Bias, Operational Bias, and Generalizability in Phase II/III Trials

Boris Freidlin, *National Cancer Institute, Bethesda, MD*
 Edward L. Korn, *National Cancer Institute, Bethesda, MD*
 Jeffrey S. Abrams, *National Cancer Institute, Bethesda, MD*

Background

Adaptive clinical trial designs improve the efficiency of the drug development process by allowing treatment decisions to be made earlier, while protecting the interests of the patients in the trials. In settings in which a randomized phase II efficacy signal is required to justify starting a large phase III trial, a phase II/III (seamless) design can sometimes be used to adaptively combine randomized phase II and phase III trial components.¹⁻⁴ After the phase II part of the trial is enrolled and the phase II outcome is sufficiently mature, a go/no-go decision is made regarding whether to continue accrual to the phase III part of the trial. The phase II patients are used in the final phase III analysis, and a separate phase III protocol need not be developed. This design offers efficiency gains relative to performing an independent phase III trial after a positive phase II trial. In general, the phase II analysis is performed confidentially solely to make a go/no-go decision, and the phase II results are not released to the public.

When the phase II outcome events (eg, progression-free survival [PFS]) are occurring slowly, or when the same end point (eg, overall survival [OS]) is used for the phase II and the phase III analyses, an accrual suspension while waiting for the phase II data to mature may be needed to minimize patient exposure to an ineffective therapy and reap the efficiency benefits of the design.² Although suspending/restarting accrual in a phase III trial represents a logistical challenge, this approach is being used in the National Cancer Institute National Clinical Trials Network trials (eg, NCT02502266: Cediranib Maleate and Olaparib or Standard Chemotherapy in Treating Patients With Recurrent Platinum-Resistant or -Refractory Ovarian, Fallopian Tube, or Primary Peritoneal Cancer; NCT02339571: Nivolumab and Ipilimumab With or Without Sargramostim in Treating Patients With Stage III-IV Melanoma That Cannot Be Removed by Surgery; NCT02152982: Temozolomide With or Without Veliparib in Treating Patients With Newly Diagnosed Glioblastoma Multiforme; and NCT01810913: Radiation Therapy With Cisplatin, Docetaxel, or Cetuximab After Surgery in Treating Patients With Stage III-IV Squamous Cell Head and Neck Cancer). However, the US Food and Drug Administration (FDA) recently raised a new concern regarding the use of phase II/III designs when it responded to a question about whether phase II patients could be used in the final phase III analysis of a planned trial that had an accrual suspension (End of Phase 2 Meeting Minutes, Reference ID 4096437):

“No, as currently designed this [an accrual suspension] would not be appropriate. In general, an adaptive design may be acceptable if the trial is well designed. In the currently proposed design, the potential 18-month delay from the time of completion of enrollment in the ‘Phase II’ portion of the trial and initiation of the ‘Phase III’ portion may introduce bias. FDA recommends that DCTD [Division of Cancer Treatment and Diagnosis] consider modifying the trial design to a single randomized trial with an interim analysis of PFS [progression-free survival] for futility or for sample size re-estimation with proper alpha adjustment.”

In this commentary, we dissect and evaluate the FDA's concern and find that it does not sufficiently justify a general prohibition of using phase II/III trial designs with accrual suspension.

The Proposed Trial and Alternative Trial Designs

The phase II/III trial under consideration specified the maximum sample size of 468 patients (randomly assigned at a ratio of 2:1 between the experimental and control arms) on the basis of a phase III design targeting an OS hazard ratio (HR) of 0.73 (80% power with a 0.025 one-sided significance level), with the final analysis scheduled to occur after 364 deaths occurred. The phase II design was based on randomly assigning 171 patients targeting a PFS HR of 0.65 (85% power with a 0.15 one-sided significance level), with the phase II go/no-go decision scheduled for when 98 PFS events have been observed. To allow the phase II data to mature, the design specified an accrual suspension after the first 171 patients had been randomly assigned until the go/no-go decision was made, which was expected to take approximately 18 months. This was deemed advisable because the PFS events were expected to occur slowly, with the median PFS being 18 months.

The first and second rows of [Table 1](#) describe the operating characteristics of the proposed design and a phase II/III trial design with no accrual suspension. To address the FDA suggestion of using a design that incorporates PFS-driven futility analyses, the third row presents an OS-driven phase III design that includes PFS futility analyses (as described in Goldman et al⁵). (Presumably, this is what was meant by the FDA's suggestion of a single trial with futility monitoring for PFS.) The difference between the designs in rows 1 and 2 versus the design in row 3 is that the phase II bar for PFS in a phase II/III design will be higher than a futility monitoring bar for PFS, the former requiring some designated benefit of the experimental treatment over the control treatment. The three

Table 1. Comparison of a Phase II/III Trial Design (with and without accrual suspension) and a Phase III Design With PFS/OS Interim Monitoring

Design	Null Hypothesis		Alternative Hypothesis	
	Average Sample Size	Average Duration (months)	Average Sample Size	Average Duration (months)
Proposed design: phase II/III design with accrual suspension*	217	38	422	77
Phase II/III design with no accrual suspension†	296	32	442	61
FDA suggestion: phase III design with PFS (and OS) interim monitoring‡	450	47	467	61

NOTE. For all designs, accrual is assumed to be 10 patients per month; median overall survival (OS) for control group is 22 months.

