuleucel-t immunotherapy for castration-resistant prostate cancer. New England Journal of Medicine 2010; 363(5):411–422. [PubMed: 20818862]
6. Hodi FS, O'day SJ, McDermott DF, Weber RW, Sosman JA, Haanen JB, Gonzalez R, Robert C, Schadendorf D, Hassel JC, et al. Improved survival with ipilimumab in patients with metastatic melanoma. New England Journal of Medicine 2010; 2010(363):711–723.
7. Robert C, Long GV, Brady B, Dutriaux C, Maio M, Mortier L, Hassel JC, Rutkowski P, McNeil C, Kalinka-Warzocha E, et al. Nivolumab in previously untreated melanoma without braf mutation. New England journal of medicine 2015; 372(4):320–330. [PubMed: 25399552]
8. Self S, Prentice R, Iverson D, Henderson M, Thompson D, Byar D, Insull W, Gorbach SL, Clifford C, Goldman S, et al. Statistical design of the women's health trial. Controlled Clinical Trials 1988; 9(2):119–136. [PubMed: 3396363]
9. Lakatos E Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics 1988; 44:229–241. [PubMed: 3358991]
10. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. Biometrika 1990; 77(4):853–864.
11. Shih JH. Sample size calculation for complex clinical trials with survival endpoints. Controlled Clinical Trials 1995; 16(6):395–407. [PubMed: 8720017]
12. Fine GD. Consequences of delayed treatment effects on analysis of time-to-event endpoints. Drug Information Journal 2007; 41(4):535–539.
13. Zhang D, Quan H. Power and sample size calculation for log-rank test with a time lag in treatment effect. Statistics in Medicine 2009; 28(5):864–879. [PubMed: 19152230]
14. Hasegawa T Sample size determination for the weighted log-rank test with the fleming–harrington class of weights in cancer vaccine studies. Pharmaceutical statistics 2014; 13(2):128–135. [PubMed: 24497461]
15. He P, Su Z. A novel design for randomized immuno-oncology clinical trials with potentially delayed treatment effects. Contemporary Clinical Trials Communications 2015; 1:28–31. [PubMed: 29736436]
16. Schoenfeld D The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981; 68(1):316–319.
17. Gail MH. Sample size estimation when time-to-event is the primary endpoint. Drug Information Journal 1994; 28(3):865–877.

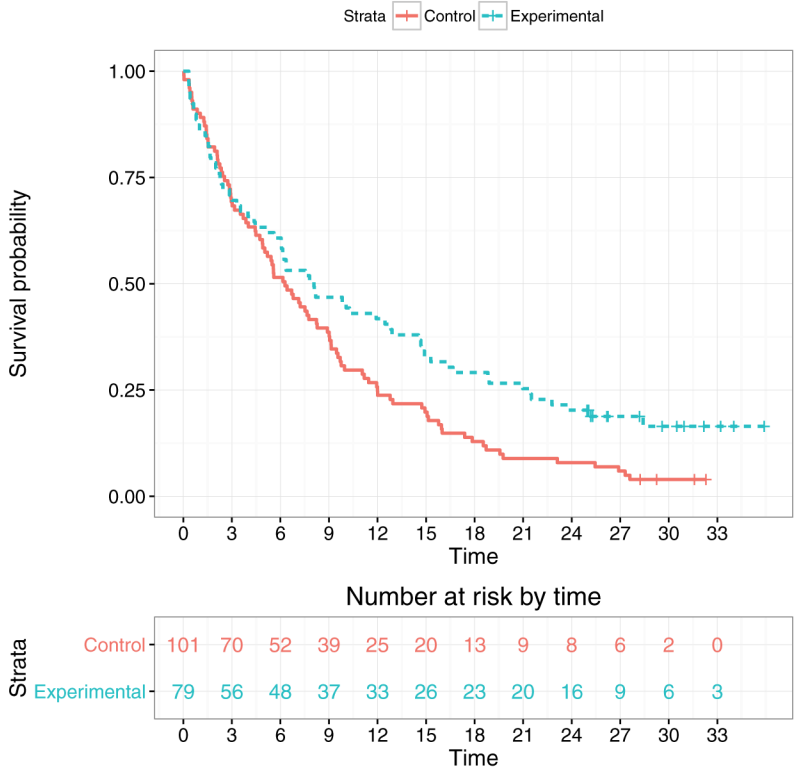

Figure 1.

