Published in final edited form as:

Stat Med. 2020 November 30; 39(27): 3914–3936. doi:10.1002/sim.8694.

Design for immuno-oncology clinical trials enrolling both responders and nonresponders

Zhenzhen Xu¹, Bin Zhu^{#2}, Yongsoek Park^{#3}

¹Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, Maryland ²Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland ³Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania

These authors contributed equally to this work.

Abstract

A typical challenge facing the design and analysis of immuno-oncology (IO) trials is the prevalence of nonproportional hazards (NPH) patterns manifested in Kaplan-Meier curves under time-to-event endpoints. The NPH patterns would violate the proportional hazards assumption, and yet conventional design and analysis strategies often ignore such a violation, resulting in underpowered or even falsely negative IO studies. In this article, we show, both empirically and analytically, that treating nonresponders in IO studies of inadequate size would give rise to a variety of NPH patterns; we then present a novel design and analysis strategy, ***P*%-responder information embedded (PRIME)**, to properly incorporate the dichotomized response incurred from treating nonresponders. Empirical studies demonstrate that the proposed strategy can achieve desirable power, whereas the conventional alternative leads to a severe power loss. The PRIME strategy allows us to quantify the impact of treating nonresponders on study efficiency, thereby enabling a proper design of IO trials with an adequate power. More importantly, it pinpoints a solution to enhance the study efficiency and alleviates the NPH patterns by enrolling more prospective responders. An R package (Immunotherapy.Design) is developed for implementation.

Keywords

cancer immunotherapy; clinical trial; dichotomized response; immuno-oncology trial; nonproportional hazards pattern; proportional hazards assumption; sample size and power calculation

1 | INTRODUCTION

While cancer immunotherapy revolutionizes cancer treatment, its unique characteristics pose new challenges in the design and analysis of immuno-oncology (IO) trials. Particularly, a variety of complex survival patterns, known as nonproportional hazards (NPH) patterns, frequently arise in the Kaplan-Meier (KM) curves under time-to-event endpoints. Such

patterns can be categorized into four distinct categories, including the *delayed effect pattern* where the KM curves for the two treatment groups do not separate until a lag time (Figure 1A), the *belly shape diminishing effect pattern* where the two curves separate for a period of time before merging to each other (Figure 1B), the *crossing hazards pattern* where the KM curve for one group starts out being worse but ends up being better than that for the other group (Figure 1C), and the *combination patterns* that mingle the aforementioned patterns in various fashions (Figure 1D–F). The NPH patterns violate the proportional hazards (PH) assumption required by conventional approaches based on the regular log-rank test and give rise to underpowered or even falsely negative studies. Hence, the question of interest becomes how to design adequate and well-controlled IO trials when the PH assumption no longer holds and, most importantly, to improve the study efficiency by mitigating the occurrence of NPH patterns via study design.

There is considerable literature proposing novel designs for IO trials with NPH patterns.^{1–7} However, each proposal has limitations. Xu et al,^{1,2} Sit,³ and Pepe⁶ address a particular category of NPH patterns but their proposals cannot be generalized to other categories. Royston^{4,5} and Uno⁷ propose the use of restricted mean survival time to account for the NPH patterns of various kinds; however, the choice of a clinically meaningful time window is challenging and the resultant treatment effect estimate is sensitive to such a choice. In addition, none of such approaches to date provides a solution to alleviate the NPH patterns at the design stage; instead, they attempt to address the problem after it has occurred with statistical analysis techniques.

To fundamentally address the question of interest requires a more comprehensive understanding of the causes behind various NPH patterns. Biological and clinical evidence^{1,2,8–13} show that the indirect mechanism of action of immunotherapy results in delayed treatment response, which serves as a major cause of delayed effect pattern. We acknowledge such a cause, but still come to question whether the lag duration observed in the KM curves reflects the underlying lag time of delayed treatment response? For instance, some IO studies display a lag duration of 6 months or even longer, would the immunotherapy truly take as long to take effect? Next, what are the causes of other patterns? The prevalence of various patterns evinces that there may be more than a working mechanism causing such an issue.

We hypothesize that treating nonresponders in studies of inadequate size would be another major cause of NPH patterns. Unlike the oncology trials of conventional cancer treatments, IO trials are characterized by treating nonresponders. Only a limited percentage of enrolled patients who harbor immunotherapy-sensitive tumors would respond to the immunotherapy and achieve improved overall survival (OS) due to reduced hazards once being treated (responders),¹⁴ whereas the rest would not respond despite being treated (nonresponders).^{15–19} Such a phenomenon signifies a dichotomized treatment group consisting of two subsets of patients: responders and nonresponders. Conventional studies, however, ignore the dichotomized response incurred by treating nonresponders in the design and analysis strategy. When it comes to calculating the sample size, conventional strategies assume that all treated patients would uniformly respond to the experimental therapy and specify a homogeneous treatment effect size in two manners: first, overspecify the effect size by

targeting only the responders, unaware that the beneficial effect in responders has been diluted by the null effect in large number of nonresponders; and second, adapt to an average, diluted hazard ratio (HR) of all treated subjects against control. Overspecifying the effect size would inevitably result in inadequate sample sizes and underpowered, falsely negative studies and adopting a diluted HR lead to unnecessarily inflated sample sizes and a severe loss of efficiency.

Xu et al²⁰ empirically explores the above hypothesis and demonstrates that when treating nonresponders in studies of inadequate size, various NPH patterns frequently arise regardless of the presence of delayed treatment effect. To salvage the power loss due to the incurred dichotomized response, they propose the parametric P-Embedded strategy, where the distribution of cumulative baseline hazard, $\Lambda_0(t)$, is modeled parametrically. Although this strategy addresses the hypothesized cause, the validity of the hypothesis is tested only empirically without a very refined theory upon which to support the causality.²⁰ Moreover, the parametric framework is sensitive to the misspecification of parametric assumptions on $\Lambda_0(t)$. These limitations motivate our interests and lead to this extension.

In this article, we first provide a theoretical justification of empirical observations that dichotomized treatment response plus inadequate sample size is sufficient to cause NPH patterns. To salvage the resultant power loss and improve the robustness performance, we next extend the parametric P-Embedded strategy to a nonparametric **P%-responder information embedded (PRIME)** approach where $\Lambda_0(t)$ is left nonparametric and modeled with the Breslow estimator.²³ The proposed nonparametric PRIME design and analysis strategy includes

1. *The PRIME EM algorithm* for treatment effect estimation
2. *The PRIME likelihood ratio test* for hypothesis testing
3. *The PRIME sample size and power calculation* for study design
4. *The PRIME profile likelihood-based approach* for prespecification of p using prior studies.

As its name implies, the proposed strategy embeds the dichotomized response in the design and analysis procedure, where the response dichotomy is measured by the baseline proportion of responders among the treatment group, p . This idea of embedding p is motivated by the practical challenges of conducting IO trials that either enroll all-comers or enrich subpopulations defined by nondefinitive biomarkers. Specifically, when IO trials enroll all-comers, the responders are not differentiated from the nonresponders and all treated subjects are assumed to uniformly respond to the experimental agent. When IO trials utilize nondefinitive biomarkers to aid in patient selection, the prospective responding patients are identified only probabilistically rather than definitively, due to the lack of definitive predictive biomarkers. In both cases, a dichotomized treatment response would inevitably arise and become difficult to eliminate. We thus propose measuring such a dichotomy by the aggregated proportion of responders among the treated, p , and taking advantage of the information on p in the design and analysis strategy to adequately power the study, where p can often be properly specified in advance via the proposed PRIME

strategy using prior studies. Statistically speaking, the individual patient's responder membership is a latent, binary random variable with the chance of success prespecified as p . The PRIME strategy hence, in essence, involves a mixture model where the treatment group is a mixture of responders and nonresponders and the individual responder membership is unknown. That differentiates the proposed mixture model from the existing mixture models including the cure model²⁴ for which all treated patients respond with a subset of them being cured.

Overall, PRIME has four attractive advantages. First, it incorporates two well-recognized features of immunotherapy into the design and analysis procedures of IO trials: the delayed treatment effect and dichotomized treatment response. Second, it allows us to quantify the impact of treating nonresponders on study efficiency; as such, it ensures an adequate study power to detect a potentially effective therapy provided that p is correctly prespecified. In the case of misspecification, the proposed approach can still achieve a considerable gain in study efficiency as opposed to its conventional alternative which ignores the dichotomized response by assuming 100% at 100%. Third, it leads to an estimate of treatment effect specific for responders by incorporating the latent responder membership into the estimation process via the expectation maximization (EM) algorithm, whereas the conventional approach estimates an averaged effect for both responders and nonresponders in the treatment group. Finally, it outlines a path toward precision immunotherapy to enhance the IO trial efficiency and mitigate the occurrence of NPH patterns. Although issues concerning the dichotomized treatment response in IO trials have been well recognized in medical literature evaluating the efficacy of immunotherapy,^{25–27} they have not been fully studied in the context of trial design and data analysis. The present article aims to address such issues from this new angle. An R package (Immunotherapy.Design) is developed to facilitate the implementation of the proposed strategy in both parametric and nonparametric fashion, which is freely available at GitHub <https://github.com/binzhulab/Immunotherapy.Design>.

2 | IMPACT OF RESPONSE DICHOTOMY AND SAMPLE SIZE ON COMPLEX SURVIVAL PATTERNS

To explore our hypothesis that the dichotomized treatment response plus inadequate sample size serves as a possible cause driving the NPH patterns, Xu et al²⁰ conducts a simulation study, in which the underlying responder membership is simulated for each individual subject and based on that, the survival outcomes for the responders and nonresponders are, respectively, drawn. The results are depicted in KM plots. Under an equal allocation ratio, enrollment rate of 1.6 subjects per day, study duration of 5 years and 100% of 20%, among 100 hypothetical trials of 200 subjects each, an astounding 89% of them manifest NPH patterns of various kinds detected by visual inspection. Three typical examples simulated under this scenario are illustrated in the upper panel of Figure 2. By contrast, once 100% increases to 90% and the sample size N remains as 200, only 20% of simulated trials demonstrate diverse NPH patterns whereas the rest displays a more clear separation of survival curves between the treatment and control arms right after the 3-months lag, followed by a seemingly PH pattern during the postdelayed phase, due to a built-in treatment time-lag effect (middle panel of Figure 2). Finally, when 100% remains as low as 20% and

N is inflated to 10-fold larger, seemingly PH patterns also emerge except that the proportionality constant appears to be much smaller than that observed under a 100% of 90% (bottom panel of Figure 2).