Abbreviations: FDA, US Food and Drug Administration; PFS, progression-free survival.

*In the phase II portion after 171 patients are enrolled, accrual is suspended for approximately 18 months until 98 PFS events have occurred; this provides the phase II analysis with 85% power for detecting a PFS hazard ratio (HR) of 0.65 at the one-sided 0.15 level. If the phase II threshold is passed, then the trial enrolls a maximum of 468 patients, with the final analysis performed when 364 deaths have occurred (using a group sequential design with three OS interim analyses for efficacy and futility at 25%, 50%, and 75% of the total number of deaths). This design provides 80% nominal power (73% overall power accounting for the phase II rule) for detecting an OS HR of 0.73 at the one-sided 0.025 level.

†There is no accrual suspension in the phase II portion. The phase II analysis takes place when 98 PFS events have occurred; this provides the phase II analysis with 85% power for detecting a PFS HR of 0.65 at the one-sided 0.15 level. If the phase II threshold is passed, then the trial enrolls a maximum of 468 patients, and the final analysis is performed when 364 deaths have occurred (using a group sequential design with three OS interim analyses for efficacy and futility at 25%, 50%, and 75% of the total number of deaths). This design provides 80% nominal power (73% overall power accounting for the phase II rule) for detecting an OS HR of 0.73 at the one-sided 0.025 level.

‡Phase III trial design with OS end point similar to the design described in previous footnotes (using a group sequential design with three OS interim analyses for efficacy and futility at 25%, 50%, and 75% of the total number of deaths) but with overall power of 73% (requiring 328 deaths with the same maximum sample size of 468 patients as in the phase II/III design). Seventy-three percent power is used to match the overall power of the phase II/III design. PFS futility analyses were added to this design at the same times as the OS interim analyses, with a stopping recommendation if the PFS HR of 0.73 could be rejected at the one-sided 0.0025 level (as described in Goldman et al⁶).

designs were simulated under the null hypothesis and the targeted alternative hypothesis (see Table 1 footnotes for details), with average sample sizes and average durations tabulated for each design.

When the experimental treatment is not beneficial (the null hypothesis), FDA's proposed design has poor characteristics compared with the phase II/III designs: a much larger average sample size and a longer average duration (Table 1). A comparison of the two phase II/III designs demonstrates that the suspension results, on average, in a longer trial (by 6 months) but with 25% fewer patients exposed to an ineffective therapy, which is the *raison d'être* for the suspension. Under the alternative hypothesis, the average sample sizes are about the same, with the suspension leading to a longer trial when compared with a design without a suspension.

Bias

In terms of protecting patients from ineffective therapies, the phase II/III design with a suspension is the best choice. However, if the trial design introduced a bias (a systematic deviation between the estimated treatment effect and its true value) that made the results of such a design less interpretable, then its use would be unjustified. For example, repeated analyses of interim data introduce bias (unless a proper multiplicity adjustment is used). A more subtle example of a design-induced bias is given by trials that use outcome-adaptive randomization (changing the randomization ratio during the study to favor the arm that appears to be doing better). If the prognosis of the patient population is improving over the course of the trial, a standard comparison of the treatment arms can make the experimental-treatment arm appear better than it truly is.⁶

Interim monitoring in standard phase III and phase II/III designs is known to result in a slight inflation of the estimated treatment effect. With a properly designed interim monitoring plan, the inflation is quite small^{7,8} and is typically ignored. (Note

that this inflation is unrelated to whether the design has an accrual suspension.) Furthermore, if the experimental and control treatments are equally effective for all patients, then a randomized phase II/III trial will preserve the type I error rate. Interestingly, the sample size re-estimation designs suggested by the FDA in the above comment can lead to bias.⁹⁻¹¹

Operational Bias and Generalizability

Perhaps the bias the FDA had in mind was operational bias, which is defined in "Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics"¹²:

"Operational Bias: Many study adaptations call for unblinding of the analysts charged with implementing the planned design revisions. Access by these analysts to the interim unblinded results raises concern about the possibility that the analysts might influence investigators in how they manage the trial, manage individual study patients, or make study assessments, bringing into question whether trial personnel have remained unequivocally objective."

