

## Phase III Clinical Trials That Integrate Treatment and Biomarker Evaluation

Boris Freidlin, Zhuoxin Sun, Robert Gray, and Edward L. Korn

Boris Freidlin and Edward L. Korn, National Cancer Institute, Bethesda, MD; and Zhuoxin Sun and Robert Gray, Eastern Cooperative Oncology Group, and Dana-Farber Cancer Institute, Boston, MA.

Published online ahead of print at [www.jco.org](http://www.jco.org) on April 8, 2013.

Supported in part by Grant No. CA023318 from the National Institutes of Health.

Authors' disclosures of potential conflicts of interest and author contributions are found at the end of this article.

Corresponding author: Boris Freidlin, PhD, Biometric Research Branch, EPN-8129, National Cancer Institute, Bethesda, MD 20852; e-mail: [freidlinb@ctep.nci.nih.gov](mailto:freidlinb@ctep.nci.nih.gov).

© 2013 by American Society of Clinical Oncology

0732-183X/13/3125w-3158w/\$20.00

DOI: 10.1200/JCO.2012.48.3826

### INTRODUCTION

Many new anticancer agents are molecularly targeted and therefore may only benefit a subgroup of a histologically defined population. Development of these agents relies on biomarkers that identify a sensitive subpopulation (ie, the assumption is made that a binary biomarker can classify patients into two subgroups so that the agent works in one subgroup but not in the other). If, at the time the definitive phase III trial is being designed, there is confidence in the ability of the biomarker to identify patients who will benefit, then the study is limited to the biomarker-positive population (ie, enrichment trial design<sup>1</sup>). If the biomarker is well developed early, then the phase II assessment of the agent can help decide whether an enrichment phase III trial is appropriate.<sup>2</sup> However, some degree of uncertainty about whether the treatment benefit extends to biomarker-negative patients is often present. In this case, a phase III trial should integrate treatment and biomarker evaluation. Accordingly, many recent integrated phase III trials have used various strategies to test for benefit in the biomarker-positive subgroup, the biomarker-negative subgroup, and the overall population. This article focuses on the fact that some of these integrated trial designs have a high risk of incorrectly recommending treatment for biomarker-negative patients when the treatment has no benefit in that subgroup. Comprehensive reviews of biomarker trial designs are given elsewhere.<sup>3-6</sup>

### INTEGRATED DESIGNS BASED ON EVALUATION OF BIOMARKER-POSITIVE AND -NEGATIVE SUBGROUPS

A clinically useful predictive biomarker differentiates the patient population into a subgroup that benefits from the therapy versus the remaining population that does not. Integrated phase III designs should provide reliable assessment of the risk-to-benefit ratio in each of the biomarker-defined subgroups to allow treatment recommendations for each subgroup. A direct way to address this is to use a parallel subgroup-specific design that is powered

to detect clinically relevant treatment effects separately in the biomarker-positive and -negative populations (design 1). To have an overall type-I error of  $\alpha$ , one must allocate  $\alpha$  between the tests of the null hypotheses of no treatment effect in each of the subgroups. For example, for an overall  $\alpha = 0.025$  error, one could use  $\alpha = 0.015$  for the biomarker-positive subgroup and  $\alpha = 0.01$  for the biomarker-negative subgroup (all  $\alpha$  levels in this presentation are one sided).

An improvement in design efficiency can be achieved by using a sequential testing procedure that uses the assumption that the new treatment will be unlikely to be effective in the biomarker-negative patients unless it is effective in the biomarker-positive patients.<sup>7,8</sup> First, test the biomarker-positive patients using the significance level  $\alpha$ ; if this test is significant, then test the biomarker-negative patients using the same  $\alpha$  (design 2). This design preserves the overall error  $\alpha$  but allows for a smaller sample size than design 1, which splits the  $\alpha$ .