In this article, we show that the causality between p, N and the NPH patterns observed empirically is guaranteed by Theorems 1 and 2, justifications of which are given in Appendix A. Specifically, these two theorems illustrate the impact of p and N on the NPH patterns via the relationship between two HRs, $h(t)$ and h , where $h(t)$ denotes the population HR function contrasting a responders-nonresponders-mixed treatment arm with a control arm, h represents the constant HR contrasting responders with nonresponders or controls, t^* indicates the treatment time-lag duration and $S_T^*(t)$, $S_C^*(t)$ refer to the residual survival functions for the treated and control subjects, respectively.

Theorem 1.

The population HR function $h(t) = h + (1 - p)(1 - h)S_C^(t)/S_T^*(t)$ varies with time within the range $h - (1 - p) \leq h(t) \leq 1 - p + ph$ where $t^* \leq t < +\infty$. When $100p\% \rightarrow 100\%$, $h(t) \rightarrow h$.*

Theorem 2.

Under the local asymptotic condition $\log\{h(t)\} = O(N^{-1/2})$ as $N \rightarrow \infty$, the log HR $\log\{h(t)\}$ is asymptotically proportional to $\log(h)$ and the proportionality $\log\{h(t)\}/\log(h) \rightarrow p$ as $N \rightarrow \infty$.

Theorem 1 provides intuitive, clinically meaningful implications with respect to the collective impact of p and N on the NPH patterns. As shown by Equation (6) in Appendix A, $h(t) = h + (1 - p)(1 - h)S_C^*(t)/S_T^*(t)$, the population HR $h(t)$ contrasting treatment with control depends on the baseline proportion of responders among the treated p , the constant HR between responders and nonresponders h , as well as a ratio involving the residual survival functions for treatment ($S_T^*(t)$) and control ($S_C^*(t)$), $S_C^*(t)/S_T^*(t)$. It is obvious that the time-variant $h(t)$ is caused by the time-varying ratio $S_C^*(t)/S_T^*(t)$, unless either 100% is 100%, or h equals to 1, or $S_C^*(t)/S_T^*(t)$ stays constant over time. The last two conditions would hold under the null treatment effect, and in that case, $h(t)$ remains as the constant 1. The first condition 100% = 100% would hold when the treatment group contains only responders and the contrast between treatment and control thus becomes the difference purely between responders (treated) and nonresponders (controls); in that event, $h(t)$ stays at the constant h . Therefore, as long as the nonresponders are included in the treatment group (ie, 100% < 100%), the NPH patterns are guaranteed to arise for any effective immunotherapy with a nonnull effect (ie, $h \neq 1$).

Although p has a dominating effect on NPH patterns, increasing the sample size N can diminish such an effect. Theorem 2 reveals a significant impact of N on the NPH patterns when 100% < 100%. Under the locally asymptotic condition, a sufficiently large N leads the time-dependent HR function $h(t)$ to a constant h^p where the PH assumption also holds but the proportionality constant is reduced from h to h^p with 100% < 100%. Naturally, as 100% rises to 100%, this proportionality constant elevates to its maximum h .

Hence, the analytic and empirical evaluations both support our hypothesis that dichotomized response plus inadequate sample size serves as one of the primary driving causes behind the NPH patterns. That motivates us to develop the PRIME strategy to properly incorporate the dichotomized response into the design and analysis of IO trials, with the aim to salvage the resultant power loss and enhance the trial efficiency.

3 | METHODS

Consider an IO trial with a time-to-event endpoint that involves N subjects randomized to the treatment group E and control group C with probabilities P_E and P_C , where $P_E + P_C = 1$. Let T_i and C_i denote the failure time of interest and censoring time since randomization for subject i , respectively, where $i = 1, 2, \dots, N$. Suppose the total study duration is τ ; let e_i refer to the calendar time when subject i arrives at the study since the study onset. We assume no dropouts except the administrative censoring mechanism. The corresponding censoring indicator for subject i is denoted by $\delta_i = I(T_i \leq C_i)$ and the observational time by $X_i = \min\{T_i, C_i\}$. Also let β represent the log HR between responders and nonresponders measuring the treatment effect and $\lambda_0(t)$, $t > 0$, denote an arbitrary baseline hazard function. We assume that the experimental treatment effect takes a lag time t^* to be manifested in the responding subjects due to the indirect mechanism of action of immunotherapy and such a lag duration is homogeneous across responders,^{28,29} whereas the treatment has no effect in the nonresponding treated subjects and they would behave just as the controls. Thus, the hazard functions for the responders, $\lambda_1(t)$, and for the nonresponders or controls, $\lambda_2(t)$, are

$$\lambda_1(t) = \lambda_0(t)e^{\beta I(t \geq t^*)},$$

$$\lambda_2(t) = \lambda_0(t).$$

Further let Z_i denote the responder membership for subject i , then Z_i is unobserved for the treated subjects but observed for the controls. Specifically, Z_i for a given treated subject i ($i \in E$) follows a Bernoulli distribution with a probability p , leaving the probability mass function for Z as $f(Z_i) = p^{Z_i I(i \in E)} (1 - p)^{(1 - Z_i) I(i \in E)}$. For control subjects, $Z = 0$. The complete-data full likelihood function can be written as

$$\begin{aligned} L(\beta, \lambda_0(t), Z; \mathbf{X}, \Delta) &= \prod_{i=1}^N \lambda_0^{\delta_i}(X_i) e^{\beta I(X_i \geq t^*) \delta_i Z_i} e^{-\beta [\Lambda_0(X_i) - \Lambda_0(t^*)] I(X_i \geq t^*) Z_i} e^{-\Lambda_0(X_i) [1 - I(X_i \geq t^*) Z_i]} \\ &\quad \times e^{-\Lambda_0(t^*) I(X_i \geq t^*) Z_i} f(Z_i) \\ &= \prod_{i=1}^N f_1^{Z_i}(y_i; \beta, \Lambda_0) f_2^{1 - Z_i}(y_i; \beta, \Lambda_0) f(Z_i) \end{aligned} \quad (1)$$

where $y_i = (X_i, \delta_i)$ and the probability density functions for responders, $f_1(\cdot)$, and nonresponders or controls, $f_2(\cdot)$, are being implicitly defined.

To estimate the cumulative baseline hazard $\Lambda_0(X)$ nonparametrically, we adopt the Breslow estimator.²³ Breslow proposes a nonparametric estimator for the cumulative baseline hazard

function by treating $\lambda_0(X)$ as piecewise constants between distinct, uncensored failure times. Specifically, for a sample with M events, we rank all failure times to M distinct times, $0 \equiv T_{(0)} < T_{(1)} < \dots < T_{(M)}$ with the corresponding multiplicities of $m_l \equiv 0, 1, \dots, M$, then assume a constant hazard λ_l between failure times $T_{(l-1)}$ and $T_{(l)}$, where $l = 1, \dots, M$. Note that $m_1 = \dots = m_M = 1$ corresponds to the case of no ties but in this article, we consider the general case accommodating the tied failure times. Based on the reparameterization of Breslow's estimator, the full likelihood function in Equation (1) thus becomes the function of β and $\lambda = (\lambda_1, \dots, \lambda_M)$,

$$\begin{aligned} L(\beta, \lambda, \mathbf{Z}; \mathbf{X}, \Delta) &= \left(\prod_{i=1}^N \prod_{l=1}^M \lambda_l^{\Delta_{li}} \right) \prod_{i=1}^N e^{\beta I(X_i \geq t^*) \delta_i Z_i} e^{-e^{\beta} \left[\sum_{l=1}^M \lambda_l (m_{li} - m_{l*}) \right] I(X_i \geq t^*) Z_i} \\ &\quad \times e^{-\left[\sum_{l=1}^M \lambda_l m_{li} \right] [1 - I(X_i \geq t^*) Z_i]} e^{-\left[\sum_{l=1}^M \lambda_l m_{l*} \right] I(X_i \geq t^*) Z_i} f(Z_i) \\ &= \prod_{l=1}^M \lambda_l^{\Delta_l} \prod_{i=1}^N e^{\beta I(X_i \geq t^*) \delta_i Z_i} e^{-e^{\beta} \left[\sum_{l=1}^M \lambda_l (m_{li} - m_{l*}) \right] I(X_i \geq t^*) Z_i} e^{-\left[\sum_{l=1}^M \lambda_l m_{li} \right] [1 - I(X_i \geq t^*) Z_i]} \\ &\quad \times e^{-\left[\sum_{l=1}^M \lambda_l m_{l*} \right] I(X_i \geq t^*) Z_i} f(Z_i) \end{aligned} \quad (2)$$

where $\Delta_{li} = I\{T_{(l)} = X_i\}$, $l = 1, 2, \dots, M$, is the indicator of whether the l^{th} failure occurs at the observational time X_i and $\Delta_l = \sum_{i=1}^N \Delta_{li}$ is the multiplicity at the distinct event time $T_{(l)}$. Furthermore, $m_{li} = (\min\{T_{(l)}, X_i\} - T_{(l-1)})I(T_{(l-1)} \leq X_i \leq \tau - e_i)$, $m_{l*} = (\min\{T_{(l)}, t^*\} - T_{(l-1)})I(T_{(l-1)} \leq X_i \leq \tau - e_i)$ measure the portion of the observational time X_i or the lag time t^* overlaps with the l^{th} event time interval $(T_{(l-1)}, T_{(l)})$, respectively.

