Design and analysis of group-sequential clinical trials based on a modestly weighted log-rank test in anticipation of a delayed separation of survival curves: A practical guidance

Dominic Magirr1 and José L Jiménez2

Abstract
Background: A common feature of many recent trials evaluating the effects of immunotherapy on survival is that non-proportional hazards can be anticipated at the design stage. This raises the possibility to use a statistical method tailored towards testing the purported long-term benefit, rather than applying the more standard log-rank test and/or Cox model. Many such proposals have been made in recent years, but there remains a lack of practical guidance on implementation, particularly in the context of group-sequential designs. In this article, we aim to fill this gap.

Methods: We illustrate how the POPLAR trial, which compared immunotherapy versus chemotherapy in non-small-cell lung cancer, might have been re-designed to be more robust to the presence of a delayed effect using the modestly-weighted log-rank test in a group-sequential setting.

Conclusion: We provide step-by-step instructions on how to analyse a hypothetical realization of the trial, based on this new design. Basic theory on weighted log-rank tests and group-sequential methods is covered, and an accompanying R package (including vignette) is provided.

Keywords
Delayed effects, group sequential designs, immuno-oncology, log-rank test, modestly-weighted log-rank test, time-to-event

Introduction
Group-sequential log-rank tests, Kaplan-Meier estimates, and Cox’s proportional hazards model are the standard tools for the analysis of confirmatory phase 3 trials with time-to-event endpoints. Although successful in general, this strategy works less well for immuno-oncology trials, where the proportional hazards assumption is often untenable, owing to a delayed separation of the survival curves. The log-rank test, although valid, may have low power if the component of the test statistic corresponding to early timepoints is contributing noise without contributing signal. In addition, the estimated beta coefficient corresponding to the treatment term in the Cox model will no longer have a straightforward interpretation.

Numerous proposals have been made to replace the log-rank test with a weighted version that is tailored towards testing purported long-term improvements in survival.1–5 Uptake has been slow, however, in part due to concerns that such tests could produce counter-intuitive results when the hazard functions on the two arms cross.6 To address such concerns, a ‘modestly-weighted’ log-rank test has been proposed,7 with the key property that if survival on the experimental drug is truly lower (or equal) to survival on control at all time-points, then the probability of claiming a statistically significant improvement is less than α. The modestly-weighted test also has considerably greater power than the standard log-rank test when there is a delayed separation of survival curves.8

In this article, we provide the guidance and tools necessary to use a modestly-weighted test in the context of a

1Advanced Methodology and Data Science, Novartis Pharma AG, Basel, Switzerland
2Global Drug Development, Novartis Pharma AG, Basel, Switzerland

Corresponding author:
José L Jiménez, Global Drug Development, Novartis Pharma AG, Fabrikstrasse 2, Basel, 4056, Switzerland.
Email: jose_luis.jimenez@novartis.com
group-sequential design. We emphasize the practical side, since, from a methodological perspective, no new concepts are required. The modestly-weighted log-rank test belongs to the class of weighted log-rank statistic studied by Fleming and Harrington, which, as shown by Tsatis, satisfy the standard independent increments assumption of group-sequential theory. We refer to Gillen and Emerson for a detailed account of the methodology.

Example: the POPLAR (NCT01903993) trial

We shall use the POPLAR trial as a starting point for our discussion. POPLAR was an open-label phase 2 randomized controlled trial of atezolizumab versus docetaxel for patients with previously-treated non-small-cell lung cancer. Key design assumptions and de-identified data are publicly available. The sample size was calculated assuming a median overall survival (OS) of 8 months for the control arm and a HR of 0.65, which translated into an assumed median OS of approximately 12.3 months for the atezolizumab arm, under an exponential model. Recruitment lasted 8 months. Three interim analyses were planned, with (two-sided) alpha levels of 0.0001, 0.0001, and 0.001. The final analysis of OS was performed when 173 deaths had occurred in the intention-to-treat (ITT) population, using a two-sided \( \alpha \) level of 4.88%. The trial enrolled a total of 287 patients.

A Kaplan-Meier estimate derived from the published data set is shown in Supplemental Figure S1. The curves display the typical late separation pattern often seen with immunotherapy agents. With the benefit of hindsight, but also based on observations from similar studies, we shall demonstrate how the trial might have been designed more robustly and efficiently, taking into account the potential for a delayed treatment effect.