The Kaplan-Meier plot of a synthetic example created based on a confidential real data source illustrates the delayed treatment effect scenario with random time lag. Each individual achieves the treatment effect at a subject-specific time t^*_{ind}, where we assume t^*_{ind}~ Uniform [3, 12] months.

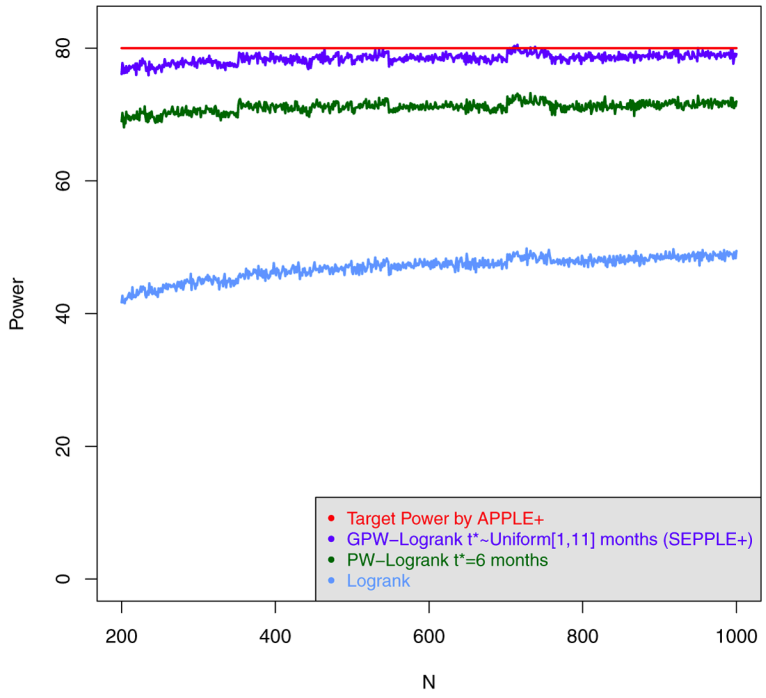

Figure 2.

Powers of APPLE+, SEPPLE+ based on GPW-Logrank test (GPW-Logrank) with $t^*_{ind} \sim [1, 11]$ months, PW-Logrank test (PW-Logrank) with $t^* = 6$ months and regular logrank test (Logrank) ignoring the delayed treatment effect, where the power of APPLE+ is set at the target 80%, under the random lag duration scenario with $T_1 = 1$, $T_2 = 11$ months; baseline hazard $h_C = 0.0067$.