3.1 | The PRIME EM algorithm

We consider maximizing the likelihood function with respect to β and λ given the latent variable $\mathbf{Z}$ and resort to the EM algorithm^{30–32} for estimating the treatment effect, while incorporating the aggregated information on the proportion of treated responders p . The *PRIME EM algorithm* works as follows:

- Step 0. Set $k = 0$. Choose initial estimates of $\mathbf{Z}$, λ and β as $\mathbf{Z}^{(0)}$, $\lambda^{(0)}$, and $\beta^{(0)}$, respectively. Typically, we can set the expectation of $Z_i^{(0)}$, $E[Z_i^{(0)}] = 1/2$ for $i \in E$, $\lambda^{(0)}$ as the estimates from the Cox PH model ignoring the response dichotomy among the treated, and $\beta^{(0)} = 0$. Repeat the following steps.
- Step 1 (M-step). Update $\beta^{(k+1)}$ and $\lambda^{(k+1)}$ sequentially:
 - Step 1.1. First update $\beta^{(k+1)}$ based on the estimates of $\lambda^{(k)}$ and expectation of $\mathbf{Z}^{(k)}$, $E[\mathbf{Z}^{(k)}]$, using the formula

$$\beta^{(k+1)} = \log \frac{C^{(k)}}{D^{(k)}}.$$

where

$$C^{(k)} = \sum_{i=1}^N I(X_i \geq t^*) E[Z_i^{(k)}] \delta_i,$$

$$D^{(k)} = \sum_{i=1}^N I(X_i \geq t^*) E[Z_i^{(k)}] \left[\sum_{l=1}^M \lambda_l^{(k)} (m_{li} - m_{l*}) \right].$$

– *Step 1.2.* Next update $\lambda^{(k+1)}$ given the current estimates of $\beta^{(k+1)}$ and $E[Z^{(k)}]$ as

$$\lambda_l^{(k+1)} = \frac{\Delta_l}{A_l^{(k)} + B_l^{(k)}}.$$

where

$$A_l^{(k)} = \sum_{i=1}^N \{I(X_i \geq t^*) E[Z_i^{(k)}] m_{li*} + [1 - I(X_i \geq t^*) E[Z_i^{(k)}]] m_{li}\},$$

$$B_l^{(k)} = e^{\beta^{(k+1)}} \sum_{i=1}^N [I(X_i \geq t^*) E[Z_i^{(k)}] (m_{li} - m_{l*})].$$

- **Step 2 (E-step).** Update $E[Z_i^{(k+1)}]$ for the treated subject i , where $i \in E$, conditional on the updated estimates of $\lambda^{(k+1)}$ and $\beta^{(k+1)}$,

$$E[Z_i^{(k+1)}] = \frac{pf_1(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)})}{pf_1(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)}) + (1-p)f_2(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)})}.$$

where

$$\begin{aligned} f_1(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)}) &= \prod_{l=1}^M (\lambda_l^{(k+1)})^{\Delta_{li}} e^{\beta^{(k+1)}} I(X_i \geq t^*) \delta_i e^{-I(X_i < t^*) \left[\sum_{l=1}^M \lambda_l^{(k+1)} m_{li} \right]} \\ &\times e^{-\left[\left(\sum_{l=1}^M \lambda_l^{(k+1)} m_{li*} \right) + e^{\beta^{(k+1)}} \left(\sum_{l=1}^M \lambda_l^{(k+1)} (m_{li} - m_{l*}) \right) \right] I(X_i \geq t^*)}, \\ &= \prod_{l=1}^M (\lambda_l^{(k+1)})^{\Delta_{li}} e^{-\left(\sum_{l=1}^M \lambda_l^{(k+1)} m_{li} \right)}. \end{aligned}$$

- **Step 3.** Repeat Step 1 to Step 2 until the log likelihood value converges, that is, $|\log L(\beta^{(k+1)}, \lambda^{(k+1)}, Z^{(k+1)}) - \log L(\beta^{(k)}, \lambda^{(k)}, Z^{(k)})| \leq \epsilon$ for the prespecified ϵ .

The estimate of β obtained at the final step, $\hat{\beta}$, is the treatment effect estimate of the *PRIME EM algorithm*. In addition, $\hat{\lambda}$, $\hat{Z}$ provide the estimates of baseline hazards and probability of being a responder.

3.2 | The PRIME likelihood ratio test

To test the significance of the estimated treatment effect, we adopt the likelihood ratio test embedded with p through the test statistic. The *PRIME likelihood ratio test* of the sharp null hypothesis of no treatment effect works as follows. For any prespecified p , we estimate the likelihood ratio test statistic by calculating the ratio of two marginal likelihoods, specifically one found by maximizing the marginal likelihood function using the maximum likelihood estimates obtained via *PRIME EM* at convergence, $\hat{\beta}$, $\hat{\lambda}$, $\hat{Z}$, and another found by calculating the marginal likelihood function under the null setting, $\beta^0 = 0, Z^0 = p$ and $\hat{\lambda}^0$, where $\hat{\lambda}^0$ can be obtained by maximizing the full likelihood function given $\beta^0 = 0$ and Z^0 .

Here, we derive a way to obtain the marginal likelihood function directly using the results from EM algorithm. The relationship between the full and marginal likelihood function is

$$f(X, \Delta, Z) = f(X, \Delta)f(Z | X, \Delta) \Rightarrow f(X, \Delta) = \frac{f(X, \Delta, Z)}{f(Z | X, \Delta)}.$$

Taking the logarithm gives us that

$$\log f(X, \Delta) = \log f(X, \Delta, Z) - \log f(Z | X, \Delta) = \log f(X, \Delta | Z) + \log f(Z) - \log f(Z | X, \Delta).$$

That is,

$$\ell_m(\beta, \lambda; X, \Delta) = \log L(\beta, \lambda, Z; X, \Delta) - \log f(Z | X, \Delta, \beta, \lambda).$$

where $\ell_m(\cdot)$ denotes the marginal log likelihood function.

Suppose that the *PRIME EM algorithm* converges at iteration K . The conditional distribution of Z given X at convergence follows a Bernoulli distribution with the probability of success $E[Z^{(K)}]$, as shown in the M step of the EM algorithm. The marginal log likelihood $\ell_m(\cdot)$ at convergence becomes

$$\begin{aligned}
 \ell_m(\beta^{(K)}, \lambda^{(K)}; \mathbf{X}, \Delta) = & \sum_{l=1}^M \Delta_l \log(\lambda_l^{(K)}) + \sum_{i=1}^N I(X_i \geq t^*) E[Z_i^{(K)}] \\
 & \left\{ \beta^{(K)} \delta_i - e^{\beta^{(K)}} \left[\sum_{l=1}^M \lambda_l^{(K)} (m_{li} - m_{l*}) \right] \right\} \\
 & + \sum_{i=1}^N \left[-I(X_i \geq t^*) E[Z_i^{(K)}] \left(\sum_{l=1}^M \lambda_l^{(K)} m_{li} \right) \right] \\
 & + \sum_{i=1}^N \left\{ -[1 - I(X_i \geq t^*) E[Z_i^{(K)}]] \left(\sum_{l=1}^M \lambda_l^{(K)} m_{li} \right) \right\} \\
 & + \sum_{i=1}^N I(i \in T) \left[E[Z_i^{(K)}] \log \left(\frac{p}{E[Z_i^{(K)}]} \right) \right. \\
 & \left. + (1 - E[Z_i^{(K)}]) \log \left(\frac{1-p}{1 - E[Z_i^{(K)}]} \right) \right].
 \end{aligned} \tag{3}$$

Thus, the likelihood ratio test statistic of the sharp null hypothesis of no treatment effect is given by $-2[\ell_m(\beta^{(K)}, \lambda^{(K)}; \mathbf{X}, \Delta) - \ell_m^0]$ where ℓ_m^0 can be obtained by replacing $\beta^{(K)}$ with β^0 , $\lambda^{(K)}$ with $\hat{\lambda}^0$ and $E[Z^{(K)}]$ with p in Equation (3).

3.3 | The PRIME sample size and power calculation

Size and power of the *PRIME likelihood ratio test* can be evaluated with simulations. Given a sample size N , the baseline proportion of responders among the treated p , the enrollment duration A , the total study duration τ , the duration of treatment time-lag t^* , and the baseline hazard h_C as well as the hazard for the responding patients after the delay h_E , the empirical power (EP) is computed for each value of the assumed treatment effect $\beta = \log(h_E/h_C)$ and a given sample size N based on the *PRIME sample size and power calculation algorithm* as follows:

Step 1 Draw subjects' enrollment times e from a Poisson process with an intensity rate $a = N/A$;

Step 2 Randomize subjects to the treatment or control groups;

Step 3 Draw the responders' memberships Z for each subject in the treatment group from a Bernoulli distribution with the probability of success of p and set $Z=0$ for the subjects in the control group;

Step 4 Simulate event times T_β from

– $T_\beta \sim \text{pexp}(h_C, h_E)$ for responders, where $\text{pexp}(\cdot)$ denotes the piecewise exponential distribution function with the rate varying at t^* .

– $T_\beta \sim \exp(h_C)$ for nonresponders and subjects in the control arm, where $\exp(\cdot)$ denotes the regular exponential distribution;

Step 5 Define the observational times $X = \min \{ T_\beta - e \}$ and the event indicators $I = \{ T_\beta - \tau - e \}$;

Step 6 Apply the *PRIME likelihood ratio test* with the prespecified p to compute the p-value p_{β} ;

Step 7 Repeat *Step 1* to *6* for a large number of B times and compute the power for the given the treatment effect β as the proportion of iterations whose p_{β} are less than or equal to α .

Worthy of note, when computing the EP given the sample size using the PRIME algorithm, the responder membership is first simulated to facilitate drawing the event times for responders and nonresponders/controls, respectively, and then assumed latent to estimate the treatment effect and perform the hypothesis testing. Thus, the simulation-based power calculation algorithm can mimic the practical scenarios with unknown responder membership.

To determine the sample size required to achieve the target power, this algorithm performs a grid search of sample size over a prespecified plausible range. The bisection method is adopted to conduct a smart search. In particular, each candidate sample size identified in the bisection search is evaluated for its EP and the one whose EP is closest to the target is then selected as the sample size required to achieve the target power by PRIME.

4 | SIMULATION STUDIES

In order to evaluate the performance of the PRIME strategy, we conduct simulation studies to examine the treatment effect estimation obtained from the *PRIME EM Algorithm* and the sample size and power relationship derived from the PRIME design and analysis strategy, under various magnitudes of p . To compare, we also explore the performances of the existing alternatives such as the standard practice. For all simulations, the EP is computed using 1000 replications and the significance level is set at two-sided 0.05.

4.1 | Evaluation #1: The property of treatment effect estimation of PRIME EM algorithm

We first evaluate the properties of treatment effect estimate, $\hat{\beta}$, based on the PRIME EM Algorithm incorporating p in terms of bias, empirical standard error and coverage of 95% confidence intervals.

Table 1 presents the simulation results. The *PRIME EM Algorithm* generally performs well. There is virtually no evidence of bias in the estimation of β and the empirical coverage of 95% confidence intervals are close to the nominal level in all cases considered. Worthy of note, the empirical standard errors of estimates from the PRIME strategy increase as p decreases, due to the fact that only the treated subjects who respond can contribute to the detection of a purified treatment effect specific for responders.