Methodology

Weighted log-rank tests

To perform a weighted log-rank test, we scan over the ordered event times \( t_1, \ldots, t_k \), and take a weighted sum of the observed minus expected events on one of the treatment arms, assuming that the survival distributions on the two arms are identical. Let \( n_{i,j} \) denote the number of patients at risk on treatment \( i = 0, 1 \) just prior to time \( t_j \), and let \( O_{i,j} \) denote the observed number of events on treatment \( i = 0, 1 \) at time \( t_j \), with \( E_{i,j} = O_{i,j} \times n_{i,j} / n_j \), where \( n_j = n_{0,j} + n_{1,j} \) and \( O_{i,j} = O_{0,j} + O_{1,j} \). Then the weighted log-rank test statistic is

\[
U_W := \sum_j w_j (O_{i,j} - E_{i,j}) \sim N(0, V_W)
\]

where

\[
V_W = \sum_j w_j^2 \frac{n_{i,j} n_{1,j} O_j (n_j - O_j)}{n_j^2 (n_j - 1)}
\]

Intuitively, if the treatment is beneficial, we should see fewer events on the experimental arm than would be expected assuming the curves are identical. We are hoping that \( U_W \ll 0 \), and, in particular, that the one-sided \( p \)-value, \( p := \Phi(U_W / \sqrt{V_W}) \), is less than, for example, \( \alpha = 0.025 \). Weights are pre-specified to boost the chances that \( p < \alpha \), given the anticipated treatment effect. The standard log-rank test uses \( w_j = 1 \), which is the most powerful choice under proportional hazards. Under a delayed-separation scenario, a popular alternative is the Fleming-Harrington-(0,1) test, with \( w_j = 1 - S(t_j -) \), where \( S(t_j -) \) is the Kaplan-Meier estimate of the pooled sample just prior to time \( t_j \). Care is necessary, however, since although the Fleming-Harrington-(0,1) test controls the type 1 error rate when survival curves are identical, it offers no guarantees regarding the direction of the effect. To put it another way: it offers a valid \( \alpha \)-level test when the null hypothesis is identical survival, \( H_0: S_0(t) = S_1(t) \) for all \( t \), but not when the null hypothesis is inferior (or identical) survival, \( H_0: S_0(t) \leq S_1(t) \) for all \( t \). A safer choice that controls \( \alpha \) also under \( H_0 \) is a ‘modestly-weighted’ log-rank test, which uses \( w_j = 1 / \max \{ S(t_j -), S(t_j) \} \).

Heuristically, the modestly-weighted test can be thought of as similar to an average landmark analysis from time \( t^\star \) to the end of follow up. This interpretation is helpful at the design stage when pre-specifying \( t^\star \). If there are several candidate landmark times which might be of interest, then the earliest such timepoint may be a good candidate for \( t^\star \). More discussion on the choice of \( t^\star \) is given for a specific example below.

Group-sequential weighted log-rank tests

For a group-sequential version of the weighted log-rank test, we must consider the joint distribution of \( U_W^{(1)}, \ldots, U_W^{(K)} \), where \( U_W^{(k)} \) denotes the test statistic at analysis \( k \). As shown by Tsatis, asymptotically under \( H_0 \)

\[
\begin{pmatrix}
U_W^{(1)} \\
U_W^{(2)} \\
\vdots \\
U_W^{(K)}
\end{pmatrix}
\sim N
\begin{pmatrix}
0 \\
0 \\
\vdots \\
0
\end{pmatrix},
\begin{pmatrix}
U_W^{(1)} & U_W^{(1)} & \cdots & U_W^{(1)} \\
U_W^{(2)} & U_W^{(2)} & \cdots & U_W^{(2)} \\
\vdots & \vdots & \ddots & \vdots \\
U_W^{(K)} & U_W^{(K)} & \cdots & U_W^{(K)}
\end{pmatrix}
\]

A group-sequential test can be defined via the \( K \) critical values, \( c_1, \ldots, c_K \) such that

\[
p \left( \bigcap_{k=1}^K \frac{U_W^{(k)}}{\sqrt{V_W^{(k)}}} > c_k \right) \leq 1 - \alpha
\]

There are many different ways to choose such critical values. One flexible approach is to use a Hwang-Shih-DeCani alpha-spending function. In this
Table 1. Relationship between number of patients per arm, expected total number of events, and power; using the standard log-rank test (LR) and the modestly weighted log-rank test (MWLR). Assuming uniform recruitment over 8 months, time-to-event distributions as given in Supplemental Figure S2, with analysis performed 21 months after the start of the trial. For both scenarios the survival distribution on the control arm is assumed to be exponential with median 8 months. The distribution on the experimental arm is assumed to be exponential with median 12.3 months (‘0 m delay’), or two-piece exponential with rate log(2)/8 prior to 4 months, and rate log(2)/16.6 after 4 months.

| Patients per arm | Total events  | Power LR | Power MWLR (t* = 6) |
|-----------------|---------------|----------|---------------------|
|                 | 0 m delay     | 4 m delay| 0 m delay           | 4 m delay |
| 150             | 207           | 203      | 0.87                | 0.84      | 0.87                | 0.91      |
| 155             | 214           | 210      | 0.88                | 0.85      | 0.88                | 0.91      |
| 160             | 221           | 217      | 0.89                | 0.86      | 0.89                | 0.92      |
| 165             | 228           | 223      | 0.90                | 0.87      | 0.90                | 0.93      |
| 170             | 235           | 230      | 0.91                | 0.88      | 0.90                | 0.94      |
| 175             | 241           | 237      | 0.92                | 0.89      | 0.91                | 0.94      |
| 180             | 248           | 244      | 0.92                | 0.90      | 0.92                | 0.95      |

