"6 choose 4": A framework to understand and facilitate discussion of strategies for overall survival safety monitoring

Godwin Yung

F. Hoffmann-La Roche AG, South San Francisco, CA, US

and

Kaspar Rufibach

F. Hoffmann-La Roche AG, Basel, Switzerland

and

Marcel Wolbers

F. Hoffmann-La Roche AG, Basel Switzerland

and

Mark Yan

F. Hoffmann-La Roche AG, Mississauga, CAN

and

Jue Wang

F. Hoffmann-La Roche AG, South San Francisco, CA, US

December 9, 2025

Abstract

Advances in anticancer therapies have contributed significantly to decreasing death rates in certain disease and clinical settings. However, they have also made it difficult to power a clinical trial in these settings with overall survival (OS) as the primary efficacy endpoint. Therefore, two approaches have recently been proposed for the pre-specified analysis of OS as a safety endpoint (Fleming et al., 2024; Rodriguez et al., 2024). We provide in this paper a simple unifying framework that includes the aforementioned approaches—and a few others—as special cases. By highlighting each approach's focus, priority, tolerance for risk, and strengths or challenges for practical implementation, this framework can help facilitate discussions between stakeholders on "fit-for-purpose OS data collection and assessment of harm" (American Association for Cancer Research, 2023). We apply this framework to a real clinical trial in large B-cell lymphoma to illustrate its application and value. Several recommendations and open questions are also raised.

1 Introduction

With the continued advancement of anticancer therapies, there are now many disease and clinical contexts in which patients' lives are extended and multiple treatment options are available following disease progression. Although these successes should be celebrated, they have also made it challenging for clinical trials to have adequate power to detect statistically significant improvements in overall survival (OS)—the gold standard endpoint in oncology (Merino et al., 2024). In July 2023, a workshop titled "Overall Survival in Oncology Clinical Trials" was held jointly by the U.S. Food and Drug Administration (FDA), American Association for Cancer Research (AACR), and American Statistical Association (ASA) to discuss challenges with the timely evaluation of OS for novel therapies (American Association for Cancer Research, 2023). The workshop cited lack of OS data, lack of plans for further OS data collection, and lack of pre-specified OS analyses as emerging challenges for regulatory authorities to interpret OS data in trials where OS is not a primary or key secondary endpoint. A call to action was made to "encourage thoughtful and comprehensive planning during trial design for fit-for-purpose OS data collection and assessment of harm given the disease and clinical context".

Following that call-to-action, two statistical approaches were proposed for the pre-specified analysis of OS as a safety endpoint. Fleming et al. (2024) proposed a monitoring guideline inspired by ones used by the FDA in cardiovascular safety trials. Rodriguez et al. (2024) suggested using confidence intervals to decide whether or not there is sufficient evidence to rule out harm. These approaches are similar to a certain degree. For example, both consider similar sets of parameters: definition of harm measured in terms of the OS hazard ratio, type I error, etc.. However, upon closer look, it can be unclear even to the experienced statistician how these approaches relate to each other exactly, how they differ, which approach one should use in practice, and why.

This paper proposes a framework for OS safety analysis that includes as special cases Fleming et al. (2024) and Rodriguez et al. (2024). Importantly, this framework explains what flexibility and restrictions exist for *any* approach analyzing OS as a safety endpoint under the traditional framework of hypothesis testing, and clarifies for each approach their focus, priority, and tolerance for risk over the course of a trial. By increasing understanding

3

of different approaches and the relationship between them, this framework can help facilitate "thoughtful planning" by sponsors and improve communication between stakeholders in order to achieve alignment on clinical trial design and analysis.

The rest of the paper is organized as follows. In Section 2, we present our framework for OS safety monitoring and provide five examples of approaches that fall under this framework, including Fleming et al. (2024) and Rodriguez et al. (2024). In Section 3, we illustrate and compare the five approaches by applying them to POLARIX, a real clinical trial studying patients with untreated large B-cell lymphoma. Section 4 concludes with recommendations for practice and open questions for further deliberation or research.

2 Proposed framework

Consider a trial with n patients, $n\pi$ randomized to the experimental arm and $n(1-\pi)$ randomized to the control arm. A prospectively designed hypothesis test for OS detriment—as measured by the OS hazard ratio (HR)—involves six parameters:

1. θ_0 the OS HR under the null hypothesis H_0 of OS detriment

2. θ_1 the OS HR under the alternative hypothesis H_1 of no OS detriment

3. d the number of deaths

4. θ^* the threshold for decision making between H_0 vs. H_1

5. α the one-sided type I error rate

6. β the type II error rate

For the purpose of safety monitoring in indolent cancer trials, we may limit ourselves to considering $\theta_0 > 1.0$ and moderate to neutral θ_1 (e.g., between 0.7 and 1.0). Fleming et al. (2024) characterized θ_0 as "the smallest unacceptable detrimental OS HR" and θ_1 as "a plausible OS HR consistent with benefit that is reasonably expected from the experimental intervention".