### INTEGRATED DESIGNS BASED ON EVALUATION OF OVERALL POPULATION AND BIOMARKER-POSITIVE SUBGROUPS

An alternative, indirect approach to integrating treatment and biomarker evaluation is by formally assessing treatment benefit in the overall population and in the biomarker-positive patients but not in the biomarker-negative patients. A parallel version of this overall/biomarker-positive design tests both the overall and the biomarker-positive populations simultaneously (design 3).<sup>9-12</sup> With this approach, the degree of confidence in the predictive value of the biomarker determines whether the required trial sample size is dictated by the biomarker-positive or overall population questions. If there is high confidence, then the sample size is determined by the requirement to have a sufficient number of biomarker-positive patients to detect the expected benefit. Conversely, if the confidence in the biomarker predictive value is limited, then the sample size is determined by the requirement to have a sufficient number of patients to detect a clinically relevant benefit in the overall population. To minimize the sample size, more of the  $\alpha$  may be allocated

to the test that is determining the sample size. For example, in the S0819 trial,<sup>10</sup> which is evaluating cetuximab in non–small-cell lung cancer (NSCLC), there was evidence that the benefit could be limited to the epidermal growth factor receptor (EGFR) –positive subgroup. Therefore, the design ensured a sufficient number of EGFR-positive patients to assess the treatment in that subgroup and allocated 80% of the  $\alpha$  to that test.<sup>10</sup>

Note that the design recommends the treatment for all patients if the overall assessment is positive, even if the test in the biomarker-positive group fails to show efficacy (Table 1, design 3). This may seem counterintuitive, but it is reasonable because the biomarker-positive subgroup may be too small to identify a moderate treatment effect that is uniformly present in both the biomarker-positive and -negative subgroups.

Some studies<sup>3,13</sup> have employed a sequential version of the overall/biomarker-positive design when there is confidence in the predictive value of the biomarker. First, test the biomarker-positive subgroup using significance level  $\alpha$ ; if the test is significant, then test the overall population using the same  $\alpha$  (design 4). This sequential procedure preserves the study-wise error  $\alpha$  and allows for a smaller sample size than the parallel assessment. However, this procedure is

problematic for determining whether the treatment is beneficial in the biomarker-negative subgroup.

Sometimes, a fallback procedure is suggested that first tests the overall population and then, if the overall test is negative, tests the biomarker-positive population (design 5). In terms of sample sizes and study outcomes, this approach is identical to design 3. However, this approach could be useful in settings where a biomarker will be developed or assessed only if the overall population results are negative.<sup>14</sup>

**CHOICE OF INTEGRATED DESIGNS**

A major concern with the overall/biomarker-positive designs is that when the benefit of the new therapy is limited to the biomarker-positive patients, it is possible that the design may inappropriately lead to a recommendation for treatment of the biomarker-negative patients.<sup>15</sup> This is because when there is no treatment effect in the biomarker-negative subgroup, there may be an observed effect in the overall population as a result of the large

**Table 1.** Trial Designs With Integrated Biomarkers

Design No.	Design/Assessment Strategy	Outcome of Treatment Assessment*			Design Conclusion: Treatment Beneficial (yes or no)	
		Overall	Biomarker Positive	Biomarker Negative	Biomarker Positive	Biomarker Negative
Designs assessing biomarker-positive and -negative subgroups separately (subgroup-specific approach)						
1	Parallel assessment of biomarker-positive and -negative subgroups	NA	–	–	No	No
		NA	+	–	Yes	No
		NA	–	+	No	Yes
		NA	+	+	Yes	Yes
2	Sequential assessment: first biomarker-positive subgroup; if positive, then biomarker-negative group	NA	–	NA	No	No
		NA	+	–	Yes	No
		NA	+	+	Yes	Yes
		NA	–	–	No	No
Designs assessing overall population and biomarker-positive subgroup separately (overall/biomarker-positive approach)						
3	Parallel assessment of overall population and biomarker-positive subgroup	–	–	NA	No	No
		–	+	NA	Yes	No
		+	–	NA	Yes	Yes
		+	+	NA	Yes	Yes
4	Sequential assessment: first biomarker-positive subgroup; if positive, then overall population	NA	–	NA	No	No
		–	+	NA	Yes	No
		+	+	NA	Yes	Yes
5	Fallback design: first overall population; if negative, then biomarker-positive subgroup	–	–	NA	No	No
		–	+	NA	Yes	No
		+	NA	NA	Yes	Yes
Designs combining sequential subgroup-specific approach and overall evaluation (hybrid design)						
6	Sequential assessment: first biomarker-positive subgroup; if positive, then biomarker-negative subgroup; if negative, then overall population	–	–	NA	No	No
		NA	+	–	Yes	No
		NA	+	+	Yes	Yes
		+	–	NA	Yes	Yes

Abbreviation: NA, not applicable.