First, consider concerns that the operational deviations that result from the knowledge of interim results can lead to error in estimating the true treatment effect in the study population. The relevant issue is not whether there is potential for this type of operational bias in phase II/III trials, but rather whether phase II/III designs increase potential for this bias relative to a standard phase III trial design. In both designs, the trial statisticians will have confidential access to unblinded interim results. In phase II/III designs (with or without accrual suspension), continuing to the phase III stage makes it more generally known that the phase II bar has been passed. (Note that there is no access to the actual phase II results.) However, in a standard phase III trial, continuing accrual makes it more generally known that the futility boundary has been passed. More importantly, in the settings in which phase II/III designs are conducted, the appropriate comparator for the phase

II/III design is a randomized phase II trial followed by a phase III trial. In this situation, the phase III trial is conducted with full knowledge of the phase II results, so all the concerns about potential for operational bias resulting from the investigators' and patients' knowledge of the phase II results would be applicable to the phase III trial. Actually, the concerns would be higher for the phase II trial followed by a separate phase III trial because, unlike in phase II/III design, the detailed results of the phase II analysis would be known.

Another FDA concern with the proposed design was the effect of temporal population trends on the results because of the suspension (End of Phase 2 Meeting Minutes, Reference ID 4096437):

"FDA stated that the primary concern is that as science and/or standard of care evolve over time, an 18-month pause in enrollment may result in a shift in patient characteristics between patients enrolled before the pause in accrual and patients enrolled after enrollment is re-initiated after analysis of PFS for the 'Phase II' portion of the study."

This concern is about the generalizability of the trial results. Although it is a legitimate consideration in interpreting trial results, this concern would similarly apply to a standard phase III trial that took an extra 18 months to complete its enrollment because of slow accrual. In any event, differences in the patient population accrued over 47 months instead of 65 months are probably swamped by the difference between patients who participate in clinical trials and the general patient population, a difference investigators acknowledge and reluctantly accept.

Conclusion

When there is not enough existing evidence to warrant conducting a phase III trial, a phase II/III trial design can lead to finding improved treatments more quickly than performing a phase II trial followed by an independent phase III trial. In some situations, however, an accrual suspension will be necessary to avoid accruing most, or all, of the phase III patients while waiting for the phase II data to mature. Having an accrual suspension in a phase II/III study does not increase potential for a biased estimate of the treatment effect (relative to standard trial designs). Theoretical concerns that the suspension will lead to additional generalizability difficulties do not seem to be of a magnitude of practical importance. There are well-known drawbacks to using

phase II/III designs,^{2,3,12-14} and they are not appropriate for all situations, but whether they have an accrual suspension should not be a deciding factor.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Disclosures provided by the authors are available with this article at jco.org.

AUTHOR CONTRIBUTIONS

Manuscript writing: All authors

Final approval of manuscript: All authors

REFERENCES

1. Bretz F, Schmidli H, König F, et al: Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biom J* 48: 623-634, 2006
2. Korn EL, Freidlin B, Abrams JS, et al: Design issues in randomized phase II/III trials. *J Clin Oncol* 30:667-671, 2012
3. Wang M, Dignam JJ, Zhang QE, et al: Integrated phase II/III clinical trials in oncology: A case study. *Clin Trials* 9:741-747, 2012
4. Redman MW, Goldman BH, LeBlanc M, et al: Modeling the relationship between progression-free survival and overall survival: The phase II/III trial. *Clin Cancer Res* 19:2646-2656, 2013
5. Goldman B, LeBlanc M, Crowley J: Interim futility analysis with intermediate endpoints. *Clin Trials* 5:14-22, 2008
6. Korn EL, Freidlin B: Outcome: Adaptive randomization—Is it useful? *J Clin Oncol* 29:771-776, 2011
7. Goodman SN: Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Ann Intern Med* 146:882-887, 2007
8. Korn EL, Freidlin B: Stopping clinical trials early for benefit: Impact on estimation. *Clin Trials* 6:119-125, 2009
9. Bauer P, Posch M: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections, by H. Schäfer and H.-H. Müller. *Stat Med* 23:1333-1334, 2004
10. Magirr D, Jaki T, Koenig F, et al: Sample size reassessment and hypothesis testing in adaptive survival trials. *PLoS One* 11:e0146465, 2016
11. Freidlin B, Korn EL: Sample size adjustment designs with time-to-event outcomes: A caution. *Clin Trials* 14:597-604, 2017
12. U.S. Department of Health and Human Services, Food and Drug Administration: Guidance for industry: Adaptive design clinical trials for drugs and biologics [draft for comment]. Rockville, MD, 2010 <https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf>
13. Emerson SS, Fleming TR: Adaptive methods: Telling "the rest of the story". *J Biopharm Stat* 20:1150-1165, 2010
14. Cuffe RL, Lawrence D, Stone A, et al: When is a seamless study desirable? Case studies from different pharmaceutical sponsors. *Pharm Stat* 13:229-237, 2014

DOI: <https://doi.org/10.1200/JCO.2017.77.0479>; published at jco.org on April 26, 2018.

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

Bias, Operational Bias, and Generalizability in Phase II/III Trials

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Boris Freidlin

No relationship to disclose

Jeffrey S. Abrams

No relationship to disclose

Edward L. Korn

No relationship to disclose