Two classical equations characterize the approximate relationship between the six parameters:

$$\alpha = Pr(\hat{\theta} < \theta^* | \theta_0) = \Phi\left(\log(\theta^*/\theta_0)\sqrt{\pi(1-\pi)d}\right) \tag{I}$$

$$1 - \beta = Pr(\hat{\theta} < \theta^* | \theta_1) = \Phi\left(\log(\theta^*/\theta_1)\sqrt{\pi(1-\pi)d}\right) \tag{II}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. These equations assume proportional hazards. More specifically, they assume that the observed log hazard ratio $\log(\hat{\theta})$ follows a normal distribution with variance $1/(\pi(1-\pi)d)$ and mean $\log(\theta_0)$ or $\log(\theta_1)$ (Schoenfeld, 1981).

Given six parameters and Equations (I)-(II) characterizing the relationship between them, our proposed framework is as follows: *at each analysis time, users input or "choose" 4 of the 6 parameters (not all from the same equation) and solve for the remaining 2.* This one-sentence framework communicates clearly what flexibility and restrictions exist in any OS monitoring strategy. It highlights a strategy's focus and associated risks; for example, choosing β and solving for α at an early interim analysis implies a focus on having the trial continue in case the experimental therapy is effective, while understanding the risk of a false positive. And as we shall illustrate, it includes as special cases Fleming et al. (2024), Rodriguez et al. (2024), and other approaches for OS safety monitoring.

Before proceeding, there are three important points that deserve clarification. First, α and β here refer to error rates associated with a single test as opposed to a group-sequential test and/or multiple endpoints. More specifically, they refer to *marginal* type I and type II error rates for hypothesis testing of OS *safety* at interim and final analyses. They should not be confused with *overall* error rates across all analysis times. (Note that Fleming et al. refer to α and β as false positive and false negative error rates. We prefer the terms marginal type I and type II error rates to more clearly distinguish them from overall error rates.) They should also not be confused with error rates for primary and secondary endpoints with a formal testing plan. Whether OS should be formally tested as a safety endpoint and how to do so in relation to other formally tested endpoints is context specific and deserves careful consideration in practice. However, a thorough discussion of this topic is beyond the scope of this paper.

Second, the proposed framework should not be misunderstood as suggesting a single

one-time calculation. Just because a user inputs certain parameters does not mean they should accept whatever the other parameters end up being. All six parameters are important and an iterative process is likely needed to arrive at a configuration that various stakeholders can agree on. The intention of our framework is not to encourage focus on some parameters while ignoring others, but rather to elucidate and encourage stakeholder discussion of the relative priorities and trade-offs between the parameters. Our framework may also help clarify how certain parameters are more informed by data or clinical/scientific considerations compared to other parameters.

Third, the word "choose" should not be misunderstood as suggesting that sponsors are always at liberty to specify the value of a parameter. Similar to the previous note, certain parameter values may be driven by trial conduct, clinical consideration, or regulatory context. For example, in a trial of indolent cancer where timing of the first analysis is based on progression free survival (PFS), the number of OS events at this time cannot be prespecified. It will be what-it-is when the prespecified number of PFS events is observed. At the time of study design, we can input a *predicted* number of OS events to explore operating characteristics of a safety monitoring strategy. However, when the actual trial takes place, the input will need to be updated based on the *observed* number of OS events. If the observed number of OS events is quite different from the predicted, then the entire strategy and configuration may need to be revisited to ensure that stakeholders are realigned on the focus and risks.

Special case #1: Fleming et al. (2024)

Fleming et al.'s OS monitoring guideline can be understood under the proposed framework as choosing $(\theta_0, \theta_1, \beta, d)$ at interim analyses (IAs) and $(\theta_0, \theta_1, \alpha, d)$ at the final analysis (FA). Their approach places more emphasis on trial continuation while evidence is still accumulating, which could make sense in trials with limited OS events at early IAs. This is later counter-balanced by a focus on ruling out OS detriment with reasonably low type I error rate α when more OS events have accumulated.

While representing Fleming et al.'s guideline in such a way makes it clear and simple to understand, it is worth reiterating that users should avoid strictly interpreting parameters

as being either an input or output. Beyond the mechanistic calculations implied by Equations (I)-(II), all parameters are important at each analysis time, whether it's α to protect patients from harm, or β to ensure that trials of effective drugs are not prematurely terminated. Therefore, an iterative process is needed to find a "sweet spot" that balances the various parameter values, and we encourage users to think of Fleming et al. and all other approaches under this framework as having different starting lines, different implications with respect to implementation, and (perhaps slightly) different focuses and tolerances for risk.

Special case #2: Rodriguez et al. (2024)

Rodriguez et al. proposed to calculate the $100 \times (1 - 2\alpha)\%$ confidence interval (CI) for the estimated OS HR, at a time when OS data is relatively mature. Failure to exclude θ_0 with the upper confidence bound is then seen as an indication of OS detriment. Their approach is equivalent to choosing a single $(\theta_0, \theta_1, \alpha, \beta)$ and solving for (d, θ^*) (Shan, 2023). In other words, like traditional hypothesis testing and sample size calculation, Rodriguez et al. use a CI for decision making precisely at a time when a desirable set of operating characteristics can be achieved.