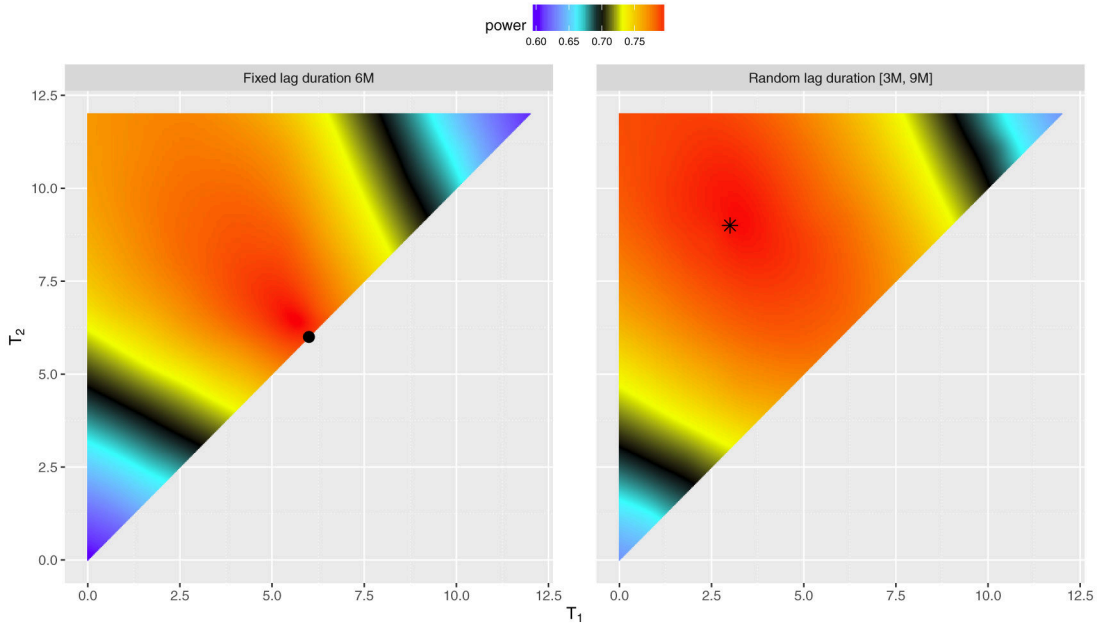

Figure 3.

Powers of GPW-Logrank test and PW-Logrank test when the domain of lag time varies, under the fixed lag duration scenario with t^* = 6 months and random lag duration scenario with t^*_{ind}~ Uniform[3, 9] months. Sample size N = 591 under the fixed scenario and N = 630 under the random scenario; these sample sizes can provide 80% power if the lag model is correctly specified under the specific lag duration scenario. The black dot refers to the scenario where the maximum power is achieved when the lag domain is correctly specified at the truth $T = T_2 = t^*$ under the fixed lag duration scenario; the black star indicates the scenario where the maximum power is achieved when the lag domain is correctly specified at the truth $T_1 = 3$ and $T_2 = 9$ under the random lag duration scenario. $\lambda_2 = 0.72$; h_C = 0.002; $\alpha = 0.05$.