In time-to-event settings, power is driven by the number of events rather than the number of patients. The number of events is a function of the recruitment assumptions, time-to-event distributions, and the duration of follow up. We thus have considerable flexibility, in theory at least, in how we design the trial to meet objective equation (5). It may sometimes be feasible to fix (approximately) the duration of recruitment and follow-up. In this case, we would adjust the recruitment rate, or, equivalently, the total number of patients, until equation (5) is satisfied. For example, the POPLAR trial specified 8 months of recruitment, plus a minimum follow-up time of 13 months, bringing the total trial duration to 21 months. Given these assumptions, as well as the time-to-event distributions in Supplemental Figure S2, the corresponding power of the standard log-rank test is shown in Table 1 for a series of potential sample sizes. The power has been calculated via numerical integration using the R package gsdelayed (available at github.com/dominicmagirr/gsdelayed), that we specifically developed to illustrate all the steps presented in this article. Computational details have already been described elsewhere,7 but the basic idea is to calculate the expected number of events, as well as the expected average hazard ratio, based on the design assumptions, which can then be used as inputs to standard sample size formulae. The expected number of events at each analysis under the two design scenarios are also included in Table 1. Once a design has been chosen, one can then proceed to design the trial, to either keep the calendar time of the analyses fixed and allow the number of events to deviate from the plan, or instead to keep the number of events fixed and allow for the POPLAR trial design, also described in Table 1. Our challenge is to find a design such that

$$p\left(U_w / \sqrt{V_w} < \Phi^{-1}(\alpha); H_1\right) = 1 - \beta$$  \hspace{1cm} (5)

As an alternative to an alpha-spending function where the information fraction is given by $V_{w}^{(k)} / \hat{V}_{w}^{(k)}$, one might also simply pre-fix $\alpha_1^* \leq \ldots \leq \alpha_K^* := \alpha$. We shall evaluate both approaches in this article. Note, however, that far more possibilities exist. The information fraction could be based on the number of events, or calendar time. In-depth discussion on alpha-spending functions can be found in Jennison and Turnbull,11 for example.

### Design

**Sample size calculation: fixed sample**

We now consider the alternative hypothesis, denoted by $H_1$. Supplemental Figure S2 shows two potential alternative hypotheses that may have been considered
the calendar times of the analyses to deviate from the
plan. In the implementation section below we shall opt
for the latter approach.

Table 1 indicates that under the proportional
hazards (‘0 m delay’) alternative, a sample size of 165
patients per arm would be sufficient to achieve 90% power,
based on a trial duration of 21 months, with an
expected number of events of 228. However, under the
non-proportional hazards (‘4 m delay’) assumption, 180
patients per arm would be required with an
expected number of events of 244. If, instead of the
standard log-rank test, we use the modestly-weighted
log-rank test with \( r^* = 6 \), then the corresponding
required number of patients per arm is 165 under pro-
portional hazards and 150 under non-proportional
hazards. The corresponding expected number of events
are 228 and 203, respectively.

The choice of \( r^* = 6 \) requires some explanation.
Note first that \( r^* = 0 \) is the same as the standard log-
rank test, and for values of \( r^* \) close to 0 there will be lit-
tle difference between these two tests. Also, as \( r^* \to \infty \),
the weights reduce to \( w_j = 1/\hat{S}(t_j -) \), which Grey and
Tsiatis\(^\text{19}\) show are the optimal weights under a
proportional-distributions cure-rate model. Loosely
speaking, this is similar to comparing the survival prob-
abilities at the end of follow-up. For intermediate val-
ues of \( r^* \), the modestly-weighted log-rank test is similar
to an average landmark analysis from time \( t^* \) until the
end of follow-up.\(^8\) This means that if we have several
different landmark times where we are potentially inter-
ested in the difference in survival curves, then the earli-
est such timepoint is a good candidate for \( r^* \). Note, in
particular, that if we anticipate a delay of 4 months,
this does not imply that we should choose \( r^* = 4 \). If we
are confident about the delay then a somewhat later \( r^* \)
will have higher power. However, if there is some
uncertainty regarding the delay, then choosing \( r^* \) closer
to zero protects the power in case proportional hazards
does indeed hold. In addition to heuristic arguments,
investigating operating characteristics for a range of \( r^* \)
is also helpful, as shown in our simulation study below.