Rodriguez et al. can be generalized to multiple readouts by choosing the same set of four parameters $(\theta_0, \theta_1, \alpha, \beta)$ at each analysis time. (Users will of course have to choose at each time a different set of parameter *values*, e.g., start with a larger α then decrease it over time.) However, in trials with very low OS event rate, it may be difficult if not impossible to reach the required number of events d within a feasible time frame. In addition, timing of the first IA (or first multiple IAs) in trials of indolent cancers is often driven by an intermediate or surrogate endpoint, not OS. There may not be enough OS data at this time to thoroughly evaluate for potential harm. Even if there is, the observed number of deaths will likely differ from the prespecified number d, and it is unclear from this strategy what modifications will be made to address this difference.

Special case #3: Safety monitoring with standard confidence intervals

An alternative approach to Rodriguez et al. (2024) that one might consider is to monitor OS safety using $100 \times (1 - 2\alpha)\%$ CIs again but with flexibility for number of events, i.e., by choosing $(\theta_0, \theta_1, \alpha, d)$ at each analysis time. (The choice between θ_1 or β is flexible and depends on the risk one would like to quantify.)

An example of this approach is an independent data monitoring committee (IDMC) who may not have been provided with detailed instructions for OS safety monitoring and is continually using 95% CIs for evaluation. As commented by Dr. Mikkael Sekeres at the FDA-AACR-ASA workshop, "It is surprising how little guidance you're often given when you're on an IDMC... IDMC should be given good guidance on how to consider overall survival... We think that IDMC should be given specific tasks to look at overall survival at different stages of a trial and also statistical guidance on the confidence with which you can declare that a drug is either too harmful or is safe in moving forward" (American Association for Cancer Research, 2023).

However, it is challenging for 95% CIs to rule out harm early on in a trial when OS data is immature and CIs are wide. Expecting them to do so could result in early termination of an effective treatment. Other standard confidence intervals (e.g., 90% or 80%) may be more appropriate for safety monitoring, but specifying α requires careful planning. In particular, one should consider whether or not to change α if a different number of deaths than d is observed, and if so, how.

Special case #4: Discrete thresholds

Another approach—one that we have seen study teams consider—is to choose $(\theta_0, \theta_1, d, \theta^*)$ at each analysis time, with θ^* being some discrete threshold such as 1.0, 1.1, or 1.2. There are pragmatic reasons for choosing θ^*. For one, a discrete threshold is easy to communicate, understand, and follow. This could be beneficial for an IDMC that is tasked with assessing OS detriment. In addition, the OS HR point estimate is a clinically important result that can have a strong influence on the public's perception of a drug, so there could be a desire to cap it. Exceeding $\theta^* = 1.0$, in particular, implies that OS detriment *was observed* in the

trial overall.

However, it is unclear by choosing $(\theta_0, \theta_1, d, \theta^*)$ what risks one is willing to tolerate and how uncertainty due to small number of events is taken into consideration, if at all. Similar to approach #3, one should consider whether or not θ^* needs to be changed if a different number of deaths than d is observed. $\theta^* = 1.0$ is also not appropriate if θ_1 is close to 1.0, since β will be approximately 0.50 regardless of how many OS events there are.

Special case #5: FDA guidance for evaluating cardiovascular risk in type 2 diabetes (2008)

Our final example is to choose $(\theta_1, \alpha, \beta, d)$ at IAs and $(\theta_0, \theta_1, \alpha, d)$ at the FA. This approach is based on monitoring guidelines that the FDA has used in trials of type 2 diabetes mellitus (T2D) to evaluate cardiovascular risk (FDA Center for Drug Evaluation and Research, 2008), guidelines which also served to inspire approach #1 by Fleming et al.. The key difference between approach #5 and #1 is that, while approach #1 addresses the question "With what certainty can we rule out a particular level of OS detriment?", approach #5 places more emphasis on type I error control at IAs and addresses the question "What OS detriment can we rule out with a high level of certainty?".

It is worth noting though that both approaches #1 and #5 lead to the same calculated thresholds θ^*, because the thresholds are uniquely determined by the parameters θ_1, β, and d. Thus, users do not actually need to choose between the two. Rather, they can simultaneously use both approaches to understand uncertainty and to evaluate benefit-risk. This can be especially useful when a clinically relevant θ_0 has been defined a priori, but few deaths have been observed to enable robust inference with respect to θ_0.

3   Case study

POLARIX was a Phase 3 randomized controlled trial that compared Pola-R-CHP and R-CHOP in patients with untreated large B-cell lymphoma (LBCL). At the time of the primary analysis for PFS and first IA for OS, the trial provided statistically significant evidence of PFS benefit (PFS HR estimate 0.73; 95% CI 0.57 to 0.95), but immature and

inconclusive OS data despite all patients having a minimum of 2 years follow-up (OS HR estimate 0.94, 95% CI 0.67 to 1.33). Subsequently, an FDA ODAC meeting was held to discuss the benefit-risk of Pola-R-CHP, taking into consideration the uncertainty around OS due to low number of deaths. For a detailed account of POLARIX and the ODAC's eventual vote in favor of Pola-R-CHP, we refer the reader to the briefing document (FDA ODAC, 2023).

