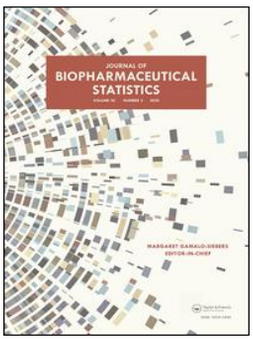

Journal of Biopharmaceutical Statistics

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

Statistical and practical considerations in designing of immuno-oncology trials

Pralay Mukhopadhyay , Wenmei Huang , Paul Metcalfe , Fredrik Öhrn , Mary Jenner & Andrew Stone

To cite this article: Pralay Mukhopadhyay , Wenmei Huang , Paul Metcalfe , Fredrik Öhrn , Mary Jenner & Andrew Stone (2020): Statistical and practical considerations in designing of immuno-oncology trials, Journal of Biopharmaceutical Statistics, DOI: 10.1080/10543406.2020.1815035

To link to this article: https://doi.org/10.1080/10543406.2020.1815035

Published online: 10 Sep 2020.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
https://doi.org/10.1080/10543406.2020.1815035

Taylor & Francis
Taylor & Francis Group

Statistical and practical considerations in designing of immuno-oncology trials

Pralay Mukhopadhyay[a], Wenmei Huang[b], Paul Metcalfe[c], Fredrik Öhrn[d], Mary Jenner[e], and Andrew Stone[f]

[a]Oncology Biometrics, AstraZeneca, Gaithersburg, USA; [b]Biostatistics, Moderna, Cambridge, USA; [c]Oncology Biometrics, AstraZeneca, Melbourn, UK; [d]Early Biometrics and Statistical Innovation, AstraZeneca, Mölndal, Sweden; [e]Qi Statistics, Kent, West Malling, UK; [f]Stone Biostatistics Limited, Cranage, Crewe, UK

ABSTRACT

The novel mechanism of action of immunotherapy agents, in treatment of various types of cancer, poses unique challenges during the designing of clinical trials. It is important to account for possibility of a delayed treatment effect and adjust sample size accordingly. This paper provides an analytical approach for computing sample size in the presence of a delayed effect using a piece-wise proportional hazards model. Failing to account for an anticipated treatment delay may result in considerable loss in power. The overall hazard ratio (HR), which now represents the average HR across the entire treatment period, can remain a meaningful measure of average benefit to patients in the trial. We show that, special consideration needs to be given for the designing of interim analyses related to futility, so as not to increase the probability of incorrectly stopping an effective agent. It is shown that the weighted log-rank test, using the Fleming-Harrington class of weights, can be used as supportive analysis to better reflect the impact of a delayed effect and possible long-term benefit in a subset of the overall population.

ARTICLE HISTORY
Received 21 August 2020
Accepted 22 August 2020

KEYWORDS
Delayed treatment effect; futility; interim analysis; piece-wise proportional hazards model; weighted log-rank test

1. Background

The recent development of immunotherapy agents in the treatment of advanced cancers and the promising efficacy results observed in various tumor types is changing the way patients are treated with these diseases. Immunotherapy utilizes the body's immune system to fight the cancer. In simple terms, a healthy immune system can detect and destroy foreign objects in our body, including cancer cells. However, the tumor can escape detection of the immune system by activating certain inhibitory pathways, such as the programmed cell death 1 (PD-1) or its ligand, (PD-L1), which prevents the cancer cells from being recognized as a foreign entity. Inhibition or blockade of pathways such as PD-1 allows the body's own immune system to detect and destroy cancer cells (Pardoll 2012). And therapeutic agents that target the PD-1 or PD-L1 pathways have proven to be very effective in treating several types of cancer (Borghaei et al. 2015; Brahmer et al. 2015; Rizvi et al. 2015).

The novel mechanism of action of these agents has also challenged researchers to some extent on the classical paradigm of study design, analysis and interpretation of clinical trials. A few characteristics of these agents, often reflected in the clinical data are (i) a delayed separation of the Kaplan-Meier (KM) curves questioning the assumption of proportional hazards (PH) in sizing of trials (Robert et al. 2015), (ii) a long and flat tail of the KM, reflecting a subset of patients deriving long term benefit with these agents, suggesting additional consideration needs to be given to follow-up as well as in data

CONTACT Pralay Mukhopadhyay Pralay.Mukhopadhyay@astrazeneca.com Head of Late Stage IO, Oncology Biometrics, AstraZeneca, Gaithersburg, MD 20878

analysis to understand this effect better, and (iii) the utility of RECIST (Eisenhauera et al. 2009) based endpoints, such as tumor response and progression free survival (PFS) as a good measure of clinical benefit compared to overall survival. In this article, we provide a statistician's perspective on these issues and discuss ways of designing and analyzing trials with immunotherapy agents. Our focus in this paper will be on issues (i) and (ii) only. The challenges with issue (iii) would be the subject of a separate paper.

While traditional methods for sizing of trials with time to event endpoints use the PH assumption, if a delayed effect is present, this may lead to considerable loss in power of the log-rank test (or equivalent Cox regression), if the delay is not considered during the design of the trial (Chen 2013, Lin 2020). Under PH, the hazard functions are assumed to be proportional over time, i.e., the relative hazard is assumed to be constant. Now consider a scenario where no treatment benefit is observed over a period $t \leq T$, i.e., the HR = 1 for this lag period. Following this lag period, for $t > T$, the HR (experimental: control) is assumed to be c (a constant < 1). Therefore, the HR throughout the entire follow-up period can be assumed to be piece-wise proportional (1 if $t \leq T$, and c if $t > T$), rather than just c for the entire follow-up period. Hence, in this example, if one were to estimate the average HR over the entire follow-up period, this will now be a value between c and 1. Thus if we size the trial assuming a HR of c, the study will be under powered.

Consider the same scenario with the treatment effect being piece-wise proportional. The overall average HR can now be seen to be decreasing over time instead of being constant; therefore, any benefit in treatment effect will only be reflected after a minimum amount of follow-up. While follow-up and sufficient maturity of data is critical in any study, the possibility of the HR decreasing over time will further highlight the importance of follow-up in trials with a possibly delayed treatment effect. For example, if the trial is analyzed prematurely, then the treatment estimate may be very close to 1 and will not reflect the benefit that would have been observed eventually. Therefore, careful consideration needs to be given in terms of timing of any analysis. This becomes even more important while planning for an interim analysis for early decision making, especially futility. For trials with a delayed effect, there is concern about the risk of early stopping to incorrectly accept the null hypothesis. It is then evident that use of conventional futility boundaries that ignore the delay would inflate the type II error, since the probability of crossing a futility boundary would be increased.

Another hallmark signature of immunotherapy agent is the long-term durability of response and disease stabilization in a subset of patients. Sometimes this is reflected in the tail of the KM curve which tends to become very flat. A good example will be the 10-year overall survival follow-up data in the Ipilimumab trial (Hodi et al. 2010). This further reiterates the need for adequate follow-up to capture the impact of the patients with long term benefit, on the overall treatment effect. In addition, it also raises the question, if the contribution of these long-term responders needs to be better reflected in our analysis. Since these long-term responders would have likely progressed or died in a matter of months instead of being alive without the disease for several years one can justify the need for capturing this benefit more appropriately in our analysis. If we now consider the traditional log-rank test, which only depends on the order of occurrence of the events rather than any consideration on the timing, and also puts the same weight on every event regardless of its rank order, the question becomes if we should consider analyses, such as using a weighted log-rank test, that will put more weight on late versus early events. This may in turn reflect better the contribution of these long-term responding patients on the overall treatment benefit.