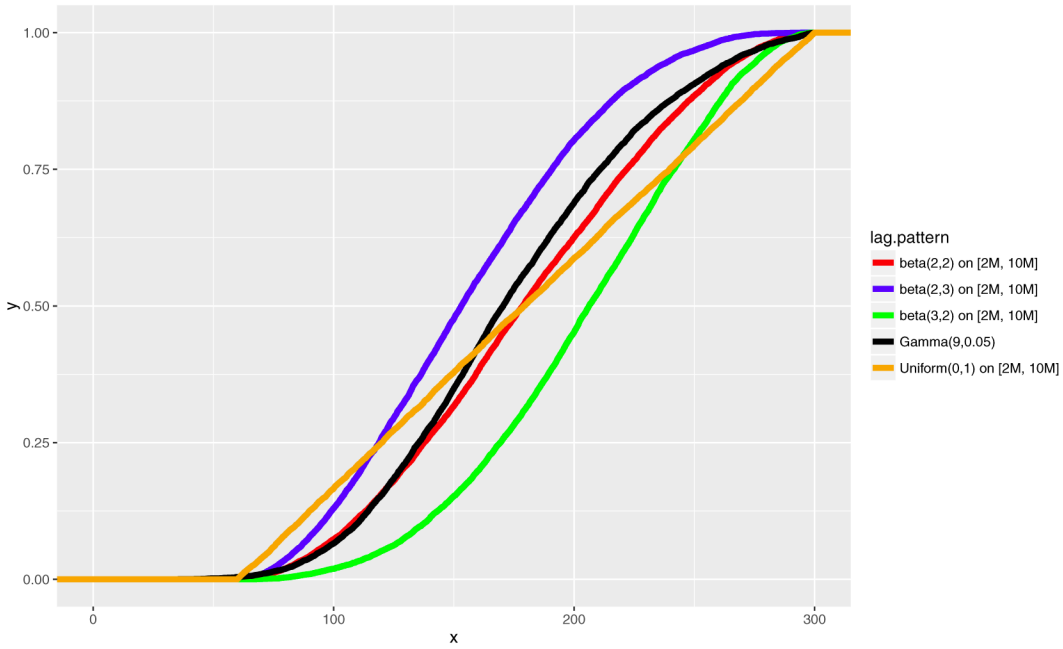

Figure 4.
CDFs of various lag distributions investigated.

Table 1.

The empirical power comparison of Generalized Piecewise Weighted Logrank test (GPW-Logrank) and Piecewise Weighted Logrank test (PW-Logrank) when the lag duration scenario or lag domain or both are mis-specified. The fixed lag duration scenario sets $t^* = 6$ months and random lag duration scenario specifies $t^*_{ind} \sim$ Uniform$[T_1 = 3, T_2 = 9]$ months. Sample size $N = 591$ under the fixed scenario and $N = 630$ under the random scenario; these sample sizes can provide 80% power if the lag model is correctly specified under the specific delayed scenario. $\lambda_2 = 0.72$; $h_C = 0.002$; $\alpha = 0.05$.

Tests	**Power**	
	True parameter setting	
Mis-specified parameter setting	**Fixed scenario $t^* = 6$**	**Random scenario $[T_1, T_2] = [3, 9]$**
PW-Logrank with $t^m = 1$ months	63%	66%
PW-Logrank with $t^m = 3$ months	68%	72%
PW-Logrank with $t^m = 6$ months	79%	78%
PW-Logrank with $t^m = 9$ months	70%	75%
PW-Logrank with $t^m = 11$ months	64%	69%
GPW-Logrank with $[T_1^m, T_2^m] = [1,11]$ months	76%	79%
GPW-Logrank with $[T_1^m, T_2^m] = [1,9]$ months	76%	79%
GPW-Logrank with $[T_1^m, T_2^m] = [3,11]$ months	78%	79%
GPW-Logrank with $[T_1^m, T_2^m] = [3,9]$ months	78%	80%

Table 2.

The impact of mis-specifying the lag time distribution on the design using APPLE+ and analysis using GPWLogrank test (GPW-Logrank) under random lag duration scenario, where the target power is set at 80%. For comparison purpose, we also evaluate the impact of mis-specifying the constant lag time on APPLE & PW-Logrank test (PW-Logrank) under the random lag duration scenario. The true domain is [$T_1 = 2,\ T_2 = 10$] months and is correctly specified in APPLE+ & GPW-Logrank test. The middle of true domain is specified in APPLE & PW-Logrank test. Baseline hazard $h_C = 0.0028$.Total accrual duration $A = 1$ year. Total study duration $\tau = 3$ years. Type I error rate $\alpha = 0.05$.

λ_2	Design & Analysis Methods	Lag Domain (months)	Sample Size	Empirical Power(%)
S1: t^*_{ind} ~ Beta(2, 3) on [2*M*, 10*M*]				
0.56	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	219 201	83% 78%
0.60	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	278 255	82% 78%
0.64	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	360 329	83% 77%
0.68	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	477 435	83% 78%
0.72	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	651 592	83% 78%
S2: t^*_{ind} ~ Beta(3, 2) on [2*M*, 10*M*]				
0.56	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	219 201	75% 78%
0.60	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	278 255	75% 70%
0.64	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	360 329	77% 70%
0.68	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	477 435	76% 71%
0.72	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	651 592	76% 71%
S3: t^*_{ind} ~ Beta(2, 2) on [2*M*, 10*M*]				
0.56	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	219 201	79% 74%
0.60	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	278 255	79% 75%
0.64	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	360 329	81% 74%
0.68	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	477 435	80% 75%
0.72	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1 = 2, T_2 = 10$ $t^m = 6$	651 592	79% 75%
S4: $t^*_{ind} \sim \Gamma$ (9,0.05)				