Fleming et al. (2024) illustrated their OS monitoring guideline by retrospectively applying it to POLARIX (Strategy 1, Table 1). Under our proposed framework, their guideline is equivalent to choosing $(d, \theta_0, \theta_1, \beta) = (89, 1.30, 0.80, 0.1)$ at IA1, $(d, \theta_0, \theta_1, \beta) = (110, 1.30, 0.80, 0.1)$ at IA2, $(d, \theta_0, \theta_1, \beta) = (131, 1.30, 0.80, 0.1)$ at IA3, and $(d, \theta_0, \theta_1, \alpha) = (178, 1.30, 0.80, 0.025)$ at FA. For the rationale behind some of these parameters, we refer the reader to Fleming et al. (2024). We now illustrate and compare the remaining four approaches described in Section 2 by choosing similar parameter values wherever the set of input parameters overlap with Fleming et al.'s, and choosing conventional/intuitive values wherever they do not.

If there is a desire to control both type I and type II error rates, then one approach is to follow Rodriguez et al. (2024) by choosing $(\theta_0, \theta_1, \alpha, \beta)$ over time. Strategy 2 in Table 1 specifies standard type I error rates that decrease over time (0.05 and 0.025) and type II error rate 0.10. The resulting thresholds θ^* are similar to those of Strategy 1 at IA3 and FA. However, the actual timing of IA in POLARIX was driven by PFS, not OS; the trial targeted 228 PFS events, eventually observing 241 PFS events along with 110 OS events. Should POLARIX have assessed for potential OS detriment then, and if so, how? Or should it have waited until 145 and 178 OS events before assessing for potential OS detriment? The latter may seem odd considering that in reality POLARIX formally assessed OS for efficacy at IA under the group sequential framework.

Strategy 3 gives back flexibility on the time of safety monitoring by using 70%, 80%, and 90% CIs to rule out harm early on in POLARIX. It is equivalent to choosing $(d = 89, \theta_0, \theta_1, \alpha = 0.15)$, $(d = 110, \theta_0, \theta_1, \alpha = 0.1)$, $(d = 131, \theta_0, \theta_1, \alpha = 0.05)$, and $(d = 178, \theta_0, \theta_1, \alpha = 0.025)$. The results are more relaxed thresholds and higher probabilities of claiming OS safety when Pola-R-CHP is effective, similar to those of Strategy 1.

Table 1: Various OS monitoring strategies applied to the POLARIX trial and depicted under the proposed framework. For each strategy, the four input or "chosen" parameters at a given stage are indicated in bold. Overall operating characteristics if safety thresholds are strictly adhered to are also provided. The threshold θ^* is considered "met" if the observed OS log hazard ratio falls below it, indicating that the experimental treatment is safe, or at least there is insufficient evidence to claim OS detriment. Strategy 1 is identical to the case study presented by Fleming et al. (2024), minus one potential interim analysis at 60 deaths for simplicity.

OS monitoring strategy	Stage	# of deaths (d)	Treatment effect under H_0 (θ_0)	Treatment effect under H_1 (θ_1)	HR threshold for decision making (θ^*)	Prob. of meeting threshold(s) under H_0 (α)	Prob. of meeting threshold(s) under H_1 ($1-\beta$)
1. Fleming et al. (2024)	IA1	**89**	**1.30**	**0.80**	1.050	0.157	**0.900**
	IA2	**110**	**1.30**	**0.80**	1.021	0.103	**0.900**
	IA3	**131**	**1.30**	**0.80**	1.001	0.067	**0.900**
	FA	**178**	**1.30**	**0.80**	0.969	**0.025**	0.900
	Overall	—	—	—	—	0.017^a	0.819^b
2. Rodriguez et al. (2024)	IA	145	**1.30**	**0.80**	0.990	**0.050**	**0.900**
	FA	178	**1.30**	**0.80**	0.969	**0.025**	**0.900**
	Overall	—	—	—	—	0.020^a	0.869^b
3. Standard CIs	IA1	**89**	**1.30**	**0.80**	1.044	**0.150**	0.895
	IA2	**110**	**1.30**	**0.80**	1.018	**0.100**	0.897
	IA3	**131**	**1.30**	**0.80**	0.975	**0.050**	0.872
	FA	**178**	**1.30**	**0.80**	0.969	**0.025**	0.900
	Overall	—	—	—	—	0.016^a	0.805^b

Table 1 (continued)

OS monitoring strategy	Stage	# of deaths (d)	Treatment effect under H_0 (θ_0)	Treatment effect under H_1 (θ_1)	HR threshold for decision making (θ^*)	Prob. of meeting threshold(s) under H_0 (α)	Prob. of meeting threshold(s) under H_1 ($1-\beta$)
4. Discrete thresholds	IA1	**89**	**1.30**	**0.80**	**1.100**	0.215	0.933
	IA2	**110**	**1.30**	**0.80**	**1.050**	0.131	0.923
	IA3	**131**	**1.30**	**0.80**	**1.000**	0.067	0.899
	FA	**178**	**1.30**	**0.80**	**1.000**	0.040	0.932
	Overall	—	—	—	—	0.026[a]	0.854[b]
5. FDA guidance in T2D (2008)	IA1	**89**	1.59	**0.80**	1.050	**0.025**	**0.900**
	IA2	**110**	1.48	**0.80**	1.021	**0.025**	**0.900**
	IA3	**131**	1.41	**0.80**	1.001	**0.025**	**0.900**
	FA	**178**	1.30	**0.80**	0.969	**0.025**	0.900
	Overall	—	—	—	—	0.017[c]	0.819[b]

[a]Probability of meeting all thresholds (interim and final) under H_0.