There are literature dealing with clinical trials with delayed treatment effect. Zucker and Lakatos (1990) proposed two weighted log-rank types of statistics designed to have good efficiency over a wide range of lag functions, which can be applied in situations where a delayed effect is expected but cannot be specified precisely in advance. However, the authors do not provide an analytic approach for sample size and power calculation. Zhang and Quan (2009) investigated the asymptotic distribution of the two-sample log-rank test statistic under the lag-time model for analyses in both intent-to-treat (ITT) population and non-ITT population, under the assumption that patient accrual is a step function, and illustrated the calculation of asymptotic power. It is worth mentioning that in order

to perform the non-ITT analysis, the lag-time needs to be predefined so that patients discontinued prior to the lag-time will be treated as non-informative censoring. Xu et al. (2017) proposed a new weighted log-rank test, called the piecewise weighted log-rank test, and developed approaches for sample size and power calculation when there is a delayed treatment effect. More specifically, the power is determined by the events accumulated after the onset of treatment effect.

In this paper, we propose a sample size calculation using the piece-wise PH model under general assumptions on accrual and follow-up in the ITT population. Section 2 provides the details of the sample size computation using the piece-wise PH model under general assumptions on accrual and follow-up. Section 3 uses simulations to understand the impact of a delayed treatment effect on power of the test, by looking at different magnitude of delay (relative to the median survival time in the control arm). Section 4 evaluates through simulations, the operating characteristics (OC) of the statistical tests at the IA, in presence of a delayed treatment effect. We describe and compare the performance of two approaches to futility stopping that control type II error in the delayed response setting. Section 5 investigates the power of the un-weighted versus weighted log-rank test in the presence of a treatment delay, under non-PH alternatives. The Fleming-Harrington's G $^{(\rho,\ \tau)}$ class of weights is considered for this evaluation. The power of the tests with different sets of weights is evaluated under varying assumptions of treatment delay. Section 6 provides a discussion of these various findings and how it may impact on our current thinking on trial designs and data analysis. Overall conclusions are provided in Section 7.

2. Sample size computation in the presence of a delayed treatment effect

2.1. Sample size

Powering of a trial with a delayed treatment effect can easily be performed by applying standard sample-size approaches for time-to-event data. However, in contrast to proportional hazards, the power depends on both the proportion and number of events observed.

Throughout the paper we assume that there is a delayed treatment effect and specifically a piecewise proportional hazards model with an alternative hypothesis H_1 given by:
$HR_1 = 1$ *for* $t < T$,

$$HR_2 = x\ (<1)\ for\ t \geq T, \tag{1}$$

where T denotes the lag-time until there is a benefit of therapy. HR_1 and HR_2 the ratios of the hazard functions (experimental: control) before and after the lag respectively. It is straightforward to extend the approach to a more complicated piecewise model that had additional time periods and any value for the HR in that period; in this paper we concentrate on a simpler model to highlight some of the key design considerations.

The overall average HR on an interval (0, t^*), where $t^* > T$,

$$\int_0^{t^*} h_1(t)/h_2(t) d\ S_1(t) S_2(t)$$

as defined by Kalbfleisch and Prentice (1981) and Schemper (1992), can be simplified as (the derivation can be found in the Appendix):

$$\overline{HR} = \exp(p_1 \ln(HR_1) + p_2 \ln(HR_2)) \tag{2}$$

where p_1 and p_2 are the proportion of events observed during the time interval (0, t^*) before and after the lag-time, h_1 and h_2 are the hazard functions, and S_1 and S_2 are the survival functions, respectively.

So, in the case of model (1)

$$\overline{HR} = \exp(p_2 \ln(HR_2)) \tag{3}$$

The average HR is an important concept in the interpretation of trial results with a delayed treatment effect and is a meaningful measure representing the average benefit over the period of observation.

Having determined $\overline{HR}$ for a given T, HR_1, HR_2, and N (the total number of patients recruited), standard sample size approaches (Schoenfeld 1983) can be used so that:

$$e = \frac{(1+r)^2}{r} * \frac{[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)]^2}{ln^2(\overline{HR})} \tag{4}$$

where e is the total number of events, α is the two-sided type 1 error and 1-β the power, and r the ratio of patients randomized in the experimental arm compared to control.

2.2. Estimation of average hazard ratio for a given follow-up

In order to estimate the expected average hazard ratio the proportion of events expected in each trial period first needs calculating. The calculation is complicated by the fact that patients are not recruited instantaneously.

We first assume that events follow a piecewise weibull distribution with:

$$S(t) = \begin{bmatrix} \exp(-\lambda_1 t^\gamma) & t < T \\ \exp(-\lambda_1 T^\gamma - \lambda_2 (t^\gamma - T^\gamma)) & t \geq T \end{bmatrix} \tag{5}$$

where λ_1 and λ_2 are the scale parameters before and after T respectively and γ the shape parameter, noting that a piecewise exponential distribution is a special case with γequal to 1. In practice a common value of λ_1 and λ_2 will be assumed for the control arm and derived from historical data with $\lambda = (\ln(2)/m)^{1/\gamma}$ and median m. Whereas for the experimental arm, a long-term survival probability can be used to estimate λ_2 assuming the same λ_1 and shape parameter,γ as the control arm.

It is also necessary to create a p.d.f., g(s), for the recruitment time and associated c.d.f. G(s). Here we assume that the recruitment time could be nonuniform and is given by Carroll (2009):

$$g(s) = \frac{ks^{k-1}}{B^k}; G(s) = \frac{s^k}{B^k} \tag{6}$$

where B is the total recruitment time and k (>0) a measure of nonuniformity, with k = 1 corresponding to uniform recruitment and k = 2 often representing a more realistic rate of recruitment.

The proportion of events occurring by time t can then be estimated as follows:

$$\begin{aligned} p(event\ by\ time\ t) &= \int_0^{\min(t,\ B)} g(s)(1 - S(t-s))ds \\ &= G(\min(t,B)) - \int_0^{\min(t,\ B)} g(s)S(t-s)\ ds \\ &= \left(\frac{\min(t,B)}{B}\right)^k - \frac{k}{B^k}\int_0^{\min(t,\ B)} s^{k-1}S(t-s)\ ds \end{aligned} \tag{7}$$

substituting the Weibull survival function from (5). For a specified maximum follow-up time, the number, and hence the proportion of events in each time-period can then be calculated by numerically integrating (7) using standard quadrature methods. Closed form solutions exist for a piecewise exponential distribution.

It is worth noting that, unlike proportional hazards, when patients are censored will have a bearing on the average HR and hence the power of the study. In particular, even if there is complete follow-up

of all patients so only administrative censoring, if the minimum follow-up time is less than T, some patients will be censored before T. In this case, the proportion of events observed before T, and consequently the average HR and power will be different to another study, with the same piecewise HRs and total events, which has a different duration of follow-up.