4.2 | Evaluation #2: The size and power of the PRIME likelihood ratio test

Size and power of the *PRIME likelihood ratio test* are evaluated with simulations, where the power is computed based on 1000 simulations. We report the results of such simulations for HRs ranging from 1.0 to 0.3 in Figure 3. The proposed test leads to the correct nominal level when the treatment has no effect and results in an adequate power as the treatment has a stronger effect. As p increases, for a given sample size, PRIME testing procedure becomes

more efficient, achieving substantially larger power under $100p\% = 50\%$ than under $100p\% = 30\%$.

4.3 | Evaluation #3: The quantitative impact of treating nonresponders on study efficiency

To quantify the impact of treating nonresponders on study efficiency, we calculate the sample size required to achieve the target power under various magnitudes of p , using the *PRIME sample size and power calculation algorithm* incorporating both the dichotomized response and lag effect. The trajectory of N as p varies reveals the impact of p on study efficiency and further implies the benefits of enhancing p on improving the efficiency.

In this evaluation, we first set the value of $100p\%$ as small as 20% and the HR for responders vs nonresponders/controls as significant as 0.01 after a 2-months lag. Such a design setting is motivated by the seemingly unique characteristics of cancer immunotherapy: long-term survival for a small subset of treated subjects and delayed onset of benefit.^{12,33–35} We then specifically, determine the design parameters, aiming to mimic the features of a well-cited Nivolumab Non-Small-Cell Lung Cancer (NSCLC) study.³³ With a 29-months of study and 15.8-months of enrollment period, the Nivolumab study demonstrates an average HR for all treated subjects against controls of 0.73 with approximate 20% treated subjects responding to Nivolumab after a 2-months of delay. Adapting to this design setting, we estimate the underlying HR for responders vs controls via simulations: Given a candidate responders-vs-nonresponders/controls HR, we first simulate the survival outcomes using *Steps 1–5* of the *PRIME sample size and power calculation algorithm* outlined in Section 3.3; We then fit a Cox PH model to estimate the diluted treatment-vs-control HR; The above process is repeated for 1000 times and the average of diluted HRs is recorded. Next, we vary the candidate values of responders-vs-nonresponders/controls HRs within a plausible range and search for the one leading to the observed diluted treatment-vs-controls HR at 0.73. That yields a responders-vs-nonresponders/controls HR of 0.01.

To explore the impact of p on study efficiency, we vary $100p\%$ from 20% to 50% in an incremental step of 10% while fixing the effect size of responders-vs-nonresponders/controls HR at 0.01, where the range of p remains coherent with the recent results of IO studies.^{33,34,36–38}

For comparison, we also evaluate the relationship between N and p using the conventional design strategy. The conventional alternative ignores both the lag effect and dichotomized response, thereby usually adopting a diluted HR between treatment and control either observed in prior studies or reported in the literature. In this evaluation, such a diluted treatment-vs-control HR is calculated via simulation, given the responders-vs-nonresponders/controls HR of 0.01 and under the same design setting of Nivolumab NSCLC study except that $100p\%$ would vary between 20% and 50%. In particular, we first simulate the survival outcomes while mixing up $100p\%$ responders with $100(1-p)\%$ nonresponders in the treatment population then estimate the HR using a Cox PH model; this process is repeated for 1000 times and the average of HR estimates is recorded as the diluted HR.

Table 2 shows that when p increases as the study treats more therapeutic-sensitive patients, the average, diluted HR would gradually decrease. As such, both the conventional and PRIME designs become more efficient by requiring fewer subjects to achieve the target power. For example, when only 20% of treated subjects react to the experimental therapy, the average HR for all treated subjects against controls is 0.74 even though the HR for responders vs controls is 0.01, leaving the conventional design to require 455 subjects to detect such a significant treatment effect with the target power. When half of treated subjects respond, on the other hand, the average HR for the treated decreases to 0.46 and thus the conventional design only entails 73 subjects. A similar declining trend is observed in the PRIME strategy as p rises. This result makes intuitive sense and the heuristic argument is as follows. If p is small, the significant benefit of treated responders is diluted by the null effect of a large number of treated nonresponders when the treatment effect is evaluated by comparing the entire treatment group with controls; thus, to detect such an effect with the desired power, investigators would need to enrich the study without knowing which subset to enrich and that leaves the magnitude of enrichment inefficiently large. In sharp contrast, enhancing p by treating more, if not all, responders, can dramatically reduce the study size and improve the study efficiency.

Even though both designs gain efficiency as p increases, the proposed PRIME strategy outperforms its conventional alternative under the same magnitude of p . For instance, PRIME requires 117 subjects whereas the conventional design, 189 subjects, given a 100% of 30%. This implies that a highly dichotomized treatment population with a small subset of responders seriously violates the PH assumption as manifested in various NPH patterns, causing a severe loss of efficiency for the conventional design based on the log-rank test. On the other hand, such a dichotomized treatment population is properly recognized by PRIME in the design and analysis procedures, thereby yielding a dramatic efficiency gain.

This evaluation underscores two possible strategies to improve IO trials' efficiency: First, properly recognize the response dichotomy by performing the PRIME design; Second, enroll more therapeutic-sensitive subjects through the precision design for immunotherapy studies.

4.4 | Evaluation #4: How treating nonresponders affecting study efficiency vary by trial parameters

The trial parameters, such as the magnitude of treatment effect (β), trial duration (τ) as well as lag duration (t^*), may affect the role that treating nonresponders plays on study efficiency. In what follows, we further explore such an effect under various scenarios. The general design setting adopted in this evaluation also mimics the well-cited Nivolumab NSCLC study,³³ except that the trial parameters of interest such as β , τ or t^* are allowed to vary one at a time and within a practically reasonable range to explore its impact.

4.4.1 | Trial duration—As the trial duration varies, the relationship between the response dichotomy and required sample size follows a similar pattern (Figure 4). To attain the target power under a given p , if the study runs in a shorter duration, a larger size N is required and conversely, if the study enrolls fewer subjects, a longer duration becomes

necessary. This interdependence between trial duration and study size diminishes as p increases. For instance, when $100p\%$ reaches 60%, N does not differ much as the trial duration varies among 19, 29, or 39 months. As $100p\%$ decreases, however, both study duration and sample size would need to be inflated in order to achieve the target power. For example, in a 19-months study with a 36.8 subjects/month enrollment rate, a mixed treatment group containing only 20% of responders makes it impossible to achieve the target power no matter how much N is inflated. This is due to the study duration constraint that prevents the latecomers from being followed up sufficiently long until their events take place. In this case, if the study duration is extended from 19 months to 29 or 39 months, the target power can then be achieved by properly increasing N . Implicit in this finding is the sweeping benefits of enhancing p : it makes the IO trials run faster and smaller.

4.4.2 | Magnitude of treatment effect—As the treatment effect diminishes, the association that the smaller the p , the larger the N becomes more apparent, as reflected by the steeper trends in Figure 5. For example, when the HR contrasting responders with nonresponders or controls is 0.01, a mixed treatment group consisting of 60% of responders requires 35 subjects to achieve the target power whereas a treatment group with 20% of responders needs 265 subjects. When the HR is 0.05, the same decrease in $100p\%$ from 60% to 20% gives rise to a more sizable increase in N from 42 to 348. The most extreme case occurs when the HR becomes 0.1, where a reduction in $100p\%$ from 60% to 20% results in inflation in N from 49 to 431. Thus, enhancing p by treating more prospective responders shows to be a sensible approach to improve the study efficiency and its benefits becomes more apparent for less effective treatment regimens than for more efficacious ones.

4.4.3 | Lag duration—Figure 6 demonstrates roughly similar trends between p and N in spite of the lag duration of various lengths, although N is inflated slightly more due to a smaller p under a longer lag. Given the same p , the presence of a time-lag effect or a lag of longer duration would generally require a larger study size to achieve the target power in IO trials, as expected.

4.5 | Evaluation #5: Impact of misspecifying p

Apparently, the PRIME design and analysis procedures depend upon the prespecification of p , but in practice, the true value of p is unknown. Then how robust is the proposed strategy against the misspecification of p ?

To address the robustness property of the proposed strategy, we examine the EP when p is misspecified in both design and analysis stages. Explicitly, we first compute the required sample size N using the *PRIME sample size and power calculation algorithm* under a misspecified p^m , while assuming the HR of responders vs nonresponders/controls is correctly specified; next, given N , we calculate the EP that the proposed testing procedure can actually achieve also under p^m . The EP is computed as follows. We first draw the response membership for each treated subject from a binomial distribution with the probability of success specified as the true p , then simulate the survival outcomes for the responding patients from a piecewise exponential distribution and for the nonresponding patients or controls from a regular exponential distribution, respectively. Finally, we

compute the p-values using the *PRIME likelihood ratio test* based on the misspecified p^m . The aforementioned process is repeated 1000 times to obtain the EP. Note that p is misspecified at both design and analysis stages but correctly specified during the data generation process, to mimic the practical scenarios.

By contrast, the conventional strategy based on the regular log-rank test is also performed. As alluded to in Section 4.3, the conventional strategy usually adopts a diluted HR between treatment and control and ignores the dichotomized treatment response. Thus, we assume the diluted HR is misspecified in the conventional approach. As the diluted HR involves the HR of responders and the proportion of the treatment population those responders account for (p), once p is misspecified, so would be the diluted HR. In this evaluation, we construct such a misspecified HR in the simulations by first drawing the survival outcomes while mixing up $100p^m\%$ responders with $100(1-p^m)\%$ nonresponders in the treatment arm then estimating the HR between treatment and control; this process is repeated for 1000 times and the average HR estimate yields a misspecified diluted HR. Given the misspecified HR, we use the conventional strategy to calculate the sample size required to achieve the target power and evaluate the EP that such a design can actually achieve under various misspecification scenarios.

Table 3 illustrates the impact of misspecifying p . As expected, overspecifying or underspecifying p can lead to underpowered or overpowered studies, respectively, and the magnitude of deviation from the target power is severe when p is small but diminishes when p gets large. For example, the PRIME design requires 49 subjects in order to achieve the target power when $100p\%$ is misspecified to 50%. If the truth falls below the misspecified at 40% or above it at 60%, then the *PRIME likelihood ratio test* given these many subjects could be slightly underpowered (EP = 63.0%) or overpowered (EP = 91.2%), where the deviation from the target power is less than 20%. By contrast, if $100p\%$ is small, even the same magnitude of misspecification of 10% could lead to a more severe deviation from the target. For instance, if the truth lies at 10% or 30% but misspecified to be 20%, the PRIME design necessitates 265 patients and given these many patients the resulting EP is 28.7% or 98.4%. This evaluation reveals a substantial benefit of increasing $100p\%$ from another angle: treating more responders can mitigate the misspecification risk for the PRIME strategy.