To summarize, if we are confident in the 4 month
delay assumption, and require 90% power, then there is
an approximate 20% saving in the number of patients
per arm from using the modestly-weighted log-rank test
instead of a standard log-rank test. Even if we are not
certain about the delayed effect, and would prefer to
choose the sample size such that there is at least 90% power under both proportional hazards and non-
proportional hazards alternatives, there is still an
approximate 10% reduction from using the modestly-
weighted test.

Adding an interim analysis (efficacy)

To add an interim analysis for efficacy, two choices are
necessary: the timing of the interim analysis, and the
amount of \( \alpha \) to spend. For our example based on
POPLAR, unless the interim analysis is done very
early, all patients will have already been recruited. The
main incentive for adding a stopping boundary for effi-
cacy is to reduce the expected duration of the study.
Typically, however, there is a trade-off: the more we
reduce the expected duration of the trial, the more we
reduce the overall power. If we decide to increase the
maximum sample size to recover 90% power, we must
trade off a shorter expected duration versus a longer
maximum duration.

In Table 2, expected duration and power are dis-
played for 10 potential designs. Calculations are per-
formed via numerical integration using the R package
gdelayed as described above. Among this (small) selec-
tion of examples, the three-stage design with interim
analyses at 11 and 16 months stands out as an appealing
option, based on a Hwang-Shih-DeCani spending func-
tion with \( \gamma = -4 \). This design reduces the expected
duration of the study by 3.4 months with barely any
reduction in power compared to a single-stage design. In
addition to power and expected duration under the
alternative hypothesis, many other metrics may be con-
sidered, such as the expected duration under the null
hypothesis, or averaged over a range of scenarios. The
topic is covered in-depth elsewhere.\(^11,12\)

Adding an interim analysis (futility)

Regulatory guidance generally steers towards futility
stopping rules that are non-binding.\(^20\) that is, the

---

Table 2. Expected duration and power of various design options. Based on a modestly-weighted log-rank test with \( r^* = 6 \), sample size of 150 per arm, and uniform recruitment over 8 months. Survival distributions are assumed to follow the ‘4 m delay’ scenario in Supplemental Figure S2. ‘Single-stage’ refers to final analysis. ‘Two-stage’ refers to interim + final analyses. ‘Three-stage’ refers to two interim + final analyses.

| Design     | Analysis times | Total events | E(months duration) \( \gamma = -4 \) | \( \gamma = -1.5 \) | \( \gamma = 1 \) | Power \( \gamma = -4 \) | \( \gamma = -1.5 \) | \( \gamma = 1 \) |
|------------|----------------|--------------|--------------------------------------|-------------------|-------------------|---------------------|-------------------|-------------------|
| Single-stage | 21             | 203          | 21                                   | 21                | 21                | 0.91                 | 0.91              | 0.91              |
| Two-stage   | 11, 21         | 123, 203     | 20.1                                 | 19.4              | 18.8              | 0.90                 | 0.89              | 0.86              |
| Two-stage   | 16, 21         | 170, 203     | 17.9                                 | 17.6              | 17.4              | 0.90                 | 0.88              | 0.86              |
| Three-stage | 11, 16, 21     | 122, 170, 203| 17.6                                 | 17.0              | 16.7              | 0.90                 | 0.88              | 0.83              |
stopping rule is not considered when calculating the efficacy boundary to guarantee an \( \alpha \)-level test. Adding non-binding futility rules reduces both the type I error probability and the power.

There are several ways that a futility rule could be specified.\(^{21}\) We could, for example, consider a beta-probability and the power.

The latter has been implemented in gsdelayed. On the \( z \)-statistic scale or on the average-hazard-ratio scale. The latter has been implemented in gsdelayed.

For time-to-event trials with an anticipated delayed separation of the survival curves, a formal futility analysis may have limited value. Unless the interim analysis occurs very early, most patients will have been recruited. In addition, a stringent rule would risk stopping inappropriately before the treatment effect has been given a chance to emerge. This is not to say that the trial would never be stopped early. An independent data safety and monitoring board (DSMB) will stop the trial if the experimental drug is clearly harmful.\(^{25}\)

### Implementation

We shall now walk through a hypothetical realization of the three-stage trial design from Table 2. We emphasize that this realization is not based on the results of the POPLAR study. Figure 1 shows how the expected number of events corresponds to calendar time under the 4 month delay alternative. We see that the first interim, second interim and final analyses at months 11, 16 and 21, correspond to 122, 170 and 203 events, respectively. As described in the design section, we now have a choice of either fixing the calendar time points of the analyses, or fixing the number of observed events that will trigger each analyses. Here, we opt for the latter approach. Having done so, the planned stopping boundaries are shown in Figure 2(d).