2.3. Associated measures

When designing a trial, it is helpful to also provide estimates of other associated measures to allow other researchers to gauge the extent of clinical benefit associated with the piecewise proportional hazards model. Additionally, it is informative to calculate and present the smallest treatment effect that would be statistically significant, often referred to as the critical value. For survival data it is possible to define the critical value for the HR exactly if the trial is analyzed when the observed number of events closely matches the number assumed in the design.

The variance of $\ln(\overline{HR})$ is given by $\frac{(1+r)^2}{re}$, therefore the critical value for$\overline{HR}$, which corresponds to the hazard ratio which has the upper limit (UL) of its confidence interval equal to 1, i.e.,

$$\ln(UL) = \ln\left(\overline{HR}_{critical}\right) - \Phi^{-1}(\alpha/2) \times \sqrt{\frac{(1+r)^2}{re}} = 0,$$

which gives us the critical value as below:

$$\exp\left[\sqrt{\frac{(1+r)^2}{re}}\Phi^{-1}(\alpha/2)\right)] \tag{8}$$

It is also possible to calculate the critical hazard ratio, HR_{2crit}, in the second time-period using an iterative approach. An initial estimate of HR_{2crit} is made by re-arranging (3) to give

This initial estimate is used to derive an updated survival function and second time-period scale parameter$\lambda_{2e} = \lambda_{2c}.HR_{2crit1}$ for the experimental arm and (7) used to find updated estimate of p_2 and hence a new value for derived HR_{2crit}. The process is continued until the value of HR_{2crit} remains sufficiently constant.

A commonly used statistic is the median survival within each treatment arm. The value associated with the alternative hypothesis can be calculated by:

$$\left(\frac{\ln(2)}{\lambda_1}\right)^{1/\gamma} \; if \; S(T) \leq 0.5$$

$$\left(\left(\frac{\ln(2) - \lambda_1 T^{\gamma}}{\lambda_2}\right) + T^{\gamma}\right)^{1/\gamma} \; if \; S(T) > 0.5$$

For the control group λ_{1c} and λ_{2c}would be set to the values assumed when estimating the duration of the trial. For the experimental group, $\lambda_{1e} = \lambda_{1c}$and $\lambda_{2e} = \lambda_{2c}$. HR_2.

The median values expected when the average HR corresponds to statistical significance can be calculated in the same manner but deriving the scale parameters using HR_{2crit}.

Whilst it is informative to present the expected medians, this measure can be particularly misleading with a delayed treatment effect as it will tend to underestimate the overall benefit. Therefore, it can be helpful to additionally present the expected survival probabilities associated with longer term survival and these can be derived directly from (5). Those associated with the critical value would use $\lambda_{2e} = \lambda_{2c}.HR_{2crit}$. Indeed, these survival probabilities can be used as means to calculate HR_2 to size the trial based on an expected T and clinically meaningful difference in long term survival.

3. Impact of delayed treatment effect on power of study

We now investigate powering of trials in the presence of a delayed treatment effect. Consider a two-arm trial with 1:1 randomization, where the primary objective is to compare overall survival (OS) between the two arms. For now, we will assume PH, and that the OS time in the control and experimental arms follow an exponential distribution with a median of 7 and 14 months respectively. Therefore, the target HR will be 0.5. This trial will require 91 deaths to achieve 90% power, with a critical HR for statistical significance of 0.66, using a two-sided 0.05 level log-rank test. If we further assume a uniform accrual of 20 patients per month over 6.5 months, then the trial can be conducted with 130 patients and will require 20.9 months (6.5 months accrual time and 14.4 months follow-up) to achieve the target number of events (70% maturity).

Now using (1), let us assume that x = 0.5 and T = 2. Therefore, we are assuming that the experimental arm follows a piecewise exponential distribution and the HR is no longer assumed to be constant over the entire study period (Figure 1); The average HR over the 20.9 months follow-up period can be estimated to be 0.6, instead of 0.5 using the methods provided in Section 2. Had the trial been conducted with the planned 91 death events, the power of the test would now be 69% instead of 90%. Therefore, to maintain the same power, we will require either a larger sample size and/or a longer duration of follow-up. One option would be to conduct the trial with 240 patients. If we enroll these patients in 12 months (uniform accrual of 20 patients per month), and follow them for another 11 months, (overall study duration of 23 months), we would expect to observe 168 deaths from 240 patients (70% maturity). The study will now achieve 90% power to detect an overall HR of 0.6, with a critical HR for statistical significance of 0.74, instead of 0.66.

Figure 2 below shows the potential loss in power, under different assumptions of a delayed effect, ranging from no delay to a delay of up to 6 months, before observing any treatment benefit. For the purpose of illustration, we assume two scenarios, with the treatment effect after the delay being a HR of 0.5 and 0.625 respectively. As expected, there is loss in power, sometimes considerable, with increased

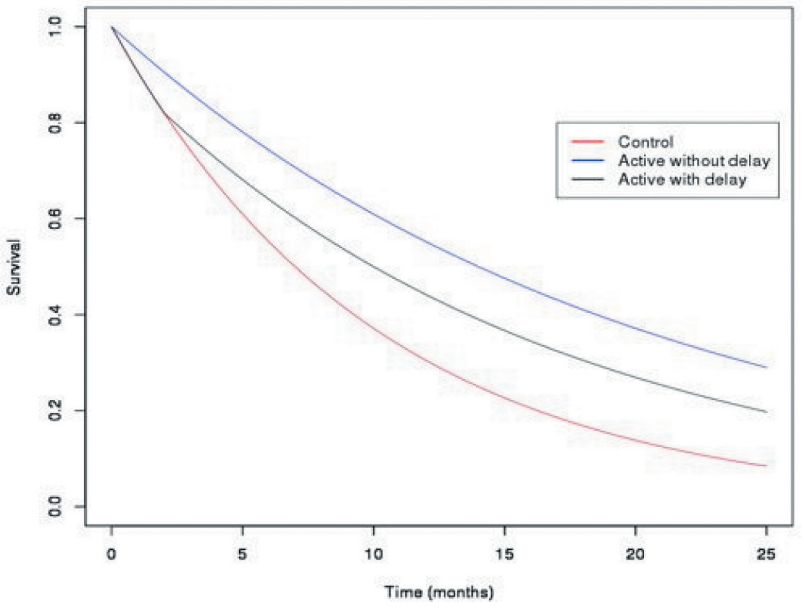

Figure 1. Hypothetical survival curves showing delayed versus immediate separation. This figure shows hypothetical survival curves, under the scenario of an immediate separation (Blue vs. Red curve, with an assumed HR of 0.5) or a delayed separation after 2 months (Black vs. Red curve, with an assumed HR of 0.5 after the separation). In the case of a delayed separation, the overall HR after 23 months of follow-up is estimated to be 0.6.