Under all circumstances considered, the conventional strategy based on the regular log-rank test ignoring treating nonresponders and assuming a misspecified diluted HR requires a much larger study size relative to PRIME, while suffering the underpowered or overpowered risk as much as PRIME when p is misspecified. Intuitively thinking, the conventional approach is misspecified from two perspectives. First, it ignores treating nonresponders by assuming that all enrolled patients would uniformly respond to the given treatment regimen, which could be regarded as overspecifying $100p\%$ to 100%, a severe overspecification that could well lead to a serious damage to study efficiency. Second, it misspecifies p constituting the diluted treatment-vs-control HR. These misspecification result in an inflation in study size for the conventional approach as opposed to PRIME and unfortunately, such an inflation cannot protect it against the misspecification risk. Specifically, it cannot salvage the conventional approach from the underpowered risk when the diluted HR involves an overspecified p and makes it prone to the overpowered risk when

the diluted HR is based on an underspecified p . For instance, when 100% constituting the diluted HR is misspecified as 20%, the conventional design requires 455 subjects but can only achieve 31.1% power empirically if the true 100% falls 10% below the misspecified 20% or otherwise, 98.7% power if the truth lies 10% above the misspecified value. We therefore conclude that, although PRIME does not improve on protecting against the misspecification risk as opposed to its conventional alternative, it at least gain in study efficiency without a sacrifice on robustness.

5 | REDESIGNING NIVOLUMAB NSCLC STUDY USING PRIME DESIGN

In this section, we consider redesigning, retrospectively, the Nivolumab NSCLC Study³³ using the proposed PRIME strategy. Assuming that the Nivolumab study is originally designed in a conventional manner, such a comparison reveals the relative efficiency gain that the PRIME design could otherwise bring when the PH assumption is violated.

The Nivolumab study is designed to evaluate the possible differential effects of Nivolumab, a PD-1 immune-checkpoint-inhibitor antibody, vs chemotherapy Docetaxel in previously treated metastatic nonsquamous NSCLC. The primary endpoint is OS and the observed median OS is 9.4 months in the Docetaxel group, resulting in a baseline hazard of 0.074 under the exponentially distributed event time assumption. The observed average treatment-vs-control HR is 0.73. Under an equal allocation ratio, a total study duration and minimum follow-up time for OS are 29 and 13.2 months, respectively. The observed median time to response is 2.1 months, which is reflected in the delayed separation of survival curves by 2-months (Figure 1C). To achieve the target power of 90%, the study requires 582 subjects based on the conventional design.

To redesign the study using the PRIME strategy, we need to prespecify p as well as the HR among the treated patients who respond to nivolumab. Brahmer et al³³ report that the objective response rate among the nivolumab group is approximately 20% and an average HR contrasting nivolumab and docetaxel is 0.73. We thus set 100% at 20% and estimate the underlying HR for the treated responders such that a mixture of 20% of responders and 80% of nonresponders in the nivolumab group can lead the average HR between nivolumab and docetaxel to 0.73, via the following simulation-based grid searching algorithm: For each candidate value of responders-vs-nonresponders/controls HR, λ_p , in a plausible range between 0.01 to 0.15,

1. *Step 1.* We randomize patients to the treatment and control arms at 1:1 ratio to obtain the treatment assignments;
2. *Step 2.* We then generate the survival outcome for each patient assuming the HR for the treated responders vs nonresponders/controls as λ_p and for the treated nonresponders vs controls as 1, respectively, and determine the censoring status under the administrative censoring mechanism;
3. *Step 3.* We next carry out the Cox PH model to estimate the average HR λ_{avg} , in the same manner as how the data is analyzed in the published study;³³

4. *Step 4.* Steps 1 to 3 are repeated for 1000 times and for each time, λ_{avg} is recorded.

The mean of the average HR λ_{avg} is estimated across the 1000 replications. Among the candidate values of $\lambda'_p S$, a HR of 0.01 whose mean value is closest to the observed value of 0.73 is then selected as the underlying HR for treated responders, when such responders account for 20% of the treatment population. In other words, we can assume that an average, diluted HR of 0.73 can be observed if the underlying HR among treated responders is 0.01. Furthermore, based on a $100p\%$ of 20% and a HR of 0.01 among treated responders, we apply the PRIME strategy to design and analyze the study and conclude that a total of 400 subjects can adequately achieve 90% of power.

Apparently, PRIME leads to a significant efficiency gain relative to the conventional design, by reducing the required sample size from 582 to 400 subjects while achieving the same degree of 90% power. Such an efficiency gain can be justified as follows. The survival patterns in Figure 1C³³ displays a crossing hazards pattern, where the survival curve of the Nivolumab group plummets faster at the beginning of the study, than that of the Docetaxel group but declines much slower toward the end of the study, than that of the Docetaxel group. The change in the declination rate implies that the PH assumption does not hold in the observed data. As the regular log-rank test is most efficient under the PH assumption, a serious violation of the PH assumption results in a severe efficiency loss of the conventional design based on the log-rank test. In sharp contrast, the PRIME design properly accounts for the mixture of responders and nonresponders in the treatment group during the trial design and analysis stages, thereby achieving a substantial efficiency gain.

6 | DISCUSSION

Properly designing IO trials is challenging, mainly due to the prevalence of various NPH patterns manifested in KM curves. In this article, we aim to address the NPH patterns from the integrated design and analysis perspectives. We first identify treating nonresponders as one of the possible, if not primarily, causes of the NPH patterns in addition to the indirect working mechanism of immunotherapy, and then propose a strategy to incorporate the response dichotomy thus incurred into the design and analysis strategy. The proposed strategy allows us to quantify the impact of treating nonresponders on study characteristics and thereby properly inflate the sample size to ensure an adequate study power, with the aim to reduce the risk of false-negative findings despite complex NPH patterns. Furthermore, our proposal suggests a solution to enhance the IO trial efficiency and mitigate the occurrence of NPH patterns by enrolling more prospective responders identified via predictive biomarkers.

To properly implement the PRIME strategy in practice comes upon a critical question: how to properly prespecify p ? The question can be addressed by applying the *PRIME profile likelihood-based approach* on early phase studies, if that becomes available. To do that, one can conduct a grid search of $100p\%$ over the candidate values of 10%, 20%, ..., 90% or on an even lower level of granularity. Each candidate value can be input into the *PRIME EM algorithm* to maximize the marginal likelihood function with respect to β and λ as in Equation (3) and repeating such a process for all candidates thereby constructs the profile

likelihood of p . The one that maximizes the profile likelihood function is chosen to be the estimate of p . We investigate the proposed PRIME approach using a simulation study, in which the event times are repeatedly simulated under the true $100p\%$ for 1000 times and each time, $100p\%$ is estimated via PRIME with a grid of candidate values in an incremental step of 1% over the range of 20% to 90%. Among the 1000 estimates, the median provides a consistent estimate of the true $100p\%$ regardless of where the truth lands, where the choice of median over mean is due to the skewed distribution of $100p\%$. For example, the median estimates of $100p\%$ are 21.6%, 50.8%, and 79.6% while the truth lies at 20%, 50%, and 80%, respectively. This ensures that in practice, once the study results of the early phase trials, such as the randomized phase 2 studies, become available, one can apply *PRIME EM algorithm* to estimate θ^β and *PRIME profile likelihood-based approach* to estimate $100p\%$, then use that information to facilitate a proper design of the confirmatory trial. Such a practice recommendation is consistent with that of the US Food and Drug Administration Guidance on Clinical Considerations for Therapeutic Cancer Vaccines (2011) that a randomized Phase 2 trial is highly recommended to be conducted prior to the transition to confirmatory trials, to help determine the appropriate sample size and treatment effect size. Properly specifying $100p\%$, although thus feasible, nevertheless is still challenging and misspecifying it is sometimes inevitable. As such, we empirically address the robustness property of the PRIME strategy against the misspecification risk in Section 4.5. This issue is warranted for further research.

More importantly, it is feasible to enroll a patient population with a large p , when a predictive biomarker is available in addition to the early phase study results. For instance, the PD-1/PD-L1 expression has been discovered as a predictive biomarker for cancer immunotherapy in several different malignancies including, but not limited to, melanoma, lung, kidney, and bladder cancers.³⁹ Emerging data suggest that patients whose tumors express PD-L1 positive have derived improved clinical benefits compared with those with PD-L1 expression negative. At Phase 2, if patients' PD-L1 expression levels are collected, then patients can be classified into the corresponding PD-L1 positive or negative subgroups and the *PRIME profile likelihood-based approach* and *EM algorithm* can then be performed to estimate the subgroup-specific p and θ^β . Investigators can then exclusively enroll patients associated with the larger p for Phase 3 (in this case, the PD-L1 positive patients), in order to construct a patient population of a high likelihood of therapeutic response.

The PRIME strategy, combined with predictive biomarkers, outlines a path toward precision IO trials. As the immunotherapy is not immune to the concept of targeted therapies for targeted populations, the rational administration of immunotherapy would require the rational use of predictive biomarkers to select the right patients more likely to respond. At present, the existing predictive biomarkers can only predict the patients' response probabilistically rather than definitively. That, however, makes the PRIME strategy particularly useful, as it recognizes the uncertainty in patients' responses and makes it embedded in the design and analysis procedure to detect, with an adequate power, a potentially effective therapy specific to the latent responders. With the refinement of existing biomarkers,^{40,41} identification of novel biomarkers⁴² and the use of artificial intelligence on large-scale prior studies,^{41,43–45} we believe that a more rational use of biomarkers could be ensured to refine the trial population with an increased p , and that can facilitate the proper

design of precision IO trials and help direct the administration of immunotherapy to the right patients in the real world.

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 Chief Scientist Intramural Challenge Grant of the U.S. Food and Drug Administration and the Intramural Research Program of the National Cancer Institute, National Institutes of Health, MD, USA. The authors would like to thank Drs. John Scott and Shiojwen Lee 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. In addition, the authors would like to thank Dr. Bill Wheeler for the computational support.

Funding information

Zhenzhen Xu was partially supported by the Chief Scientist Intramural Challenge Grant of the U.S. Food and Drug Administration. Bin Zhu was supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute.

APPENDIX A

In this appendix, we first show that the response dichotomy combined with the inadequate sample size can drive the NPH patterns primarily (Theorems 1 and 2), then derive the *PRIME EM algorithm*.