The planned stopping boundaries are based on the assumed joint distribution of the test statistics under the design assumptions. Since the weights in the weighted log-rank test depend on the pooled survival distribution and recruitment distribution, which are inevitably misspecified at the design stage, we must update the boundary at each analysis in light of the observed variance of the test statistic. This is achieved via the alpha-spending function. Note that deviations from the exact numbers of planned events can also be handled in this way. As mentioned in the group-sequential weighted log-rank tests section, we consider two types of spending function. First, the Hwang-Shi-DeCani alpha-spending function with \( \gamma = -4 \), based on the information fraction \( V^{(k)}_W / \tilde{V}^{(k)}_W \). Here, differences between the planned \( \tilde{V}^{(k)}_W \) and observed \( V^{(k)}_W \) lead to differences between the actual alpha spend compared to the planned alpha spend at each interim. The crucial parameter is the anticipated variance of the \( U \) statistic at the final analysis, denoted by \( \tilde{V}^{(k)}_W \). For our example, based on the design in the final row of Table 2, we find via numerical integration that \( \tilde{V}^{(3)}_W = 103.4 \). The second type of alpha-spending approach that we consider is to simply fix the cumulative alpha spend at each analysis as dictated by the design. In our case we have \( \alpha_1^* = 0.00301, \alpha_2^* = 0.0106, \alpha_3^* = 0.025 \).

We now describe how our hypothetical trial proceeds:

- **Trial recruitment begins.**
- **We conduct the first interim analysis after 122 events.** Suppose we observe the data shown in Figure 2(a). Applying the modestly-weighted test, \( U^{(1)}_W = -6.46 \) and \( V^{(1)}_W = 50.4 \). For the Hwang-Shi-DeCani alpha-spending function approach, we would plug the information fraction \( t = V^{(1)}_W / \tilde{V}^{(1)}_W = 0.487 \) into equation (3) to find \( \alpha_1^* = 0.00281 \). Or, using the fixed alpha-spending approach we would use \( \alpha_1^* = 0.00301 \). This corresponds to critical values on the \( Z \)-statistic scale of \(-2.770 \) or \(-2.747 \), respectively. In either case, since the observed \( Z \)-statistic \( U^{(1)}_W / \sqrt{V^{(1)}_W} = -0.91 \), the decision at the first interim would be to continue. This is represented graphically in Figure 2(d) by the blue ‘x’ at 122 events, where changes to the pre-planned critical value would be too small to be perceptible.
- **We conduct the second interim analysis after 170 events.** Suppose we observe the data shown in Figure 2(b). Applying the modestly-weighted test, \( U^{(2)}_W = -13.6 \) and \( V^{(2)}_W = 78.1 \), so that \( Z_2 = U^{(2)}_W / \sqrt{V^{(2)}_W} = -1.53 \). The information fraction is now \( V^{(2)}_W / \tilde{V}^{(2)}_W = 0.755 \) so that the Hwang-Shi-DeCani alpha-spending function gives

![Figure 1. Switching from a study time perspective to an expected number of events perspective. Based on the alternative hypothesis (NPH).](image-url)
If we apply fixed alpha spending we would keep \( \alpha_2^* = 0.0106 \). Based on the observed correlation between \( Z_1 \) and \( Z_2 \), and using equation (4), together with the critical value found at the first stage, this corresponds to critical values for \( Z_2 \) of \(-2.42\) or \(-2.35\), respectively. Since, in either case, \( Z_2 \) does not exceed the critical value, the decision would be to continue the trial. This is represented graphically in Figure 2(d) by the blue ‘x’ on top of 170 events.

We conduct the final analysis after 203 events. Suppose we observe the data shown in Figure 2(c). Applying the modestly-weighted test, \( V_{\text{W}}^{(3)} = -23.4 \) and \( V_{\text{W}}^{(3)} = 97.2 \), so that \( Z_3 = -2.37 \). Notice that \( V_{\text{W}}^{(3)} < V_{\text{W}}^{(3)} \). With the pre-fixed alpha-spend we would have \( \alpha_3^* = 0.025 \) and that would be the end of the trial. For the alpha-spending function with maximum information equal to \( V_{\text{W}}^{(3)} = 103.4 \), we could in principle carry on the trial to further analyses until the maximum information is reached. In practice, it may be preferable to pre-specify a maximum of three analyses, and stipulate that all remaining \( \alpha \) shall be spent at the final analysis. We assume this to be the case here, so that \( \alpha_3^* = 0.025 \).

Note that as a variation on this procedure, one could stipulate that if the maximum information has almost been reached, for example, if \( V_{\text{W}}^{(3)} / V_{\text{W}}^{(3)} > 0.95 \), say, then the trial will be stopped with all remaining alpha to be spent, otherwise the trial will continue to a further analysis – see the discussion in the simulation study section below. Based on the observed correlation between \( Z_1 \), \( Z_2 \) and \( Z_3 \), and using equation (4), together with the critical values found at the first and second analyses, this corresponds to a critical value for \( Z_3 \) of \(-2.00\) when applying the information-based

---

**Figure 2.** Kaplan-Meier curves at interim analyses 1 (a) and 2 (b), and at the final analysis (c). Also shown are the values of the z-statistics over time, superimposed on the planned stopping boundaries (d).
spending function, or \(-2.01\) when applying the fixed alpha spending approach. Since, in either case, \(Z_1\) exceeds the critical value, the final decision would be to reject the null hypothesis. This is represented graphically in Figure 2(d) by the blue ‘x’ on top of 203 events.