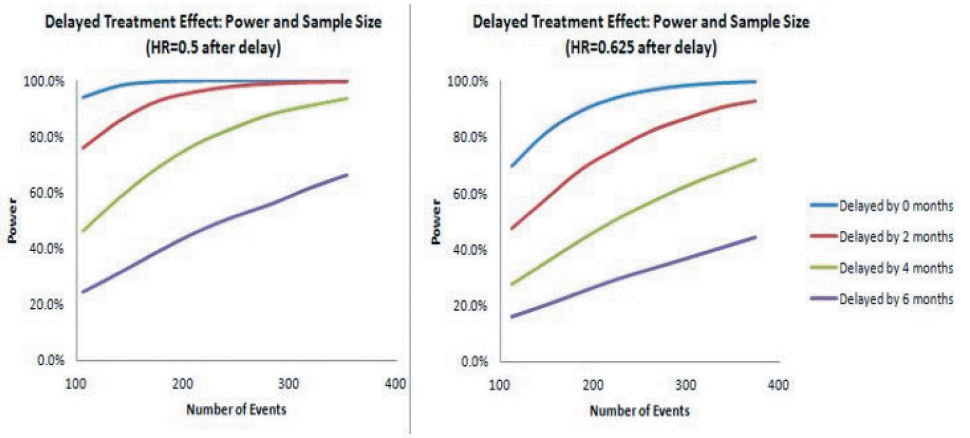

Figure 2. Impact of treatment delay on power of the study. This figure shows the power of the test (y-axis) corresponding to a specific number of events (x-axis) under different assumptions of treatment delay. The median in the control arm is assumed to be 7 months, and the corresponding HR after the delay is assumed to be 0.5 and 0.625 in the two examples. Under each scenario (example 100 events), the sample size, the required number of events, and accrual time (15 months) are fixed, keeping the maturity level at 71%, while varying the duration of follow-up to obtain the target events, with different assumptions of delay.

amount of treatment delay. The required number of events to compensate for the loss in power is considerably high as well.

Another useful observation to note is that once the average HR is determined using (3) (under the piece-wise PH assumption), the required number of events to achieve a specific power follows standard sample size calculations using (4). Therefore, the power of the test and the required number of events will be the same as under the scenario of no delayed separation, as long as the target HR and the level of significance of the two tests are the same

With PH, for the same number of events, the power is identical regardless of the number of patients recruited. This is no longer the case with a delayed effect and the same power can be achieved with a different number of events. For example, consider a trial where the HR follows (1) with x = 0.625 and T = 2, the control (treatment) arms follow an exponential (piece-wise) distribution. Table 1 show that if 480 patients are enrolled in the first 16 months (30 patients per month uniform accrual) and followed for 8.3 months, then the average HR will be 0.704 and 342 events will be observed after 24.3 months. One can then achieve 90% power to detect an average HR of 0.704. However, if we choose to enroll a larger number of patients, so that the same number of events is obtained in a shorter time period, this will result in loss of power. Note the average HR which now increases from 0.704 (study duration of 24.3 months) to 0.725 (study duration of 21.5 months). Note that, one can still achieve 90% power by conducting the trial with fewer events, and a smaller sample size, but requiring a higher level of maturity of the data. Consider a trial with a sample size of 390 patients, but with study duration of 28 months (13 months enrollment and 15 months of follow-up). This will now result in 316 events (81% maturity). Since the average HR now reduces to 0.694, the study can achieve 90%

Table 1. Sample size, follow-up and impact on power of the test (Simulation runs = 5000).

No. of Patients	Accrual (Months)	Follow-up (Months)	No. of Events (maturity)	Average HR	Estimated Study Duration (Months)	Power
480	16	8.3	342(71%)	0.704	24.3	90%
510	17	6.0	342(67%)	0.709	23.0	88%
540	18	4.3	342(63%)	0.715	22.3	87%
570	19	2.8	342(60%)	0.719	21.8	86%
600	20	1.5	342(57%)	0.725	21.5	84%

Randomization ratio = 1:1; fixed number of events; delayed separation after 2 months; HR = 0.625 after separation; uniform accrual (30 patients per month) with varying trial size and minimum follow-up; median OS (control) = 7 months; 2-sided type I error = 0.05.

power with only 316 instead of 342 events, but the trial would take 3.7 months longer than a 480-patient trial with the same power.

4. Interim analysis in the presence of a delayed treatment effect

It is very common to introduce interim analyses (IA) in trials with an objective of either stopping early for overwhelming benefit (superiority) or for lack of benefit (futility). In any trial, requirement for adequate follow-up before performing an IA is critical. However, as highlighted in the previous section, follow-up and maturity of the data will be particularly important where a delay is expected before the treatment starts to show effectiveness, especially to avoid falsely stopping a trial for an effective agent. We investigate appropriate timings for conducting IAs for futility, using simulations. For simplicity, we will consider situations that are commonly expected in practice to evaluate how the operating characteristics (OC) changes when there is a delayed effect present.

4.1. Simulation parameters

We will consider a two-arm trial with 1:1 randomization; uniform accrual of 30 patients per month; target HR of 0.625; Median OS in the control arm of 7 months. In order to highlight the increased probability of making incorrect decision, the trial has been powered under the PH assumption. This study would require 192 events (71% maturity) to achieve 90% power to detect a HR of 0.625, using a 2-sided 0.05 level test, without adjustment for any planned IA. The trial will be conducted with 274 patients. Therefore, it will take 9.1 months to accrue these patients and another 11.6 months of follow-up to achieve the necessary events. The target maturity of the data at the time of the final analysis is 192/274 (70%).

Now suppose we introduce a single IA for futility, at the design stage, either after 50% of the target number of events or 80% of the target number of events are observed. We will now evaluate the impact of a delayed treatment effect of 2 months, which was not accounted for during the study design stage in this trial. For illustration purpose, we will use the Lan-DeMets beta spending function that approximates the O'Brien Fleming boundary (Lan and DeMets 1983) to compute the stopping boundary for futility at the IA, assuming a true HR of 0.625 to calculate the type II error spent at the interim. O'Brien Fleming boundary is chosen as one may want to be cautious when stopping a trial for futility. Other spending functions can be utilized as appropriate to achieve the operating characteristics that are desirable for a given trial, but a similar overall conclusion should be reached. This trial will now require 194 events (if the IA is introduced at 50% of the events) or 204 events (if the IA is introduced after 80% of the events) to maintain overall power at 90% allowing for early rejection of H1. Therefore, the IA will be conducted after 97 out of 194 target events (164 events out of 204 target events) are observed.

We will also consider conducting the IA using a slightly different strategy. Using IA at 50% target events as an example, since the follow-up time and maturity of events is critical, instead of conducting the IA at 50% of the target events, we will consider conducting the analysis on the first 137 (50%) patients, but with a similar maturity as in the final analysis, i.e., with 97 events (71% maturity) on the first 137 patients, where the same stopping boundary computed using 50% of target events in the overall population will be used, provided that the interim analysis will be conducted based on same number of events. Since the objective is to stop for futility, we will compare the probability of stopping incorrectly to accept the null hypothesis at the interim, i.e., the false negative rate (FNR). Results provided in Tables 2 and 3 are based on 5000 simulation runs.

Based on Table 2, we can see that if the original PH assumptions (no treatment delay) do not hold true, there is an increased risk of deeming the new agent as ineffective early on and incorrectly stopping the trial. That risk is reduced if this analysis is conducted on a subset of patients, but with similar maturity as in the final analysis.

Table 2. Comparison of FNR of futility analyses under different assumptions of treatment delay (Simulation runs = 5000).