A.1 Theorem 1 proof

We first show that the response dichotomy drives the NPH Patterns and as the degree of response dichotomy dampens, the nonproportionality issue vanishes: Let $\lambda_R(t)$, $\lambda_{NR}(t)$, $\lambda_T(t)$, $\lambda_C(t)$ refer to the hazard functions and $\Lambda_R(t)$, $\Lambda_{NR}(t)$, $\Lambda_T(t)$, $\Lambda_C(t)$ the corresponding cumulative hazard functions for the responding, nonresponding, treated, and control patients, respectively. Furthermore, let $S_R(t)$, $S_{NR}(t)$, $S_T(t)$, $S_C(t)$ denote the survival functions for the responders, nonresponders, treated, and control patients. Thus, let $h = \lambda_R(t)/\lambda_{NR}(t) = \lambda_R(t)/\lambda_C(t)$, ($h < 1$), denote the constant HR between the responders and nonresponders whereas the hazard functions for the nonresponders and controls are same $\lambda_{NR}(t) = \lambda_C(t)$. In this article, we measure the chance of a treated patient responding to a given therapy with the proportion of responders among the treated patients at baseline, p , and that proportion remains unchanged throughout the lag period $[0, t^*]$. The underlying reason is that the treatment effect is not manifested itself until t^* , causing the treated patients experience events as fast as the controls do during $[0, t^*]$. Thus, we define the residual survival functions for the responders, nonresponders and control patients as

$$S_R^*(t) = S_R(t)/S_R(t^*) = \exp\left\{\int_{t^*}^t -\lambda_R(s)ds\right\},$$

$$S_{NR}^*(t) = S_{NR}(t)/S_{NR}(t^*) = \exp\left\{\int_{t^*}^t -\lambda_{NR}(s)ds\right\},$$

$$S_C^*(t) = S_C(t)/S_C(t^*) = \exp\left\{\int_{t^*}^t -\lambda_C(s)ds\right\},$$

where $S_{NR}(t) = S_C(t)$. Since the residual survival function for the treated subjects $S_T^*(t)$ can be written as the following two forms

$$S_T^*(t) = pS_R^*(t) + (1-p)S_{NR}^*(t) = p[S_C^*(t)]^h + (1-p)S_C^*(t), \quad (A1)$$

$$S_T^*(t) = \exp\{-[\Lambda_T(t) - \Lambda_T(t^*)]\} = \exp\left\{-\int_{t^*}^t \lambda_T(s)ds\right\}. \quad (A2)$$

The hazard functions $\lambda_T(t), \lambda_C(t)$ can be, respectively, derived as follows:

From the Equation (A1), we have

$$dS_T^*(t)/dt = -hp[S_C^*(t)]^{h-1}f_C(t)/S_C(t^*) - (1-p)f_C(t)/S_C(t^*),$$

since $dS_C^*(t)/dt = -\lambda_C(t)S_C^*(t) = -f_C(t)/S_C(t^*)$.

From the Equation (A2),

$$dS_T^*(t)/dt = -\lambda_T(t)\exp\left\{-\int_{t^*}^t \lambda_T(s)ds\right\} = -\lambda_T(t)S_T^*(t) = -f_T(t)/S_T(t^*).$$

Hence,

$$\lambda_T(t) = -\frac{dS_T^*(t)/dt}{S_T^*(t)} = \frac{hp[S_C^*(t)]^{h-1}f_C(t)/S_C(t^*) + (1-p)f_C(t)/S_C(t^*)}{S_T^*(t)},$$

$$\lambda_C(t) = \frac{f_C(t)}{S_C(t)} = \frac{f_C(t)/S_C(t^*)}{S_C(t)/S_C(t^*)} = \frac{f_C(t)/S_C(t^*)}{S_C^*(t)}.$$

The overall HR function between the treated and control subjects is

$$\begin{aligned}
 h(t) &= \frac{\lambda_T(t)}{\lambda_C(t)} \\
 &= \frac{\{hp[S_C^*(t)]^{h-1}f_C(t)/S_C(t^*) + (1-p)f_C(t)/S_C(t^*)\}/S_T^*(t)}{f_C(t)/S_C(t^*)S_C^*(t)} \\
 &= \frac{hp[S_C^*(t)]^h + (1-p)S_C^*(t)}{S_T^*(t)} \\
 &= \frac{h\{p[S_C^*(t)]^h + (1-p)S_C^*(t)\} + (1-p)(1-h)S_C^*(t)}{S_T^*(t)} \\
 &= h + (1-p)(1-h)\frac{S_C^*(t)}{S_T^*(t)} \\
 &= h + (1-p)(1-h)\frac{1}{p[S_C^*(t)]^{h-1} + (1-p)} \\
 &= h + \frac{1-h}{\frac{1}{p} - 1} \frac{1}{[S_C^*(t)]^{h-1} + 1}
 \end{aligned}$$

For any fixed value of h where $h < 1$, as time passes by, the residual survival function for the control group $S_C^*(t)$ decreases, $[S_C^*(t)]^{h-1}$ increases, $h(t) \propto \frac{1}{p[S_C^*(t)]^{h-1} + (1-p)}$ decreases,

leading to the overall HR function between treated and control subjects $h(t)$ decreases. Thus, $h(t)$ is a nonmonotonically increasing function of t which reaches its minimum at $t = t^*$ and maximum as t goes to infinity,

- When $t = t^* \Rightarrow S_C(t = t^*) = 1 \Rightarrow h(t) = h + (1-p)(1-h) = 1 - p + ph$
- When $t = +\infty \Rightarrow S_C^*(t) \rightarrow 0 \Rightarrow [S_C^*(t)]^{h-1} \rightarrow +\infty \Rightarrow \frac{1}{p[S_C^*(t)]^{h-1} + (1-p)} \rightarrow 0 \Rightarrow h(t) \rightarrow h$;

It follows that the HR $h(t)$ is not a constant proportion over time but varies within the range h to $1-p+ph$ as $t^* < t < +\infty$, which violates the PH assumption. It also shows that as the baseline proportion of treated responders increases, $h(t)$ decreases to h . When 100% = 100%, $h(t) \rightarrow h$. It implies that as the responders account for a larger proportion, if not all, of the treatment group, the HR between the treatment and control groups tends to a constant HR contrasting responders and nonresponders.

A.2 Theorem 2 proof

We then show that as the sample size goes to infinity, the log HR $\log\{h(t)\}$ is asymptotically proportional to $\log(h)$ and the proportionality $\log\{h(t)\}/\log(h) \rightarrow p$.

Under the regularity condition $\log\{h(t)\} = O(n^{-1/2})$ as $N \rightarrow \infty$ which implies that as the sample size N goes to infinity, the HR h is close to one. Thus, $\log(h)$ and $h - 1$ have the same order, that is, $\log\{h(t)\} \Leftrightarrow h - 1$ as $N \rightarrow \infty$, as

$$\frac{\log(h)}{h-1} = \frac{1/h}{1} = \frac{1}{h} \rightarrow 1 \text{ as } N \rightarrow +\infty,$$

where the derivatives were taken to both numerator and denominator. Then we have

$$\begin{aligned} \frac{\log\{h(t)\}}{\log(h)} &= \frac{\log\left\{h + (1-p)(1-h) \frac{1}{p[S_C^*(t)]^{h-1} + (1-p)}\right\}}{\log(h)} \\ &= \frac{h + (1-p)(1-h) \frac{1}{p[S_C^*(t)]^{h-1} + (1-p)} - 1}{\frac{h-1}{p[S_C^*(t)]^{h-1} + (1-p)}} \\ &= 1 - (1-p) \frac{1}{p[S_C^*(t)]^{h-1} + (1-p)} \\ &\rightarrow p \text{ when } h \rightarrow 1 \text{ as } N \rightarrow +\infty \end{aligned}$$

A.3 Derivation of PRIME EM algorithm

For responders,

$$\lambda_1(X_i) = \lambda_0(X_i) I(X_i < t^*) [\lambda_0(X_i) e^{\beta}]^{I(X_i \geq t^*)} = \lambda_0(X_i) e^{\beta I(X_i \geq t^*)},$$

$$S_1(X_i) = e^{-\Lambda_0(X_i) I(X_i < t^*)} - [\Lambda_0(t^*) + e^{\beta}(\Lambda_0(X_i) - \Lambda_0(t^*))] I(X_i \geq t^*),$$

and the pdf of the event time distribution for the responders becomes

$$\begin{aligned} f_1(X_i, \delta_i) &= \lambda_1(X_i)^{\delta_i} S_1(X_i) \\ &= \left(\lambda_0(X_i) e^{\beta I(X_i \geq t^*)} \right)^{\delta_i} e^{-\Lambda_0(X_i) I(X_i < t^*)} - [\Lambda_0(t^*) + e^{\beta}(\Lambda_0(X_i) - \Lambda_0(t^*))] I(X_i \geq t^*). \end{aligned}$$

For nonresponders or controls,

$$\lambda_2(X_i) = \lambda_0(X_i),$$

$$S_2(X_i) = e^{-\Lambda_0(X_i)},$$

$$f_2(X_i, \delta_i) = \lambda_2^{\delta_i}(X_i) S_2(X_i) = \lambda_0^{\delta_i}(X_i) e^{-\Lambda_0(X_i)}.$$

The full likelihood function is:

$$\begin{aligned} L(\beta, \lambda, Z; X, \Delta) &= L(\beta, \lambda, X, \Delta; Z) f(Z) \\ &= \prod_{i=1}^N f_1^{Z_i}(X_i, \delta_i) f_2^{(1-Z_i)}(X_i, \delta_i) p^{Z_i \cdot I(i \in E)} (1-p)^{(1-Z_i) \cdot I(i \in E)}, \end{aligned}$$

and after rearranging terms, it leads to

$$\begin{aligned} L(\beta, \lambda, Z; X, \Delta) &= \prod_{i=1}^N \lambda_0^{\delta_i}(X_i) \cdot e^{\beta I(X_i \geq t^*) \delta_i Z_i} \cdot e^{-e^{\beta} (\Lambda_0(X_i) - \Lambda_0(t^*)) I(X_i \geq t^*) Z_i} \\ &\quad \times e^{-\Lambda_0(X_i) [1 - I(X_i \geq t^*) Z_i]} \cdot e^{-\Lambda_0(t^*) I(X_i \geq t^*) Z_i} \cdot f(Z_i) \end{aligned}$$