\*– denotes negative outcome of treatment assessment; + denotes positive outcome of treatment assessment.

effect in the biomarker-positive patients. This concern is particularly pronounced in the sequential version of the design, which first tests the biomarker-positive subgroup and then, if it is positive, tests the overall population; once clinical benefit has been demonstrated in the biomarker-positive population, the only therapeutically relevant question is how well the treatment works in the biomarker-negative patients, not how well it works overall. In fact, it is hard to imagine not assessing the treatment effect in the biomarker-negative subgroup after seeing a significant effect in the biomarker-positive subgroup. For example, consider the sequential overall/biomarker-positive design specified for the trial of letrozole plus lapatinib versus letrozole plus placebo in breast cancer, with the biomarker defined by human epidermal growth factor receptor 2 (HER2).<sup>13</sup> Even though this sequential procedure indicated a statistically significant effect for lapatinib in the overall population, contrary to the formal decision rule of the design, the investigators rightfully concluded that the benefit was limited to HER2-positive patients because no clinically meaningful benefit was observed in HER2-negative patients. It makes no sense to use a clinical trial design based on an analysis strategy that will not be followed.

Because the subgroup-specific approach provides the best direct evidence for clinical decision making, why is it not always used? The reason is that the subgroup-specific approach can require a much larger sample size than the overall/biomarker-positive approaches, especially when one hypothesizes that the treatment effect is relatively homogeneous across the biomarker-defined subgroups. For example, suppose the biomarker has 50% prevalence and is not predictive, with the same treatment effect in the biomarker-positive and -negative subgroups. Then, assessing the treatment effect in each subgroup separately would require a trial roughly twice as large as a test of the treatment in the overall population. When the biomarker prevalence is not 50%, some patients have unavailable biomarker status, or different effect sizes are targeted for the different subgroups, then the effect on the relative sample size required for the subgroup-specific approach as compared with the overall/biomarker-positive approaches is more complicated. For example, the SATURN (Sequential Tarceva in Unresectable NSCLC) trial<sup>9,16</sup> randomly assigned 889 patients with NSCLC to erlotinib versus placebo; 70% were EGFR positive, 14% were EGFR negative, and 16% had indeterminate or missing EGFR status. The trial used an  $\alpha = 0.025$  parallel-assessment overall/biomarker-positive design that allocated  $\alpha = 0.01$  for the EGFR-positive subgroup and  $\alpha = 0.015$  for the overall population. If one used a subgroup-specific approach (sequential assessment with  $\alpha = 0.025$ ), one would need to enroll approximately 650 EGFR-negative patients, which would require screening approximately 4,600 patients and an unnecessarily large 3,200-patient EGFR-positive subgroup. There was the possibility of using a two-stage subgroup-specific design that would have suspended accrual after the required number of biomarker-positive patients was enrolled; if the biomarker-positive subgroup analysis had been positive with further follow-up, the accrual of biomarker-negative patients would have resumed. However, even this approach would have resulted in a much larger screened population compared with that in the SATURN trial.

The subgroup-specific approach will not always require larger sample sizes than the overall/biomarker-positive design, especially when the proportion of the biomarker-positive patients is small. For example, the lapatinib trial<sup>13</sup> had a sample size of 1,286; 17% were HER2 positive, 74% were HER2 negative, and 9% had missing HER2 status. The sequential-assessment overall/biomarker-positive design ( $\alpha = 0.025$ ) targeted a hazard ratio (HR) of 0.645 in the HER2-positive population (80% power) and an HR of 0.769 in the overall population (90% power).<sup>13</sup> A sequential subgroup-specific design ( $\alpha = 0.025$ ) of the same sample size would have had 80% power to detect an HR of 0.645 in the HER2-positive subgroup and 97% power to detect an HR of 0.769 in the HER2-negative subgroup.