[b]Probability of meeting all thresholds (interim and final) under H_1.

[c]Probability of meeting all thresholds (interim and final) under HR = 1.30, the HR under H_0 at FA.

If 95% CIs were to be used instead from IA1 all the way through to FA (not shown in Table 1), then Strategy 3 would result in very stringent thresholds (0.858 with 89 deaths, 895 with 110 deaths, and 0.923 with 131 deaths) and substantially lower probabilities of claiming OS safety when Pola-R-CHP is in fact effective (0.629, 0.721, and 0.793). It is for this reason that we do not recommend stringent control of type I error early on in a trial, since immature OS data and wide 95% CIs can cause unnecessary alarm and result in the wrongful termination of effective treatments.

Strategy 4 considers the use of $\theta^* = 1.1$, 1.05, 1.0, and 1.0 as the HR thresholds for decision making over time. We chose θ^* of 1.1 for IA1 and 1.05 for IA2 based on our observation that study teams have a tendency to be conservative early on in a trial when there is little data. Specific to this case study, Strategy 4 can be seen to have similar operating characteristics as Strategy 1. However, the slightly larger type I error rates of 0.215 and 0.040 at IA1 and FA might be a safety concern. More generally, this strategy does not support a transparent discussion on the various risks because it focuses on the point estimate, which can be problematic from a statistical perspective when there are few deaths.

Strategy 5 calculates what θ_0 can be ruled out at the IAs with $\alpha = 0.025$ and $\beta = 0.1$. We see in this case that $\theta_0 = 1.59$, 1.48, and 1.41 given 89, 110, and 131 deaths, respectively. As noted in Section 2, Strategies 1 and 5 result in the same calculated thresholds θ^*. Therefore, a user can use both strategies at an IA to understand with what level of certainty they can rule out $\theta_0 = 1.30$ ($\alpha = 0.157$ at IA1, $\alpha = 0.1031$ at IA2, and $\alpha = 0.067$ at IA3) and what OS detriment they can rule out with high certainty $\alpha = 0.025$ ($\theta_0 = 1.59$ at IA1, $\theta_0 = 1.48$ at IA2, and $\theta_0 = 1.41$ at IA3). Using Strategy 5 alone though may be problematic, as it doesn't allow a transparent discussion on the risks associated with a clinically relevant target, and may in fact claim a low risk but for a clinically irrelevant question.

Besides the *marginal* risks involved at each stage in a trial, stakeholders should also consider the *overall* risk of the trial should the decision making thresholds θ^* be strictly followed. Table 1 calculates for each strategy the probability of consistently being deemed safe (i.e., probability of meeting all thresholds θ^*) under H_0 and under H_1. It can be seen

that all strategies have a high probability ($> 97\%$) of being flagged as unsafe at least once (i.e., failing to meet a threshold at least once) during trial conduct if the drug is harmful. They also have a $>80\%$ probability of consistently being deemed safe if the drug is effective.

For Strategies 1, 3, 4, and 5, the family wise error rates (i.e., probability of being deemed safe at least once under H_0) are between 0.17 to 0.24 (not shown in Table 1). If there is a desire to control these rates at a lower level, say 0.05 or 0.025, then one can expect a less than 80% probability of consistently being deemed safe under H_1—a level of risk that sponsors might be unwilling to take. In the context of safety monitoring where interest is in flagging a treatment that causes OS detriment, we think the probability of consistently being deemed safe under H_0—or more precisely, its complement, the probability of being flagged as unsafe at least once if the drug is harmful—is a more meaningful quantity than the FWER.

4   Discussion

With advancements in anticancer therapy, it has become increasingly difficult for clinical trials in certain disease and clinical contexts to power for OS as an efficacy endpoint. Nevertheless, OS is an important endpoint for assessing safety, and rigorous benefit-risk assessment of novel therapies requires improved prospective planning, collection, and assessment of OS data. In this paper, we proposed a framework for the pre-specified analysis of OS as a safety endpoint (Section 2). Our framework solicits users to choose, at each analysis time, 4 out of 6 parameters from classical equations for statistical hypothesis testing. In doing so, users communicate clearly their focus, priority, and tolerance for risk over the course of a trial.

We described five special cases under the proposed framework: Fleming et al. (2024), Rodriguez et al. (2024), an approach that uses standard confidence intervals to continually rule out harm, another approach that uses discrete thresholds for decision making, and a fifth approach that is based off of previous FDA monitoring guidelines for evaluating cardiovascular risk in T2D. We illustrated use of the five approaches by applying them to the POLARIX trial (Section 3). While the main objective of our paper is to increase understanding of each approach, thereby facilitating planning by sponsors of future clinical

trials and improving communication between stakeholders, we leave here with some final comments and recommendations regarding their use.