Length of Delay (Months)	Target events at FA, % of target events at IA	FNR at IA (overall type II error) [§]	FNR at IA (overall type II error) [i]
0	194, 50%	1.7% (10.5%)	1.9% (10.2%)
2	194, 50%	10.9% (35.5%)	5.2% (32.8%)
0	204, 80%	7.4% (10.8%)	N/A*
2	204, 80%	32.1% (38.1%)	N/A*

[§]IA analysis conducted after planned target events reached.
[i]IA conducted on a subset (first 50% of enrolled patients) after planned target events reached.
* N/A = Not applicable, as it would now require 163 (80% of target) events from 137 patients.
For illustration purpose, assuming delayed separation after 2 months; HR = 0.625 after separation; uniform accrual (30 patients per month). The futility boundary for the interim analysis was calculated using a Lan-DeMets beta spending function that approximates the O'Brien Fleming boundary; A critical HR of 0.948 (0.792) was used for IA conducted after 50% (80%) of target events. Maturity is around 71% (75%) at final analysis.

Table 3. Comparison of FNR of futility analyses at IA when treatment delay is accounted for in sample size calculation (Simulation runs = 5000).

Length of Delay (Months)	Target events at FA, % of target events at IA	FNR at IA (overall type II error) [§]	FNR at IA (overall type II error) [i]
2	346, 50%	4.8% (12.8%)	1.8% (11.2%)
2	363, 80%	10.0% (12.9%)	N/A*

[§]IA analysis conducted after planned target events reached.
[i]IA conducted on a subset (first 50% of enrolled patients) after planned target events reached.
* N/A = Not applicable, as it would now require 275 (80% of target) events from 241 patients.
Sample size was calculated with consideration of delayed treatment effect. If no IA is planned, 344 events (482 patients) are needed to achieve at least 90% power at final analysis (FA) when there is a 2-month delay and assumed HR is 0.625 after the delay; overall HR is estimated to be 0.703 after 8.3 months of follow-up. The study was then adjusted for one interim analysis for futility using the Lan-DeMets beta spending function that approximates the O'Brien Fleming boundary, assuming a true fixed HR of 0.703; A critical HR of 0.961 (0.839) was used for IA at 50% (80%) target events, Maturity is around 72% (75%) at final analysis.

However, in our example, the approximate timing when 97 (50% of the 194 target events) were observed from all randomized patients was about 10 months, while it took about 18.5 months to observe the same number of events from the first 137 patients. That timing is close to when we expect to observe 92% of the target events, from all randomized patients. Under a different set of assumptions, this time difference could be even longer. Therefore, it is not surprising that there is a higher risk of a false conclusion if the futility analysis is conducted early in the trial, when we expect a delayed treatment effect. This can be mitigated by requiring sufficient follow-up and maturity. However, it is acknowledged that the practical savings from conducting the futility analysis very late in the trial (say after greater than 80% of the target events) may be limited.

It is intuitive, that the FNR will be more manageable, if the trial was originally powered assuming a 2-month delay (Table 3). If no IA is planned, the trial will now require 344 events from 482 patients (71% maturity) to achieve 90% power. Assuming an accrual rate of 30 patients per month, it will take 16 months to enroll and another 8.3 months of follow-up to achieve the target events. The overall HR over the entire treatment period will be 0.703. In order to maintain the overall power, adjusting for one IA for futility, conducted after 50% (80%) of events, we recomputed the required number of events using the Lan-DeMets (O'Brien Fleming) beta spending function, assuming a true fixed HR of 0.703. This trial will now require 346 (IA at 50% of target) events or 363 events (IA at 80% of target) events. However, note that the overall power of the study is still approximately 87% (87%) instead of 90%, as would have been expected with this adjustment.

The reduction in overall power from 90% to 87% is because we are still using unadjusted futility boundary that does not take into consideration, a delayed treatment effect and the HR not being constant over time. Again, if the IA is conducted using the alternative strategy, i.e., using the same number of 173 (50%) events, but obtained from the first 241 (50%) patients, then the overall power will

be approximately 89%, with an FNR at the IA of 1.8% instead of 4.8%. In this case, the timing of the IA using a subset of patients is approximately same as the occurrence of 84% of target events in all patients. However, the FNR using all the events is about 10%, instead of 1.8%. It is worth noting that, the probability of early termination, under the null hypothesis of no treatment benefit is approximately 60% if the IA is conducted with 50% of events (or on events from the first 50% patients enrolled) while it is 93% if the IA is conducted after observing 80% of the events. This tradeoff should be considered while making appropriate choice of futility analysis in the study.

In order to strike a reasonable balance between the risk of early stopping to incorrectly accept the null hypothesis and scheduling the futility analysis very late in the trial, an alternative is to use the full dataset but account for the lower treatment effect at the interim analysis. This can be accomplished using a beta spending approach similar to the method by Pampallona et al. (2001) and will be the subject of a separate publication. The Pampallona et al. method is flexible in the sense that it provides a very general class of boundaries, that can be made suitably aggressive or cautious depending on what is desirable in the context of a specific trial.

5. Analysis to account for possible delayed treatment effect

During the design stage, if there is a potential of a delayed treatment effect, it is recommended to take the possible delay into consideration by calculating the sample size as outlined in Section 2. However, if there is an anticipation of a delayed treatment effect, one option could be to account for it in the analysis by putting more weight on the later events.

5.1. Weighted log-rank test

Under the PH assumption, the log-rank test (LRT) is the most powerful option for comparing survival distributions. However, when the proportional hazards assumption is violated, log-rank tests may not have the maximal power in the class of all linear rank tests. A good alternative is the weighted log-rank test (WLRT) using Fleming and Harrington's $G^{\rho,\tau}$ class of weights. It can be demonstrated that WLR tests provides greater power than LRT when effects of treatment are delayed.

Assume that deaths are observed at times $t_1 < t_2 < \ldots < t_d$ and that the number of deaths at time t_i is d_i $(=d_{1i} + d_{2i})$of a total number at risk at that time of $n_i(= n_{1i} + n_{2i})$.

The weighted log-rank statistics, with weight function $W(t)$ is

$$\sum_{i=1}^{D} W(t_i)(d_{1i} - E(d_{1i}))/\sqrt{\sum_{i=1}^{D} W^2(t_i)Var(d_{1i})},$$

where $E(d_{1i}) = n_{1i} \times (d_{1i}/n_i)$ and $Var(d_{1i}) = n_{1i}n_{2i}d_i(n_i - d_i)/n_i^2(n_i - 1)$ under the null hypothesis (Hasegawa 2014). This statistic follows the standard normal distribution. The $G^{\rho,\tau}$ class of weighted log-rank tests was proposed by Fleming and Harrington (1991), with a weight function equal to

$W(T_i) = \left(\hat{S}(t_i)\right)^{\rho}\left(1 - \hat{S}(t_i)\right)$ for $\rho > 0, \tau > 0$,

where $\hat{S}(t_i)$is the Kaplan-Meier estimate of the survival function in the pooled sample at time t_i.

Figure 3 Weight functions in the Fleming-Harrington's $G^{\rho,\tau}$ class. Weights are uniform in $G^{0,0}$ and emphasize early, middle, and late differences, respectively, in $G^{1,0}$, $G^{1,1}$ and, $G^{0,1}$.

When $\rho = 0, \tau = 0$, $G^{0,0}$ corresponds to the standard log-rank; when $\rho = 1, \tau = 0$, $G^{1,0}$ corresponds to the Prentice statistics (Prentice 1978). Letting $\rho = 0$ and $\tau = 1$ would place more weight on late events, emphasizing late differences in the hazard rates and/or the survival curves.