To define a p-value based on a group sequential design, we refer the reader to standard texts on group-sequential design theory.\(^\text{11,12}\) One approach is a so-called ‘stage-wise ordering p-value’, where earlier stops for efficacy are always considered more extreme evidence against the null hypothesis than later stops for efficacy. In the hypothetical trial described above, for the fixed alpha-spending approach, the stage-wise (one-sided) p-value would be:

\[
1 - P_{t_6}(Z_1 > -2.75 \cap Z_2 > -2.35 \cap Z_3 > -2.37) = 0.015
\]

As has been noted by many authors,\(^\text{5,8,26,27}\) in the setting of non-proportional hazards, there is no single-number summary measure that can adequately capture the full information from the survival curves.

In our hypothetical realization, we might focus on the survival probabilities at 18 months (0.38 on experimental versus 0.23 on control), the median survival times (11.3 versus 9.3 months), or the restricted mean survival times up to 18 months (11.2 versus 10.3 months). Weighted hazard ratios have also been proposed in this context,\(^\text{28}\) although their interpretability has been debated.\(^\text{29,30}\)

For all summary measures, a group sequential design introduces some bias, owing to the possibility to stop early on a random high. Various methods have been proposed that attempt to account for this bias.\(^\text{31-33}\) They are rarely used in practice, however, with the justification often that the size of the bias is small, particularly if the interim analyses occur late.\(^\text{34}\) We refer the reader to Robertson et al.\(^\text{35}\) for in-depth discussion of this issue.

### Robustness to design assumptions

Robustness of the proposed approach to design assumptions can be assessed via simulations. Here, as in the implementation section, we pre-specify a design that is based on a modestly-weighted log-rank test (\(t^* = 6\)) with 150 patients per arm. First interim, second interim, and final analyses are triggered after 122, 170 and 203 events, respectively. According to the sample-size calculation in the design section (based on numerical integration), the power should be 90% when recruitment is uniform over 8 months, and the survival distributions are as described in the ‘Delay = 4 months’ scenario in Figure 3(b). We now check this calculation via simulation, and assess how the power changes if the survival and/or recruitment distributions are misspecified as in Figure 3(a)–(c). Results are presented in the \(t^* = 6\) columns of Table 3. When picking alternative scenarios, we have attempted to offset longer delays with better long-term survival, but this does not necessarily imply that the scenarios represent equal magnitudes of clinical benefit. Note also that it is not the duration of the delay period per se that impacts power, but rather the proportion of the total number of events that occur during the delay period. In these simulations we have applied the alpha-spending function approach with the information fraction indexed by the variance of the score statistic. We have also stipulated that if 95% of the maximum information is already available at the first interim analysis then the trial will stop with \(\alpha_{t^*} = 0.025\). Similarly at the second interim if 97.5% of the maximum

---

**Table 3.** Simulation study with possibly misspecified survival and recruitment distributions. For each analysis method (\(t^* = 0, 6, 12\)), the design is based on a control median of 8 months, a delay of 4 months, a sample size of 150 per arm, uniform recruitment over 8 months, analyses triggered after 122, 170 and 203 events, and a Hwang-Shi-DeCani (\(\gamma = -4\)) alpha-spending function. Recruitment scenario equation (1) is a uniform distribution over 8 months, as assumed in the design. In scenario equation (2), recruitment times \(R\) are simulated under \(P(R=r) = (r/15)^{t^*}\) for \(r=15\). See Figure 3 for full details of the survival distribution assumptions. Based on 10,000 simulations.