Of the five approaches mentioned in this paper, we think Fleming et al.'s is the most appropriate for general use due to its balance in priority over time—which shifts from a focus on ensuring that potentially effective treatments will continue to be studied when there are few OS events, to a focus on ruling out OS detriment when more OS events have accumulated. This approach also has the advantage of making clear how decision making thresholds can be calculated at any time given the observed number of OS events.

The fifth approach based on FDA's guidance in T2D can be used in tandem with Fleming et al. to better evaluate benefit-risk. Together, when deaths are still slowly accruing, the two approaches address the questions, "What OS detriment can we rule out with a high level of certainty?" and "With what certainty can we rule out a particular level of OS detriment?".

Rodriguez et al. can be used to identify the goal of a clinical trial, the point at which OS data can be considered sufficiently mature to facilitate evaluation of OS as a safety endpoint. However, it provides no guidance on how to evaluate OS at other time points, including potential interim analyses driven by the primary endpoint. Likewise, 95% CIs provide little utility for ruling out harm early on in a trial.

The other limitation of Rodriguez et al.—the limitation that planned vs. observed number of deaths are typically different—can be addressed by extending the authors' approach within our proposed framework. At the design stage, sponsors can use Rodriguez et al. to define and predict the number of PFS and OS events at an IA, seeking a balance between properly powering PFS and observing sufficiently mature OS. Subsequently, a threshold θ^* can be calculated based on the designed d. At the IA, if the observed number of deaths d' is different from the designed d, then one can use $(\theta_0, \theta_1, \alpha, d')$ or $(\theta_0, \theta_1, \beta, d')$ to update θ^* depending on what risk one could like to quantify. This is similar to the concept of group sequential designs, where stopping boundaries are updated prior to formal readouts based on the observed numbers of deaths.

More relaxed CIs (90%, 80%, etc.) or discrete thresholds may be appropriate provided that the risks involved are understood and acceptable. In fact, one can apply Fleming et al.
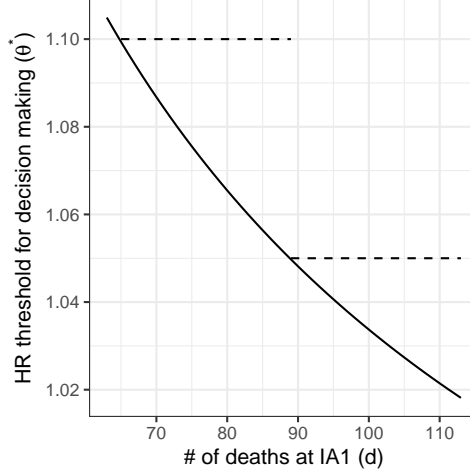
Figure 1: Fleming et al. (2024) and an approximately equivalent discrete threshold strategy applied to POLARIX at the first planned interim analysis with 89 expected OS events. Solid curve is the hazard ratio (HR) threshold for decision making based on Fleming et al. (2024) with $\theta_0 = 1.3$, $\theta_1 = 0.80$, $1 - \beta = 0.9$, and varying potential number of observed deaths d. Dashed curve indicates discrete thresholds that approximate Fleming et al. (2024) by maintaining $1 - \beta \geq 0.9$ while tolerating higher levels of α.

as a first approximation, then find an approach using standard CIs or discrete thresholds that has similar operating characteristics. Such an approach could effectively balance priorities over time *and* be easy to implement during a trial. Sponsors taking this approach would do well to pre-specify when and how to adjust the α-level or discrete threshold in case the number of events deviates significantly from expectations during trial conduct (see Figure 1 for example).

Regardless of the approach taken, the low mortality rate in indolent diseases makes rigorous OS monitoring challenging. We agree with Fleming et al. (2024) that any set of proposed thresholds should be seen as non-binding "informative guidelines" as opposed to black-and-white "rules" for continuing or stopping a trial. At a minimum, the need to monitor OS for safety should encourage stakeholders to contemplate the level of OS data that is likely to be accessible during the course of a clinical trial. Meanwhile, the additional analyses suggested by Rodriguez et al. (2024) may be helpful for gaining a better understanding of immature OS data: analysis of biologically plausible subgroups, simulation and

supplementary analyses to investigate the robustness of results under different assumptions, evaluation of other endpoints, results from other trials, use of real-world data, etc.. IDMCs may also find measures of stability or graphical summaries such as the predicted interval plot useful for understanding potential effect size estimates and associated precision with trial continuation (Betensky, 2015; Li et al., 2009).

Several questions remain to be addressed for the practical implementation of the proposed framework:

- How should θ_0 be defined, taking into account the clinical context and risk tolerance of different stakeholders? For example, patients or clinicians might tolerate greater uncertainty in OS data if a robust clinical benefit is ascertained on an intermediate/surrogate endpoint.