Breslow estimator²³ is introduced to estimate the baseline hazard $\lambda_0(t)$ nonparametrically. Assume that the baseline hazard mass is a piecewise constant λ_l during the l^{th} event time interval $[t_{(l-1)}, t_{(l)}]$ defined by the distinct adjacent event times $t_{(l-1)}, t_{(l)}$, where $l = 1, 2, \dots, M$, then we have

$$\lambda_0^{\delta_i}(X_i) = \prod_{l=1}^M \lambda_l^{\Delta_{li}},$$

where $\Delta_{li} = I\{t_{(l)} = X_i\}$, $l = 1, 2, \dots, M$, indicates whether the i^{th} observational time X_i is the l^{th} event time $t_{(l)}$ and $\Delta_l = \sum_{i=1}^N \Delta_{li}$ denotes the multiplicity corresponding to $t_{(l)}$. Please note $\Delta_1 = \dots = \Delta_M = 1$ corresponds to the case of no tie.

$$\Lambda_0(X_i) = \sum_{l=1}^M \lambda_l [(\min\{t_{(l)}, X_i\} - t_{(l-1)}) I(t_{(l-1)} \leq X_i \leq \tau - e_i)] = \sum_{l=1}^M \lambda_l m_{li},$$

where $m_{li} = (\min\{t_{(l)}, X_i\} - t_{(l-1)}) I(t_{(l-1)} \leq X_i \leq \tau - e_i)$ measures how much time the i^{th} observational time (event or censored) covers the l^{th} event time interval;

$$\Lambda_0(t^*) = \sum_{l=1}^M \lambda_l [(\min\{t_{(l)}, t^*\} - t_{(l-1)}) I(t_{(l-1)} \leq t^* \leq \tau - e_i)] = \sum_{l=1}^M \lambda_l m_{l*},$$

where $m_{l*} = (\min\{t_{(l)}, t^*\} - t_{(l-1)}) I(t_{(l-1)} \leq t^* \leq \tau - e_i)$ measures how much time the change point t^* covers the l^{th} event time interval $[t_{(l-1)}, t_{(l)}]$. It follows that

$$\Lambda_0(X_i) - \Lambda_0(t^*) = \sum_{l=1}^M \lambda_l (m_{li} - m_{l*}).$$

Then the likelihood function becomes that

$$\begin{aligned}
 L(\beta, \lambda, Z; \mathbf{X}, \Delta) &= \left(\prod_{i=1}^N \prod_{l=1}^M \lambda_l^{\Delta_{li}} \right) \prod_{i=1}^N e^{\beta I(X_i \geq t^*) \delta_i Z_i} \cdot e^{-e^{\beta} \left[\sum_{l=1}^M \lambda_l (m_{li} - m_{l*}) \right]} I(X_i \geq t^*) Z_i \\
 &\quad \times e^{\left[\sum_{l=1}^M \lambda_l m_{li} \right] [1 - I(X_i \geq t^*) Z_i]} \cdot e^{\left[\sum_{l=1}^M \lambda_l m_{l*} \right] I(X_i \geq t^*) Z_i} \cdot f(Z_i). \\
 &= \prod_{l=1}^M \lambda_l^{\Delta_l} \prod_{i=1}^N e^{\beta I(X_i \geq t^*) \delta_i Z_i} \cdot e^{-e^{\beta} \left[\sum_{l=1}^M \lambda_l (m_{li} - m_{l*}) \right]} I(X_i \geq t^*) Z_i \\
 &\quad \times e^{\left[\sum_{l=1}^M \lambda_l m_{li} \right] [1 - I(X_i \geq t^*) Z_i]} \cdot e^{-\left[\sum_{l=1}^M \lambda_l m_{l*} \right] I(X_i \geq t^*) Z_i} \cdot f(Z_i)
 \end{aligned}$$

The log-likelihood function is

$$\begin{aligned}
 l(\beta, \lambda, Z; \mathbf{X}, \Delta) &= \sum_{l=1}^M \Delta_l \log(\lambda_l) + \sum_{i=1}^N I(X_i \geq t^*) Z_i \\
 &\quad \left\{ \beta \delta_i - e^{\beta} \left[\sum_{l=1}^M \lambda_l (m_{li} - m_{l*}) \right] \right\} + \sum_{i=1}^N \left[-I(X_i \geq t^*) Z_i \left(\sum_{l=1}^M \lambda_l m_{li} \right) \right] \\
 &\quad + \sum_{i=1}^N \left\{ -[1 - I(X_i \geq t^*) Z_i] \left(\sum_{l=1}^M \lambda_l m_{li} \right) \right\} + \sum_{i=1}^N \log(f(Z_i))
 \end{aligned} \tag{A3}$$

A.4 The PRIME EM algorithm

- M-Step: First update $\beta^{(k+1)}$ based on the estimates of $\lambda^{(k)}$ and $E[Z^{(k)}]$

$$\frac{\partial l(\beta, \lambda, Z; \mathbf{X}, \Delta)}{\partial \beta} = \sum_{i=1}^N I(X_i \geq t^*) Z_i \delta_i - e^{\beta} \sum_{i=1}^N I(X_i \geq t^*) Z_i \left[\sum_{l=1}^M \lambda_l (m_{li} - m_{l*}) \right] = 0.$$

This leads to

$$\beta^{(k+1)} = \log \frac{C^{(k)}}{D^{(k)}},$$

where

$$C^{(k)} = \sum_{i=1}^N I(X_i \geq t^*) E[Z_i^{(k)}] \delta_i,$$

$$D^{(k)} = \sum_{i=1}^N I(X_i \geq t^*) E[Z_i^{(k)}] \left[\sum_{l=1}^M \lambda_l^{(k)} (m_{li} - m_{l*}) \right].$$

Next update $\lambda^{(k+1)}$ given the current estimates of $\beta^{(k+1)}$ and $E[Z^{(k)}]$ as

$$\begin{aligned} \frac{\partial l(\beta, \lambda, Z; \mathbf{X}, \Delta)}{\partial \lambda_l} &= \frac{\Delta_l}{\lambda_l} - \sum_{i=1}^N [I(X_i \geq t^*) Z_i (m_{li} - m_{l*})] e^\beta \\ &\quad - \sum_{i=1}^N [I(X_i \geq t^*) Z_i m_{l*} + (1 - I(X_i \geq t^*) Z_i) m_{li}] = 0. \end{aligned}$$

This leads to

$$\lambda_l^{(k+1)} = \frac{\Delta_l}{A_l^{(k)} + B_l^{(k)}},$$

where

$$A_l^{(k)} = \sum_{i=1}^N \{I(X_i \geq t^*) E[Z_i^{(k)}] m_{l*} + [1 - I(X_i \geq t^*) E[Z_i^{(k)}]] m_{li}\},$$

$$B_l^{(k)} = e^{\beta^{(k+1)}} \sum_{i=1}^N [I(X_i \geq t^*) E[Z_i^{(k)}] (m_{li} - m_{l*})].$$

- E-Step: Update $E[Z_i^{(k+1)}]$ for the treated subject i , where $i \in E$, conditional on the updated estimates of $\lambda^{(k+1)}$ and $\beta^{(k+1)}$

$$E[Z_i^{(k+1)}] = \frac{p f_1(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)})}{p f_1(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)}) + (1-p) f_2(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)})},$$

where

$$\begin{aligned} &f_1(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)}) \\ &= \prod_{l=1}^M (\lambda_l^{(k+1)})^{\Delta_{li}} e^{\beta^{(k+1)}} I(X_i \geq t^*) \delta_i e^{-I(X_i < t^*) \left[\sum_{l=1}^M \lambda_l^{(k+1)} m_{li} \right]} \\ &\quad \times e^{-\left[\left(\sum_{l=1}^M \lambda_l^{(k+1)} m_{l*} \right) + e^{\beta^{(k+1)}} \left(\sum_{l=1}^M \lambda_l^{(k+1)} (m_{li} - m_{l*}) \right) \right] I(X_i \geq t^*)} \\ &f_2(X_i, \delta_i | \beta^{(k+1)}, \lambda^{(k+1)}) = \prod_{l=1}^M (\lambda_l^{(k+1)})^{\Delta_{li}} e^{-\left(\sum_{l=1}^M \lambda_l^{(k+1)} m_{li} \right)} \end{aligned}$$

REFERENCES

1. Xu Z, Zhen B, Park Y, Zhu B. Designing therapeutic cancer vaccine trials with delayed treatment effect. *Stat Med.* 2017;36(4):592–605. [PubMed: 27807870]
2. Xu Z, Park Y, Zhen B, Zhu B. Designing cancer immunotherapy trials with random treatment time-lag effect. *Stat Med.* 2018;37(30):4589–4609. [PubMed: 30203592]
3. Sit T, Liu M, Shnaidman M, Ying Z. Design and analysis of clinical trials in the presence of delayed treatment effect. *Stat Med.* 2016;35(11):1774–1779. [PubMed: 26833957]

4. Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*. 2011;30(19):2409–2421. [PubMed: 21611958]
5. Royston P, Parmar MK. Augmenting the Logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Med Res Methodol*. 2016;16(1):16. [PubMed: 26869168]
6. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*. 1989;45(2):497–507. [PubMed: 2765634]
7. Uno H, Tian L, Claggett B, Wei L. A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves. *Stat Med*. 2015;34(28):3680–3695. [PubMed: 26194988]
8. Hoos A, Parmiani G, Hege K, et al. A clinical development paradigm for cancer vaccines and related biologics. *J Immunother*. 2007;30(1):1–15. [PubMed: 17198079]
9. Sliwkowski MX, Mellman I. Antibody therapeutics in cancer. *Science*. 2013;341(6151):1192–1198. [PubMed: 24031011]
10. Melero I, Gaudernack G, Gerritsen W, et al. Therapeutic vaccines for cancer: an overview of clinical trials. *Nat Rev Clin Oncol*. 2014;11(9):509–524. [PubMed: 25001465]
11. Kantoff PW, Higano CS, Shore ND, et al. Sipuleucel-t immunotherapy for castration-resistant prostate cancer. *N Engl J Med*. 2010;363(5):411–422. [PubMed: 20818862]
12. Hodi FS, O'day SJ, McDermott DF, et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med*. 2010;2010(363):711–723.
13. Robert C, Long GV, Brady B, et al. Nivolumab in previously untreated melanoma without braf mutation. *N Engl J Med*. 2015;372(4):320–330. [PubMed: 25399552]
14. Schlom J, Gulley JL. Vaccines as an integral component of cancer immunotherapy. *JAMA*. 2018;320(21):2195–2196. [PubMed: 30419097]
15. Ventola CL. Cancer immunotherapy, part 3: Challenges and future trends. *Pharm Therapeut*. 2017;42(8):514.
16. Chiriva-Internati M, Bot A. A new era in cancer immunotherapy: discovering novel targets and reprogramming the immune system. *Int Rev Immunol*. 2015;34:101–103. [PubMed: 25901856]
17. Yang Y. Cancer immunotherapy: harnessing the immune system to battle cancer. *J Clin Invest*. 2015;125(9):3335–3337. [PubMed: 26325031]
18. Grady D. Harnessing the immune system to fight cancer. *The New York Times*; 2016.
19. Sharma P, Hu-Lieskovan S, Wargo JA, Ribas A. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*. 2017;168(4):707–723. [PubMed: 28187290]
20. Xu Z, Park Y, Liu K, Zhu B. Treating non-responders: pitfalls and implications for cancer immunotherapy trial design. *J Hematol Oncol*. 2020;13(1):1–11. 10.1186/s13045-020-0847-x, <https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-0847-x#citeas>. [PubMed: 31900191]
21. Hussain M, Rathkopf D, Liu G, et al. 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. *Eur J Cancer*. 2015;51(13):1714–1724. [PubMed: 26082390]
22. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N Engl J Med*. 2015;373(2):123–135. [PubMed: 26028407]
23. Breslow NE. Discussion of the paper by d. r. cox. *J Royal Stat Soc Ser B*. 1972;34:216–217.
24. Sy JP, Taylor JM. Estimation in a cox proportional hazards cure model. *Biometrics*. 2000;56(1):227–236. [PubMed: 10783800]
25. Wang Q, Hu B, Hu X, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell*. 2017;32(1):42–56. [PubMed: 28697342]
26. Wu YM, Cie lik M, Lonigro RJ, et al. Inactivation of cdk12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell*. 2018;173(7):1770–1782. [PubMed: 29906450]