| Recruitment scenario | Treatment effect | Power | Control median = 6 m | Control median = 8 m | Control median = 10 m |
|----------------------|-----------------|-------|----------------------|----------------------|----------------------|
|                      |                 | \(t^* = 0\) | \(t^* = 6\) | \(t^* = 12\) | \(t^* = 0\) | \(t^* = 6\) | \(t^* = 12\) | \(t^* = 0\) | \(t^* = 6\) | \(t^* = 12\) |
| 1.                   | Null            | 0.028 | 0.024 | 0.027 | 0.027 | 0.025 | 0.028 | 0.023 | 0.026 | 0.024 |
|                      | Prop. Haz.      | 0.85  | 0.84  | 0.82  | 0.85  | 0.86  | 0.84  | 0.85  | 0.86  | 0.85  |
|                      | Delay = 4 m     | 0.62  | 0.75  | 0.81  | 0.62  | 0.75  | 0.81  | 0.86  | 0.85  | 0.85  |
|                      | Delay = 8 m     | 0.18  | 0.22  | 0.39  | 0.18  | 0.22  | 0.39  | 0.92  | 0.95  | 0.97  |
| 2.                   | Null            | 0.024 | 0.024 | 0.026 | 0.026 | 0.027 | 0.025 | 0.93  | 0.96  | 0.99  |
|                      | Prop. Haz.      | 0.85  | 0.85  | 0.81  | 0.85  | 0.85  | 0.83  | 0.86  | 0.86  | 0.85  |
|                      | Delay = 4 m     | 0.62  | 0.74  | 0.81  | 0.62  | 0.74  | 0.81  | 0.86  | 0.95  | 0.97  |
|                      | Delay = 8 m     | 0.19  | 0.23  | 0.41  | 0.19  | 0.23  | 0.41  | 0.93  | 0.96  | 0.99  |
information is already available. The trial stops after a maximum of three analyses with $\alpha_3^* = 0.025$. This procedure has the advantage of avoiding numerical instability, as well as mimicking more closely what would happen in practice. In addition, in the unlikely event that the variance of the score statistics decreases from one analysis to the next, we set the boundary to $\alpha_0^*$, except for the final analysis, where we would spend all remaining alpha in any case, and cap at 1 the correlations in equation (1).

Table 3 shows that the operating characteristics are reasonably robust to misspecification of the control event rate, timing of the separation, and recruitment assumptions. The exception is when the control event rate is higher than expected and the delay is longer than expected. This situation is particularly challenging, with approximately 60% of events occurring before there is any separation of the survival curves. If such a scenario is considered plausible, one should consider choosing a larger value of $\tau$ and increasing the sample size.

Also included in Table 3 are results from the standard log-rank test ($\tau = 0$) and the modestly-weighted test with $\tau = 12$. These two tests have been applied with the same sample size, the same analysis trigger points, as well as the same alpha-spending function. Results follow the same pattern as was discussed for a single-stage design in the design section. For long delays, a large value of $\tau$ will increase power, but we need to trade this off against some reduction in power under proportional hazards.

In Supplemental Table S1, the simulation study has been repeated under the fixed alpha-spending approach. Conclusions are broadly the same, although the power appears somewhat more robust to model misspecification than when using the information-based spending function. This makes sense, given that the information-based spending function induces variation in the amount of interim alpha spend.

**Concluding remarks**

Trials involving immunotherapy often have a delayed separation of survival curves. We could use this knowledge to increase efficiency, by focusing the test statistic on the purported long-term survival benefit, rather than using the standard log-rank test as a default. One potential barrier, however, is a lack of guidance and software for implementation in the context of group-sequential trials. In this article, we have described in detail how to design and analyse a phase 3 trial in immuno-oncology using a group-sequential modestly-weighted log-rank test. We have also discussed the scope for a formal futility analysis in the special case of a time-to-event endpoint with an anticipated delayed effect. Finally, we have illustrated how a range of single-number summary measures together help to quantify the treatment effect, which is important given that the hazard ratio lacks interpretability in this setting.

We caution that this article presents a limited number of comparisons of the modestly-weighted log-rank test to other relevant tests. To provide for an informed choice of the most appropriate test, a comprehensive comparison including the evaluation of robustness to

![Figure 3. Scenarios for the simulation study to assess robustness of methods to design assumptions. Survival on the control arm is exponential with median either (a) 6 months, (b) 8 months, or (c) 10 months. On the experimental arm, the ‘Delay = 0 months’ scenario is exponential with median (a) 9.2 months, (b) 12.3 months, or (c) 15.4 months. The ‘Delay = 4 months’ scenario is two-piece exponential with a rate equal to the control arm rate until 4 months, and then a rate of (a) $\log(2)/12.6$, (b) $\log(2)/16.6$, or (c) $\log(2)/21.6$, thereafter. The ‘Delay = 8 months’ scenario is two-piece exponential with a rate equal to the control arm rate until 8 months, and then a rate of (a) $\log(2)/25$, (b) $\log(2)/35$, or (c) $\log(2)/45$, thereafter.](image-url)
departures from the model assumptions or the misspecifications of the test parameters (e.g. τ∗) would be necessary.

Data availability
The data used to produce the Kaplan-Meier curves in Supplemental Figure S1 is publicly available in Gandara et al.15 The R code used throughout the article is part of the package gsdelayed, which includes a vignette, and is available at github.com/dominicmagirr/gsdelayed.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
José L Jiménez https://orcid.org/0000-0002-8809-2717

Supplemental material
Supplemental material for this article is available online.