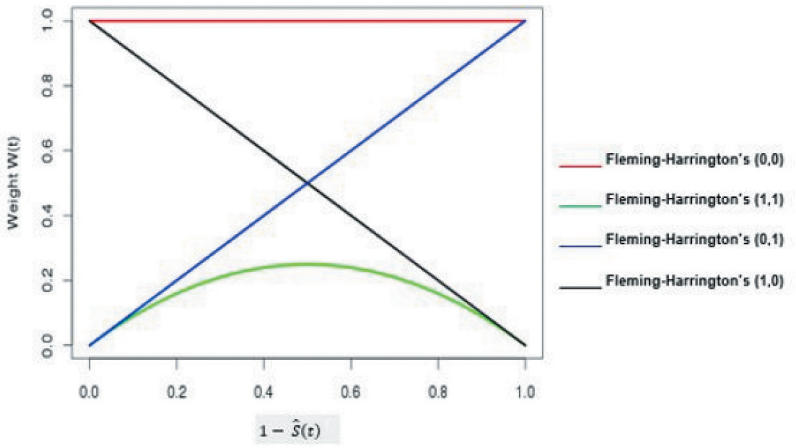

Figure 3. The weight functions used in the Fleming-Harrington's $G^{\rho,\tau}$ class.

5.2. Performance of WLRT

The performance of the WLRT is evaluated through simulations. WLRT with the Fleming-Harrington class of weights can be easily performed using the LIFETEST procedure in *SAS* 9.1 or later (SAS/STAT® 9.1 User's Guide 2004) or the FHtest package in R (Oller and Langohr 2012). The performance of various WLR tests can be evaluated using simulations. The results in Table 4 are based on 5000 simulation runs.

As an example, in a double-blind, randomized, parallel-controlled trial, 266 patients were randomly assigned in a 1:1 ratio to receive the experimental or control therapy, where the HR follows (1) with x = 0.625 for different T (=0, 2, 4 and 6-month, respectively). A uniform 15-month accrual period was assumed. The primary endpoint is OS and the median survival is assumed to be 7 months in the control arm. The analysis of OS is conducted using a two-sided 5% test, once 193 OS events are achieved. Table 4 shows the percentage of times the null hypothesis is rejected under different weights for various length of delay, T.

As expected, the standard LRT is the most powerful test when T = 0 because the PH assumption holds when there is no delayed the treatment effect. However, when there is a delayed treatment effect, LRT may not remain the most powerful test. In this particular example, when the delay is moderate (2 months), $G^{1,1}$ performs better than the other tests; while with longer delays (≥ 4 months), $G^{0,1}$ performs better than the other tests. It is not surprising that the $G^{1,0}$ test does not perform well in the presence of a delay, since more emphasis is given to the early versus late events.

6. Discussion

The goal of this paper was to highlight some of the issues one needs to consider during the designing of trials with immunotherapies for treating different types of cancer. As observed from data in recently published clinical trials, two hallmark features of these class of agents, are (i) a delayed treatment effect, generally seen as a delayed separation of the KM curve and (ii) the long-term durability of responses in

Table 4. Percentage of times rejecting null hypothesis based on 5,000 runs of simulation (%) (Simulation runs = 5000).

WLRT	Under H_0	No Delay	2-Month Delay	4-Month Delay	6-Month Delay
$G^{0,0}$	4.8	89.9	67.5	43.3	23.5
$G^{1,1}$	4.9	85.9	78.1	55.2	29.8
$G^{0,1}$	5.5	79.4	74.7	60.5	41.2
$G^{1,0}$	5.4	85.8	50.9	24.5	12.1

patients that may result in a long flat tail of the KM curve. It is important to note that it is not always the case that these features are observed with the clinical data. In fact, there are also examples of trials, see Brahmer et al. (2015) and Motzer et al. (2015), where there was no evidence of a delay. However, with the given uncertainty, building in these assumptions during the design phase may protect the trial from being under powered, in case the initial assumptions are incorrect. A detailed discussion on impact of power of the log-rank test under various non-PH alternatives has been discussed using simulations and case studies and a combination of weighted log-rank test (called the MaxCombo test) has been proposed when the type of deviation from the PH assumption (e.g., delayed separation versus crossing survival curves) is unknown (Lin et al. 2020).

We have provided an analytical approach for computing the sample size in the presence of a delayed treatment effect, under general assumptions on patient accrual and the underlying survival distribution. We have assumed nonuniform accrual (of which the simple uniform accrual is a special case) and that the survival time follow a Weibull distribution (of which, the commonly used exponential distribution is a special case). Traditionally medians are supplied by physicians and translated by statisticians into a HR, making various assumptions, in order to design a trial; we have presented an approach where only long-term survival probabilities need to be supplied and trials can be designed along with an assumption regarding the time delay. These approaches were then used to evaluate the possible loss in power, in the presence of a delay, if that is not accounted for during the design stage. We hope this presentation of a more general framework will help others design trials in a wide variety of settings. Another possibility to consider is the crossing of the survival curves, where the treatment effect may be in favor of the control arm for a certain period of time (HR > 1 prior to the lag) before switching directions. The sample size framework presented in this paper can be adapted to consider this situation. However, if this is anticipated, then the clinical significance of such a scenario, where a group of patients may actually perform worse with the experimental agent, needs to be given careful consideration before embarking in such a trial. We acknowledge, one of the limitations of our proposed method is it requires some amount of knowledge of the lag time during the design stage. This may often be obtained from available data. However when there is limited data but NPH is anticipated based on the mechanism of action of the molecule, a minimum amount of lag time can be assumed such that the trial is still adequately powered for an overall treatment effect (HR) that is clinically relevant. If the actual delay happens to be longer than what was assumed, there will still be loss in power. However, in that situation one could argue that the treatment benefit in the overall population is less likely to be clinically relevant and further exploration on subgroups with greater benefit may be necessary. If the delay happens to be smaller than originally assumed, the trial can be over powered. However, this will still be a better outcome than having a negative study. When there are such uncertainties, interim analyses for efficacy can be introduced to mitigate the risk of prolonging the trial due to increased sample size and/or follow-up. We have also highlighted the importance of follow-up and maturity of the data, if the PH assumption does not hold, and the HR estimate is expected to improve over time. Therefore, careful consideration should be given to the timing of the analysis. With PH for the same number of events the power is identical regardless of the number of patients recruited, this is no longer the case with a delayed effect and the same power can be achieved with a different number of events which provides for a greater number of design options. Another option that may be considered when there is uncertainty in the lag time during the trial design period, is to obtain a better estimate based on the actual data, and refining the assumptions on the proportion of events, p_1 and p_2 observed before and after the lag-time that will contribute to the overall HR. If it appears that the initial assumptions during the trial design period underestimated the lag-time, then the follow-up time can be adjusted to ensure enough events are contributing after the lag, in the estimation of the overall HR, to ensure that the power of the study is maintained. Such an approach will be similar in nature to an adaptive sample size re-estimation but may have more applicability in the immune-oncology space given the novel mechanism of action and the evolving understanding of both the timing and reason for the treatment lag in different tumor

types. Also, this will very likely involve an independent data monitoring committee having to make these decisions, based on a pre-specified set of rules. Therefore, like other adaptive trials, this will require careful planning prior to implementation.