- How should θ^* be defined given potential confounding by intercurrent events? If a substantial number of patients switch to subsequent therapy after tumor progression, then it may be desirable to relax θ^* since subsequent therapies can lead to a smaller OS difference (Korn et al., 2011).

- When should OS be evaluated for safety? Should it only be evaluated when OS is sufficiently mature (however we choose to define "mature")? Or should it also be evaluated at key timepoints (e.g., interim and final analyses of the primary endpoint)?

Another question that may come up is regarding non-proportional hazards (NPH). Is the proposed framework applicable under NPH? Our opinion is yes, that at least a good part of the framework can be applied *if* the overall HR is still considered a meaningful treatment effect summary; for example, analysis timing can still be based on the number of deaths d and θ^* can still be used as the threshold for signaling overall OS detriment. However, if it is known a priori that a NPH effect could exist, then operating characteristics should be evaluated carefully under potential patterns of NPH in order to ensure that whatever plan is proposed for OS safety analysis balances risk appropriately over time. Furthermore, supportive analyses such as piecewise exponential and cure rate models can be prospectively planned to help interpret trial results. This is similar to how some sponsors currently handle NPH for efficacy analysis. They conduct event-driven analyses using the logrank test and

cox HR, but consider potential NPH when calculating the required number of events. For example, see Checkmate-274 and Keynote-966 (Bajorin et al., 2021; Kelley et al., 2023).

This paper proposed a supportive decision making framework based on a single criteria θ^*. It is worth noting that Rodriguez et al. (2024) considered having a 'grey zone', which, if a trial were to fall into, would prompt further OS follow-up. Doing so could have the effect of ruling out substantial OS detriment while limiting rejection of treatments with marginal OS benefit. While we did not consider such an extension here for simplicity of comparison and presentation, we do find this idea of practical interest worth pursuing in future discussions and research.

In conclusion, it is in everyone's interest—patients, clinicians, regulators, and sponsors alike—for safe and effective cancer therapies to be developed in a timely manner. However, the low mortality rate in some disease settings can make it difficult to rigorously evaluate benefit-risk. There is a present need for stakeholders to come together, to communicate, and to align on the expectations and priorities of clinical trials with respect to the generation and interpretation of OS data. By simplifying the presentation of recent approaches for OS safety analysis and clarifying their focus, priority, and tolerance for risk over time, we believe the "6 choose 4" framework proposed in this paper can help to meet those needs.

Acknowledgements

Disclosure statement

References

American Association for Cancer Research (2023), 'FDA-AACR-ASA Workshop: Overall Survival In Oncology Clinical Trials'. Available at: https://www.aacr.org/professionals/policy-and-advocacy/regulatory-science-and-policy/events/fda-aacr-asa-workshop-overall-survival-in-oncology-clinical-trials/.

Bajorin, D. F., Witjes, J. A., Gschwend, J. E., Schenker, M., Valderrama, B. P., Tomita, Y., Bamias, A., Lebret, T., Shariat, S. F., Park, S. H., Ye, D., Agerbaek, M., Enting, D., McDermott, R., Gajate, P., Peer, A., Milowsky, M. I., Nosov, A., Antonio, J. N., Tupikowski, K., Toms, L., Fischer, B. S., Qureshi, A., Collette, S., Unsal-Kacmaz, K., Broughton, E., Zardavas, D., Koon, H. B. and Galsky, M. D. (2021), 'Adjuvant nivolumab versus placebo in muscle-invasive urothelial carcinoma', *New England Journal of Medicine* **384**, 2102–2114.

Betensky, R. A. (2015), 'Measures of follow-up in time-to-event studies: Why provide them and what should they be?', *Clinical Trials* **12**, 403–408.

FDA Center for Drug Evaluation and Research (2008), 'Guidance for industry. Diabetes mellitus – evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes'. Available at: https://downloads.regulations.gov/FDA-2008-D-0118-0029/content.pdf.

FDA ODAC (2023), 'March 9, 2023. Meeting of the Oncologic Drugs Advisory Committee – Combined FDA and Genentech Briefing Document'. Available at: https://www.fda.gov/media/165961/download.

Fleming, T. R., Hampson, L. V., Bharani, B.-D., Bretz, F., Chakravartty, A., Coroller, T., Koukouli, E., Wittes, J., Yateman, N. and Zumber, E. (2024), 'Monitoring overall survival in pivotal trials in indolent cancers', *Statistics in Biopharmaceutical Research* . doi: 10.1080/19466315.2024.2365648.