References
1. Harrington DP and Fleming TR. A class of rank test procedures for censored survival data. Biometrika 1982; 69(3): 553–566.
2. Yang S and Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics 2010; 66(1): 30–38.
3. Garès V, Andrieu S, Dupuy JF, et al. A comparison of the constant piecewise weighted logrank and Fleming-Harrington tests. Electron J Stat 2014; 8(1): 841–860.
4. Karrison TG. Versatile tests for comparing survival curves based on weighted log-rank statistics. Stat J 2016; 16(3): 678–690.
5. Roychoudhury S, Anderson KM, Ye J, et al. Robust design and analysis of clinical trials with nonproportional hazards: A straw man guidance from a cross-pharma working group. In: Statistics in Biopharmaceutical Research. Taylor & Francis, 2021, pp. 1–15.
6. Freidlin B and Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? J Clin Oncol 2019; 37(35): 3455–3459.
7. Magirr D and Burman CF. Modestly weighted logrank tests. Stat Med 2019; 38(20): 3782–3790.
8. Magirr D. Non-proportional hazards in immuno-oncology: is an old perspective needed? Pharm Stat 2021; 20: 512–527.
9. Fleming TR and Harrington DP. Counting processes and survival analysis, vol. 169. Hoboken, NJ: John Wiley & Sons, 2011.
10. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. J Am Stat Assoc 1982; 77(380): 855–861.
11. Jennison C and Turnbull BW. Group sequential methods with applications to clinical trials. Boca Raton, FL: CRC Press, 1999.
12. Whitehead J. The design and analysis of sequential clinical trials. Hoboken, NJ: John Wiley & Sons, 1997.
13. Gillen DL and Emerson SS. Information growth in a family of weighted logrank statistics under repeated analyses. Sequent Anal 2005; 24(1): 1–22.
14. Fehrenbacher L, Spira A, Ballinger M, et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. Lancet 2016; 387(10030): 1837–1846.
15. Gandara DR, Paul SM, Kowanetz M, et al. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. Nat Med 2018; 24(9): 1441–1448.
16. Rahman R, Fell G, Ventz S, et al. Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: prevalence, associated factors, and implications. Clin Cancer Res 2019; 25(21): 6339–6345.
17. Jiménez JL, Stalbovskaya V and Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. Pharm Stat 2019; 18(3): 287–303.
18. Hwang IK, Shih WJ and De Cani JS. Group sequential designs using a family of type I error probability spending functions. Stat Med 1990; 9(12): 1439–1445.
19. Gray RJ and Tsiatis AA. A linear rank test for use when the main interest is in differences in cure rates. Biometrics 1989; 45: 899–904.
20. Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics: guidance for industry, https://www.fda.gov/media/78495/download
21. Gallo P, Mao L and Shih VH. Alternative views on setting clinical trial futility criteria. J Biopharm Stat 2014; 24(5): 976–993.
22. Pampallona S, Tsiatis AA and Kim K. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. Drug Inform J 2001; 35(4): 1113–1121.
23. Lachin JM. A review of methods for futility stopping based on conditional power. Stat Med 2005; 24(18): 2747–2764.
24. Spiegelhalter DJ, Freedman LS and Blackburn PR. Monitoring clinical trials: conditional or predictive power? Control Clin Trials 1986; 7(1): 8–17.
25. DeMets DL and Lan KK. Interim analysis: the alpha spending function approach. Stat Med 1994; 13(13–14): 1341–1352.
26. Rufibach K. Treatment effect quantification for time-to-event endpoints–estimands, analysis strategies, and beyond. Pharm Stat 2019; 18(2): 145–165.
27. Jiménez JL. Quantifying treatment differences in confirmatory trials under non-proportional hazards. J Appl Stat. Epub ahead of print 3 September 2020. DOI: 10.1080/02664763.2020.1815673.
28. Lin RS and León LF. Estimation of treatment effects in weighted log-rank tests. *Contemp Clin Trials Commun* 2017; 8: 147–155.

29. Bartlett JW, Morris TP, Stensrud MJ, et al. The hazards of period specific and weighted hazard ratios. *Stat Biompharm Res* 2020; 12(4): 518–519.

30. Lin RS, Lin J, Roychoudhury S, et al. Rejoinder to letter to the editor 'The hazards of period specific and weighted hazard ratios'. *Stat Biompharm Res* 2020; 12(4): 520–521.

31. Pinheiro JC and DeMets DL. Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika* 1997; 84(4): 831–845.

32. Fan XF, DeMets DL and Lan KKG. Conditional bias of point estimates following a group sequential test. *J Biompharm Stat* 2004; 14(2): 505–530.

33. Walter SD, Guyatt GH, Bassler D, et al. Randomised trials with provision for early stopping for benefit (or harm): the impact on the estimated treatment effect. *Stat Med* 2019; 38(14): 2524–2543.

34. Freidlin B and Korn EL. Stopping clinical trials early for benefit: impact on estimation. *Clin Trials* 2009; 6(2): 119–125.

35. Robertson DS, Choodari-Oskooei B, Dimairo M, et al. Point estimation for adaptive trial designs, https://orca.cardiff.ac.uk/141478/1/2105.08836.pdf