We illustrate through simulations, the risk of conducting an IA early in the trial. In particular, the probability of the test to cross the superiority boundary can be low, while the chance for incorrectly declaring the trial as negative, can be higher. Therefore, the practical benefits of an early IA for either futility or superiority analysis can be questionable, and it is recommended that an IA is planned using a sufficiently mature dataset (e.g., 80% of the target number of events).

A few additional points to consider, if we believe the underlying PH assumption may not hold in practice. First, is the HR estimate, still a valid measure of clinical benefit, either from a statistical or a clinical perspective? From a statistical perspective, the Cox model is the most powerful test when hazard functions are proportional over the entire follow-up period. From that point, the use of the HR estimate as a valid measure has been questioned (Uno et al. 2014). However, if the same tie handling approach is used, the score test arising from the Cox model is mathematically identical to the ordinary LRT. Therefore, any inference made using the Cox model and the log-rank test is generally identical or near identical.

Can we address this theoretical concern in the following way? Let us assume that the hazard functions are piece-wise proportional at two or even more distinct time periods. The HR estimates derived using the Cox model should still be valid during these two periods. One can then define an overall measure of clinical benefit as the average HR, which is the weighted average of these estimates, weighted by the proportion of events occurring in each time period. We can then make inference based on this weighted test statistic which can be regarded as a measure of average benefit to the patient, as discussed in Section 2 of the paper.

From a clinical perspective, the meaningfulness of the average HR would generally depend on the observed data. In many cases, we believe, the HR is still a useful measure that effectively communicates the relative average benefit of receiving one therapy over the other. Exceptions always apply, depending on the shapes of the KM curves, amount of delay, whether there is a crossing of the curves, that may suggest presence of a quantitative or qualitative interaction between treatment and some other predictive factor, etc. (Mok et al. 2009). One could use standard techniques, (such as a Cox model with a time dependent covariate, Schoenfeld residuals), to evaluate the validity of the PH assumption. However, violation of the PH assumption by itself may not undermine the value of the overall HR estimate. It may then need to be supplemented with additional information, such as looking at piecewise HR estimates or consideration of other supplemental measures, such as the restricted mean survival time (RMST) (Huang and Kuan 2017; Royston and Parmar 2013). Although, RMST does not require an underlying assumption of proportional hazards and has a relatively easy interpretation as a summary measure, under NPH it suffers from similar loss in power as the LRT (Lin et al. 2020). Other drawbacks of RMST includes sensitivity of the measure to censoring at the tail of the KM curve and the magnitude of benefit (either in terms of difference or ratio) may be less well understood by clinical practitioners compared to the HR. Never the less, when NPH is observed, both RMST and milestone survival estimates can provide valuable information which helps with the holistic understanding of the trial results.

Another point worth mentioning is that, since the HR estimate only depends on the rank order of the events, and not on the actual timing of occurrence, two very different shapes of the KM curves may provide the same HR estimate. If we anticipate patients remaining disease free over a very long period or potentially getting cured, it may be also important to capture the timing of the event as well. From that point, one can look at other measures of clinical benefit, such as the RMST or milestone survival estimates to supplement the HR estimate or performing a parametric analysis in such a case. However, this aspect may be the topic of a separate discussion and not necessarily related to the discussion around validity of the HR in the presence of a treatment delay.

The second question is how much of a delay should be assumed during the sizing of the trial. One way to address this may be to consider that the fundamental objectives, when designing these trials

have not changed. Our goal is to still size these trials to be able to detect clinically meaningful differences in treatment effect, if one exists, and not just power the study to achieve statistical significance. As discussed in Section 3, the power of the trial when analyzed using a Cox model or log-rank test will depend on the average HR. Assuming a very long delay followed by a modest benefit will result in an average HR estimate that may no longer be clinically relevant. For example, consider in a late stage disease with no available treatment option, 25% of the patients treated with the new agent achieves long term remission while the remaining 75% does no better than best supportive care (BSC). Will it still be justified to approve such an agent in an all-comers population? The answer may be yes, provided (i) there is no way yet, to prospectively identify the patients who are going to benefit and (ii) the remaining 75% of patients are doing no worse than they would have in the BSC arm and (iii) the toxicity profile is acceptable. In this example, such a dichotomous population may result in a delayed separation of the KM curve. However, the outstanding benefit in one quarter of the study population may justify the risk of approving the new agent in a wider population. But if the benefit is only modest in those 25% of patients, and does not bring long term remission, one would then question the value of such an agent getting approved. However, if there was sufficient follow-up that revealed evidence of long-term cure amongst those 25% this could substantially alter the overall assessment of benefit/risk. Therefore, during the design stage, one still needs to be conscious of the average HR being targeted and ensure that the effect size is clinically relevant even in the presence of a delay.

Another point is should we be considering the anticipated delay in the analysis? From a statistical perspective, the ordinary LRT is the most powerful nonparametric test to detect PH alternatives. But that is not the case, under the presence of a treatment delay. In Section 5, we show that if a delay is observed, then the WLRT using the Fleming-Harrington class of weights performs better than the LRT. However, one of the drawbacks for the WLRT is the selection of weights, which would depend on the type of NPH expected to be observed from the data, e.g., delayed separation versus crossing of the KM curves. If there is considerable uncertainty in the type of NPH expected, one can also use a combination of weighted LRTs such as the MaxCombo test, which will be relatively agnostic to the type of NPH alternative observed and can provide more robust power (Lin, 2020). The motivation to use the WLRT should not be driven solely by increasing our chances of observing statistical significance or being able to reduce the sample size of the study as the use of an unequally weighted analysis implicitly assumes that it is more important to delay life in some patients compared to others. However, in diseases with significant unmet need, if there are patients, (with otherwise very short life expectancy) who can survive for years, the WLRT may better help in capturing the impact of treatment benefit for these patients. Furthermore, if patients are followed for a significant time cure rate models could be explored to assess whether there is a group of patients cured by therapy (Farewell 1982; Maller and Zhou 1996). Therefore, such tests can be used as additional sensitivity analyses to support the primary conclusions, provided they are clearly pre-specified in the statistical analysis plan.

7. Conclusion

The novel mechanism of action and different response kinetics of immunotherapy, in the treatment of various types of cancer, poses some unique challenges during the designing of such trials. In particular, it is important to account for the possibility of a delayed treatment effect and adjust the sample size accordingly. This paper provides an analytical approach for computing the sample size in the presence of a delayed treatment effect using a piece-wise PH model. Failing to account for an anticipated treatment delay may result in considerable loss in power. The overall HR, which now represents the average HR across the entire treatment period, remains a meaningful measure of average benefit to patients in the trial. However, based on the initial findings, this may need to be supplemented with other measures, such as long-term survival probabilities. Special consideration needs to be given for the designing of interim analyses, particularly related to futility, so as not to increase the probability of incorrectly stopping an effective agent. Weighted log-rank tests can be considered as supportive analysis

to better reflect the impact of a delayed treatment separation and long term durability of response observed in a subset of patients. The weights should be pre-specified in the statistical analysis plan.

References

Berry, G., R. Kitchin, and P. Mock. 1991. A comparison of two simple hazard ratio estimators based on the log-rank test. *Statistics in Medicine* 10:749–755. doi:10.1002/sim.4780100510.