λ_2	Design & Analysis Methods	Lag Domain (months)	Sample Size	Empirical Power(%)
0.56	APPLE+ & GPW-Logrank	T_1 =2 , T_2 = 10	219	80%
	APPLE & PW-Logrank	t^m = 6	201	74%
0.60	APPLE+ & GPW-Logrank	T_1 =2 , T_2 = 10	278	79%
	APPLE & PW-Logrank	t^m = 6	255	74%
0.64	APPLE+ & GPW-Logrank	T_1 =2 , T_2 = 10	360	80%
	APPLE & PW-Logrank	t^m = 6	329	74%
0.68	APPLE+ & GPW-Logrank	T_1 =2 , T_2 = 10	477	80%
	APPLE & PW-Logrank	t^m = 6	435	74%
0.72	APPLE+ & GPW-Logrank	T_1 =2 , T_2 = 10	651	80%
	APPLE & PW-Logrank	t^m = 6	592	73%

Table 3.

The salvage of power loss due to the lag distribution mis-specification by over-specifying the boundary(ies) of lag domain in both design using APPLE+ and analysis using GPW-Logrank test (GPW-Logrank) under random lag duration scenario, where the target power is set at 80%. For comparison purpose, we also evaluate the impact on APPLE & PWLogrank test (PW-Logrank). The true domain [$T_1 = 2,\ T_2 = 10$] months. The mis-specified domain $T_1^m = 2, t_2^m = 12$ months. Baseline hazard $h_C = 0.0028$. Total accrual duration $A = 1$ year. Total study duration $\tau = 3$ years. Type I error rate $\alpha = 0.05$.

λ_2	Design & Analysis Methods	Lag Domain (months)	Sample Size	Empirical Power(%)
		S2: t^*_{ind} ~ Beta(3, 2) on [2*M*, 10*M*]		
0.60	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1^m = 2, T_2^m = 12$ $t^m = 7$	314 277	81% 74%
0.64	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1^m = 2, T_2^m = 12$ $t^m = 7$	406 357	81% 75%
0.68	APPLE+ & GPW-Logrank APPLE & PW-Logrank	$T_1^m = 2, T_2^m = 12$ $t^m = 7$	539 471	81% 75%

Table 4.

The impact of mis-specifying the lag domain in GPW-Logrank (GPW-Logrank) test versus PW-Logrank (PWLogrank) test on the study design and analysis when the random delayed effect is present. The target power is set at 80%. $t^*_{ind} \sim$ Uniform$[T_1 = 2,\ T_2 = 10]$ months. Hazard ratio $\lambda_2 = 0.64$. Baseline hazard $h_C =$ 0.002. Total accrual duration $A = 1$ year. total study duration $\tau = 3$ years. Type I error rate $\alpha = 0.05$.

Design & Analysis Methods	Mis-specified Lag Domain (months)	Sample Size	Empirical Power(%)
APPLE+ & GPW-Logrank APPLE & PW-Logrank	T_1^m =1, T_2^m = 11 t^m = 6	366 329	80% 72%
APPLE+ & GPW-Logrank APPLE & PW-Logrank	T_1^m =1, T_2^m = 10 t^m = 5.5	345 316	76% 70%
APPLE+ & GPW-Logrank APPLE & PW-Logrank	T_1^m =2, T_2^m = 11 t^m = 6.5	382 342	81% 75%
APPLE+ & GPW-Logrank APPLE & PW-Logrank	T_1^m =3, T_2^m = 9 t^m = 6	353 329	78% 72%