27. Schmid P, Adams S, Rugo HS, et al. Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *N Engl J Med*. 2018;379(22):2108–2121. [PubMed: 30345906]
28. Finke LH, Wentworth K, Blumenstein B, Rudolph NS, Levitsky H, Hoos A. Lessons from randomized phase iii studies with active cancer immunotherapies—outcomes from the 2006 meeting of the cancer vaccine consortium (cvc). *Vaccine*. 2007;25:B97–B109. [PubMed: 17916465]
29. Anagnostou V, Yarchoan M, Hansen AR, et al. Immuno-oncology trial endpoints: capturing clinically meaningful activity. *Clin Cancer Res*. 2017;23(17):4959–4969. [PubMed: 28864724]
30. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*. 1993;80(2):267–278.
31. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *J Royal Stat Soc Ser B (Methodol)*. 1977;39(1):1–22.
32. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. New York: Wiley & Sons; 1987.
33. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. *N Engl J Med*. 2015;373(17):1627–1639. [PubMed: 26412456]
34. Larkin J, Chiarion-Sileni V, Gonzalez R, et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N Engl J Med*. 2015;373(1):23–34. [PubMed: 26027431]
35. Robert C, Thomas L, Bondarenko I, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N Engl J Med*. 2011;364(26):2517–2526. [PubMed: 21639810]
36. Schadendorf D, Hodi FS, Robert C, et al. Pooled analysis of long-term survival data from phase ii and phase iii trials of ipilimumab in unresectable or metastatic melanoma. *J Clin Oncol*. 2015;33(17):1889. [PubMed: 25667295]
37. Ribas A, Hamid O, Daud A, et al. Association of pembrolizumab with tumor response and survival among patients with advanced melanoma. *JAMA*. 2016;315(15):1600–1609. [PubMed: 27092830]
38. Topalian SL, Sznol M, McDermott DF, et al. Survival, durable tumor remission, and long-term safety in patients with advanced melanoma receiving nivolumab. *J Clin Oncol*. 2014;32(10):1020. [PubMed: 24590637]
39. Patel SP, Kurzrock R. Pd-1l expression as a predictive biomarker in cancer immunotherapy. *Mol Cancer Ther*. 2015;14(4):847–856. [PubMed: 25695955]
40. Samstein RM, Lee CH, Shoushtari AN, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics*. 2019;51:202–206. [PubMed: 30643254]
41. Jiang P, Gu S, Pan D, et al. Signatures of t cell dysfunction and exclusion predict cancer immunotherapy response. *Nat Med*. 2018;24(10):1550–1558. [PubMed: 30127393]
42. Sacher AG, Gandhi L. Biomarkers for the clinical use of pd-1/pd-l1 inhibitors in non–small-cell lung cancer: a review. *JAMA Oncol*. 2016;2(9):1217–1222. [PubMed: 27310809]
43. Cristescu R, Mogg R, Ayers M, et al. Pan-tumor genomic biomarkers for pd-1 checkpoint blockade–based immunotherapy. *Science*. 2018;362(6411):eaar3593. [PubMed: 30309915]
44. Auslander N, Zhang G, Lee JS, et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat Med*. 2018;24(10):1545–1549. [PubMed: 30127394]
45. Łuksza M, Riaz N, Makarov V, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*. 2017;551(7681):517–520. [PubMed: 29132144]

FIGURE 1.

Real study examples on complex survival patterns. A, Delayed treatment effect pattern: Nivolumab in Previously Untreated Melanoma without BRAF Mutation (N = 418); adapted from Reference 13. B, Belly shape diminishing effect pattern: A randomized noncomparative phase 2 trial of cixutumumab (IMC-A12) or ramucirumab (IMC-1121B) plus mitoxantrone and prednisone in men with metastatic docetaxel-pretreated castration-resistant prostate cancer; adapted from Reference 21. C, Crossing hazards pattern: Nivolumab vs Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer (N = 582); adapted from Reference 22. D, Delayed effect and crossing hazards combination pattern: Nivolumab vs Docetaxel in Advanced Squamous-Cell Non-Small-Cell Lung Cancer (N = 582); adapted from Reference 22. E, Delayed effect and belly shape diminishing effect combination pattern: Sipuleucel-T Immunotherapy for Castration-Resistant Prostate Cancer; adapted from Reference 11. F, Various belly shape diminishing effect combination pattern: A randomized noncomparative phase 2 trial of cixutumumab (IMC-A12) or ramucirumab (IMC-1121B) plus mitoxantrone and prednisone in men with metastatic docetaxel-pretreated castration-resistant prostate cancer; adapted from Reference 21

FIGURE 2.

Selected toy examples of nonproportional hazards (NPH) patterns of hypothetical trials simulated under various degrees of response dichotomy and sample size. The vertical line refers to the 3-months lag time. The top, middle, bottom rows of the panel plot are simulated under $100p\%=20\%$ and $N=200$, $100p\%=90\%$ and $N=200$, $100p\%=20\%$ and $N=2000$, respectively.

FIGURE 3.

Size and power evaluation of the *PRIME likelihood ratio test*. The horizontal gray dashed line refers to the 5% Type I Error Rate. Sample size is 200. Lag duration is 1 month. Study duration is 5 years. Enrollment rate is 15.8 subjects every month. Baseline hazards is 0.117. PRIME, P%-responder information embedded

FIGURE 4.

The impact of treating nonresponders on study efficiency as the trial duration varies. Target power is 80%. Lag duration is 2 months. Hazard ratio for responding patients is 0.01. Enrollment rate is 36.8 subjects/month. Type I error rate $\alpha = .05$

FIGURE 5.

The impact of treating nonresponders on study efficiency as the magnitude of treatment effect varies. Target power is 80%. Lag duration is 2 months. Study duration is 29 months. Enrollment rate is 36.8 subjects/month. Type I error rate $\alpha = .05$

FIGURE 6.

The impact of treating nonresponders on study efficiency as the lag duration varies. Target power is 80%. Hazard ratio for responding patients is 0.01. Study duration is 29 months. Enrollment rate is 36.8 subjects/month. Type I error rate $\alpha = .05$

TABLE 1

The property of the treatment effect estimation

Hazard ratio	PRIME EM		
e^{β}	Bias	ESE	95% CI
100p%=60%			
0.6	0.008	0.213	95.8%
0.3	0.004	0.231	96.6%
0	0.001	0.239	95.7%
100p%=30%			
0.6	0.052	0.453	95.9%
0.3	0.009	0.447	98.4%
0	0.068	0.492	95.3%

Note: Lag duration is 1 month. Total study duration is 5 years.

Enrollment rate is 16 subjects per month. Sample size $N=261$. $\alpha = .05$.

Abbreviations: CI, coverage interval; ESE, empirical standard error; EM, expectation maximization; PRIME, P%-responder information embedded.

TABLE 2

The impact of dichotomized response on study efficiency in term of the sample size required to achieve the target power (N)

Dichotomy 100p(%)	PRIME design		Conventional design	
	HR after t^* Responders vs nonresponders/Controls	N	Average HR Treatment vs Controls	N
20	0.01	265	0.74	455
30		117	0.63	189
40		74	0.54	109
50		49	0.46	73

Note: Response dichotomy is measured by the proportion of responders (100p%) among the treatment arm at baseline.

Target power is 80%. Lag duration is 2 months. Baseline hazard is 0.0737. Total study duration is 29 months.

Enrollment rate is 36.8 subjects/month. Type I error rate $\alpha = .05$.

Abbreviations: HR, hazard ratio; PRIME, P%-responder information embedded.

TABLE 3

The impact of misspecifying p on the performances of the PRIME and conventional strategies

PRIME strategy under $100p^m(\%)$							Conventional strategy under $100p^m(\%)$					
Design			Likelihood ratio test				Design		Log-rank test			
			$p^m > p^*$		$p^m < p^*$				$p^m > p^*$		$p^m < p^*$	
$100p^m(\%)$	HR	N	$p^*(\%)$	EP(%)	$p^*(\%)$	EP(%)	Average HR	N	$p^*(\%)$	EP(%)	$p^*(\%)$	EP(%)
20		265	10	28.7	30	98.4	0.74	455	10	31.1	30	98.7
30	0.01	117	20	47.3	40	94.0	0.63	189	20	53.8	40	96.9
40		74	30	60.4	50	95.0	0.54	109	30	63.6	50	95.7
50		49	40	63.0	60	91.2	0.46	73	40	69.1	60	95.5

Note: Target power is 80%. Lag duration is 2 months. Hazard ratio (HR) for responding patients is 0.01. Total study duration is 29 months. Enrollment rate is 36.8 subjects/month. Type I error rate $\alpha = .05$. Empirical power (EP) is evaluated based on 1000 replications. $p^*(\%)$: true baseline proportion of responders among the treated. $100p^m(\%)$: misspecified baseline proportion of responders among the treated.

Abbreviation: PRIME, P%-responder information embedded.