Borghaei, H., L. Paz-Ares, L. Horn, D. R. Spigel, M. Steins, N. E. Ready, L. Q. Chow, E. E. Vokes, E. Felip, E. Holgado, et al. 2015. Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. *The New England Journal of Medicine* 373:1627–1639. doi:10.1056/NEJMoa1507643.

Brahmer, J., K. L. Reckamp, P. Baas, L. Crinò, W. Eberhardt, E. Poddubskaya, S. Antonia, A. Pluzanski, E. Vokes, E. Holgado, et al. 2015. Nivolumab versus docetaxel in advanced squamous-cell non–small-cell lung cancer. *The New England Journal of Medicine* 373:123–135. doi:10.1056/NEJMoa1504627.

Carroll, K. J. 2009. Back to basics: Explaining sample size in outcome trials, are statisticians doing a thorough job? *Pharmaceutical Statistics* 8:333–345. doi:10.1002/pst.362.

Chen, T. 2013. Statistical issues and challenges in immuno-oncology. *Journal for ImmunoTherapy of Cancer* 1:18. doi:10.1186/2051-1426-1-18.

Eisenhauera, E. A., P. Therasseb, J. Bogaertsc, L. H. Schwartzd, D. Sargente, R. Fordf, J. Danceyg, S. Arbuckh, S. Gwytheri, M. Mooneyg, et al. 2009. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* 45:228–247. doi:10.1016/j.ejca.2008.10.026.

Farewell, V. T. 1982. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 38:1041–1046. doi:10.2307/2529885.

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

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

Hodi, F., S. O'Day, D. McDermott, R. Weber, J. Sosman, J. Haanen, R. Gonzalez, C. Robert, D. Schadendorf, J. Hassel, et al. 2010. Improved survival with ipilimumab in patients with metastatic melanoma. *The New England Journal of Medicine* 363:711–723. doi:10.1056/NEJMoa1003466.

Huang, B., and P. Kuan. 2017. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. *Pharmaceutical Statistics* 17 (3):202–213. doi:10.1002/pst.1846.

Kalbfleisch, J. D., and R. L. Prentice. 1981. Estimation of the average hazard ratio. *Biometrika* 68:105–112. doi:10.1093/biomet/68.1.105.

Lan, K. K. G., and D. L. DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–663. doi:10.2307/2336502.

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

Maller, R., and X. Zhou. 1996. *Survival analysis with long-term survivors*. Wiley Series in Probability and Statistics. England: Chichester.

Mok, T., Y. Wu, S. Thongprasert, C. Yang, D. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose, et al. 2009. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. *The New England Journal of Medicine* 361:947–957. doi:10.1056/NEJMoa0810699.

Motzer, R., B. Escudier, D. McDermott, S. George, H. Hammers, S. Srinivas, S. Tykodi, J. Sosman, G. Procopio, E. Plimack, et al. 2015. Nivolumab versus everolimus in advanced renal-cell carcinoma. *The New England Journal of Medicine* 373:1803–1813. doi:10.1056/NEJMoa1510665.

Oller, R., and K. Langohr. 2012. FHtest: Tests for right and interval-censored survival data based on the Fleming-Harrington class (R package version 0.85). http://CRAN.R-project.org/package=FHtest.

Pampallona, S., A. A. Tsiatis, and K. Kim. 2001. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal* 35:1113–1121. doi:10.1177/009286150103500408.

Pardoll, D. M. 2012. The blockade of immune checkpoints in cancer immunotherapy. *Nature Reviews Cancer* 12:252–264. doi:10.1038/nrc3239.

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

Ray S. Lin et al. Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards. A comparative Analysis. *Statistics in Biopharmaceutical Research* 2020; 00:1-12.

Rizvi, N., J. Chaft, A. Balmanoukian, S. Goldberg, R. Sanborn, K. Steele, M. Rebelatto, Y. Gu, J. Karakunnel, and S. Antonia. 2015. Tumor response from durvalumab (MEDI4736) + tremelimumab treatment in patients with

advanced non-small cell lung cancer (NSCLC) is observed regardless of PD-L1 status. *Journal for ImmunoTherapy of Cancer* 3 (Suppl 2):193. doi:10.1186/2051-1426-3-S2-P193.
Robert, C., G. Long, G. Brady, C. Dutriaux, M. Maio, L. Mortier, J. Hassel, P. Rutkowski, C. McNeil, E. Kalinka-Warzocha, et al. 2015. Nivolumab in previously untreated melanoma without braf mutation. *The New England Journal of Medicine* 372:320–330. doi:10.1056/NEJMoa1412082.
Royston, P., and M. Parmar. 2013. Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology* 13:152. doi:10.1186/1471-2288-13-152.
SAS/STAT® 9.1 User's Guide. 2004. Cary, NC: SAS Institute.
Schemper, M. 1992. Cox analysis of survival data with non-proportional hazard functions. *The Statistician* 41:455–465. doi:10.2307/2349009.
Schoenfeld, D. A. 1983. Sample-size formula for the proportional-hazards regression model. *Biometrics* 39:499–503. doi:10.2307/2531021.
Uno, H., B. Claggett, L. Tian, E. Inoue, P. Gallo, T. Miyata, D. Schrag, M. Takeuchi, Y. Uyama, L. Zhao, et al. 2014. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*. 32(22):2380–2385. doi:10.1200/JCO.2014.55.2208.
Xu, Z., B. Zhen, Y. Park, and B. Zhu. 2017. Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine* 36:592–605. doi:10.1002/sim.7157.
Zhang, D., and H. Quan. 2009. Power and sample size calculation for log-rank test with a time lag in treatment effect. *Statistics in Medicine* 28:864–879. doi:10.1002/sim.3501.
Zucker, D. M., and E. Lakatos. 1990. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 77:853–864. doi:10.1093/biomet/77.4.853.

Appendix. Derivation of Average Hazard Ratio in the Presence of Non-Proportional Hazards

The derivation is provided in the case of a piecewise proportional hazards model with two periods. The result by Berry et al. (1991) can easily be extended to additional periods.

$$\ln(\mathrm{HR})\mho/\mathrm{V} \tag{1}$$

where

$$U = \sum_i \left(d_{1j} - {}^{n_{1j}d_j}/_{n_j}\right)$$

is the usual log-rank numerator, and

$$V = \sum_i \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

is the usual log-rank denominator.

Furthermore,

$$V\frac{\tilde{re}}{(1+r)^2} \tag{2}$$

is the reciprocal of the variance for the ln(HR), where r is the randomization ratio and e is the number of events observed.

Crucially, U and V can be partitioned into summations before and after a change in the HR.

Therefore,

$$\ln(HR)\tilde{}\frac{U_1 + U_2}{V_1 + V_2} \tag{3}$$

where U_i and V_i are the corresponding values for period i where hazards are proportional.

From (1) and (2) we know that:

$$U_i \frac{r\tilde{e}_i}{(1+r)^2} \ln(HR_i).$$

Substituting into (3) gives

$$\ln(HR)^{\sim} \frac{e_1 \ln(HR_1) + e_2 \ln(HR_2)}{e_1 + e_2}.$$

Therefore,

$$\overline{HR} = \exp(p_1 \ln(HR_1) + p_2 \ln(HR_2))$$

,

where $p_i$is the proportion of events in each period.