Kelley, R. K., Ueno, M., Yoo, C., Finn, R. S., Furuse, J., Ren, Z., Yau, T., Klümpen, H. J., Chan, S. L., Ozaka, M., Verslype, C., Bouattour, M., Park, J. O., Barajas, O., Pelzer,

U., Valle, J. W., Yu, L., Malhotra, U., Siegel, A. B., Edeline, J., Vogel, A., Akce, M., Diaz, I. A., Alves, G., Anand, S., Arslan, C., Asselah, J., Assenat, E., Aubin, F., Bai, L. Y., Bai, Y., Bates, S., Begbie, S., Ben-Aharon, I., Beri, N., Berres, M. L., Blanc, J. F., Borbath, I., Bordonaro, R., Brandi, G., Burgoyne, A., Butthongkomvong, K., Cao, K., Carballido, M., Camandaroba, M., Chang, S. L., Chen, J. S., Chen, M. H., Chen, X., Cheng, A., Chiu, T. J., Choi, H. J., Chon, H. J., Collignon, J., Gracian, A. C., Davis, S., de Carvalho, R. S., de Groot, D. J., Demols, A., Vos, J. D., Diab, M., Easaw, J., Eatock, M., Elias, R., Eskens, F., Falcone, A., Fernandez, P., Finn, R., Franke, F., Furukawa, M., Gbolahan, O., Geboes, K., Geneser, K. L., Geng, Z., Geva, R., Gillmore, R., Goetze, T., Gou, H., Grasselli, J., Gu, S., Gumus, M., Mohammad, N. H., Hao, C., Harputluoglu, H., Hatoum, H., Heinemann, V., Ho, W. K., Hsu, C., Hubert, A., Hwang, J., Inanc, M., Iseas, S., Jeyasingam, V., Fonseca, P. J., Joubert, W., Juengsamarn, J., Kaen, D., Kanai, M., Kasper-Virchow, S., Kazemi, G., Kelleher, F., Kelley, R., Kim, J. W., Kim, J. G., Abrahao, A. B. K., Klumpen, H., Kochenderfer, M., Kose, F., Lam, H. C., kun Lee, C., Lee, H. W., Lee, M., Lee, M. A., Lee, W. M. S., Sourd, S. L., Li, D., Li, W., Liang, H., Liang, T., Lim, C. S., Lingerfelt, B., Lopez, C., Low, J., Mercade, T. M., Malka, D., Mao, Y., Masi, G., McCune, S., McDermott, R., McWhirter, E., Mendez, G., Milella, M., Mizutani, T., Moniz, C., Morales, L., Martin, A. J. M., Nervi, B., Ngamphaiboon, N., Oh, S. C., Oksuzoglu, B., Outlaw, D., Ozguroglu, M., Ozyilkan, O., Painemeal, C., Pan, Y., Peng, C., Petorin, C., Pezet, D., Power, D., Qin, S., Roohullah, A., Ryu, H., Salman, P., Sasidharan, R., Satho, T., Schulze, K., Scott-Brown, M., Segovia, R., Seufferlin, T., Siena, S., Sinapi, I., Smolenschi, C., Song, T., Sookprasert, A., Soparattanapaisarn, N., Starling, N., Stein, S., Stemmer, S., Su, H., Sugimoto, R., Suksombooncharoen, T., Tam, V., Tan, A. L., Tan, C. K., Tanasanvimon, S., Tonini, G., Tortora, G., Tsuji, A., Uribe, R., Venerito, M., Mata, H. V., Victorino, A. P., Wade, J., Waldschmidt, D. T., Wang, L., Isahk, W. Z. W., Wasan, H., Weschenfelder, R., Wong, C. Y., Wong, Y. F., Yalcin, S., Weber, P. Y., Yang, X., Yasui, H., Yazici, O., Yen, C. J., Ying, J., Yu, W. and Zhao, H. (2023), 'Pembrolizumab in combination with gemcitabine and cisplatin compared with gemcitabine and cisplatin alone for patients with advanced biliary tract cancer (keynote-966): a randomised, double-blind, placebo-controlled, phase 3 trial', *The*

Lancet **401**, 1853–1865.

Korn, E. L., Freidlin, B. and Abrams, J. S. (2011), 'Overall survival as the outcome for randomized clinical trials with effective subsequent therapies', *Journal of Clinical Oncology* **29**, 2439–2442.

Li, L., Evans, S. R., Uno, H. and Wei, L. (2009), 'Predicted interval plots (PIPS): A graphical tool for data monitoring of clinical trials', *Statistics in Biopharmaceutical Research* **1**, 348–355.

Merino, M., Kasamon, Y., Theoret, M., Pazdur, R., Kluetz, P. and Gormley, N. (2024), 'Irreconcilable differences: The divorce between response rates, progression-free survival, and overall survival', *Journal of Clinical Oncology* **41**, 2706–2713.

Rodriguez, L. R., Gormley, N. J., Lu, R., Amatya, A. K., Demetri, G. D., Flaherty, K. T., Mesa, R. A., Pazdur, R., Sekeres, M. A., Shan, M., Snapinn, S., Theoret, M. R., Umoja, R., Vallejo, J., Warren, N. J. H., Xu, Q. and Anderson, K. C. (2024), 'Improving collection and analysis of overall survival data', *Clinical Cancer Research* . doi: 10.1158/1078-0432.CCR-24-0919.

Schoenfeld, D. (1981), 'The asymptotic properties of nonparametric tests for comparing survival distributions', *Biometrika* **68**, 316–319.

Shan, M. (2023), 'Assessment of OS data for safety as pre-specified endpoint for trials in indolent/early-stage cancers'. Presented at FDA-AACR-ASA Workshop. Available at: https://www.aacr.org/wp-content/uploads/2023/07/Session-2- Slides.pdf.