If the required sample size to perform a subgroup-specific approach is feasible, then this design is recommended. If not, a hybrid design that sequentially tests the treatment effect in the subgroups and overall may be useful (design 6). First, the biomarker-positive subgroup is tested at a reduced level  $\alpha_1$ . If it is significant, then the biomarker-negative subgroup is tested at the level  $\alpha$ . If the biomarker-positive test is not significant, then the overall population is tested at the level of  $\alpha_2 = \alpha - \alpha_1$ . This design controls the overall type-I error at level  $\alpha$  when the treatment does not work in either subgroup. In addition, in contrast to the overall/biomarker-positive approach, this design only tests the overall population when no significant effect is detected in the biomarker-positive subgroup; this allows controlling the probability of erroneously concluding overall benefit when the overall effect is driven by the biomarker-positive patients. In particular, the probability of erroneously recommending the treatment for biomarker-negative patients depends on the choice of  $\alpha_1$  and the proportion of the biomarker-positive patients;  $\alpha_1$  should be chosen to ensure that the probability of such recommendations is sufficiently low. For example, in a phase III trial of blinatumomab in acute lymphoblastic leukemia (E1910), the presence of a specified quantity of minimum residual disease (MRD) was thought to be predictive of whether the treatment would work. The trial ( $N = 285$ ) was designed to first test the MRD-positive subgroup (expected to be approximately one third of the population) using  $\alpha_1 = 0.02$ . If the treatment effect were statistically significant for this MRD-positive subgroup, then the MRD-negative subgroup would be tested using  $\alpha = 0.025$ . Otherwise, the overall population would be tested using  $\alpha_2 = 0.005$ . It could be shown that if the treatment benefit were limited to the MRD-positive patients, then the probability that this design would recommend blinatumomab for the MRD-negative patients would be  $< 0.026$  (this would be true regardless of the magnitude of the treatment effect in the MRD-positive patients, assuming that one third of the study population was MRD positive). On the other hand, with the overall/biomarker-positive designs described in the previous section, the probability of erroneously recommending the agent for MRD-negative patients could be as high as 100% (if the agent worked well in the MRD-positive patients).

## DISCUSSION

Integrated trial designs that test only the biomarker-positive and the overall populations may formally recommend a treatment for the biomarker-negative subgroup, even though it does not work

for these patients. When the required sample size is feasible, a trial design that evaluates both the biomarker-positive and -negative subgroups is preferred. Sometimes, a hybrid design that assesses the treatment effect in the biomarker-specific subgroups and overall can lower required sample sizes while minimizing the probability of recommending an ineffective treatment for the biomarker-negative subgroup. For all designs, the final study report should include estimates of the treatment effect in the biomarker-positive and -negative subgroups.

## REFERENCES

1. Simon R, Maitournam A: Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Res* 10:6759-6763, 2004
2. Freidlin B, McShane LM, Polley MY, et al: Randomized phase II trial designs with biomarkers. *J Clin Oncol* 30:3304-3309, 2012
3. Mandrekar SJ, Sargent DJ: Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *J Clin Oncol* 27:4027-4034, 2009
4. Freidlin B, McShane LM, Korn EL: Randomized clinical trials with biomarkers: Design issues. *J Natl Cancer Inst* 102:152-160, 2010
5. Simon R: Clinical trials for predictive medicine. *Stat Med* 31:3031-3040, 2012
6. US Food and Drug Administration: FDA draft guidance 2012: Enrichment strategies for clinical trials to support approval of human drugs and biological products. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf>
7. Douillard JY, Siena S, Cassidy J, et al: Randomized, phase III trial of panitumumab with infu-

sional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: The PRIME study. *J Clin Oncol* 28:4697-4705, 2010

8. Peeters M, Price TJ, Cervantes A, et al: Randomized phase III study of panitumumab with fluorouracil, leucovorin, and irinotecan (FOLFIRI) compared with FOLFIRI alone as second-line treatment in patients with metastatic colorectal cancer. *J Clin Oncol* 28:4706-4713, 2010

9. Cappuzzo F, Ciuleanu T, Stelmakh L, et al: Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: A multicentre, randomised, placebo-controlled phase 3 study. *Lancet Oncol* 11:521-529, 2010

10. Redman MW, Crowley JJ, Herbst RS, et al: Design of a phase III clinical trial with prospective biomarker validation: SWOG S0819. *Clin Cancer Res* 18:4004-4012, 2012

11. Scagliotti GV, Vynnychenko I, Park K, et al: International, randomized, placebo-controlled, double-blind phase III study of motesanib plus carboplatin/paclitaxel in patients with advanced nonsquamous non-small-cell lung cancer: MONET1. *J Clin Oncol* 30:2829-2836, 2012

## AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The author(s) indicated no potential conflicts of interest.

## AUTHOR CONTRIBUTIONS

**Manuscript writing:** All authors

**Final approval of manuscript:** All authors

12. Boyer MJ, Janne PA, Mok T, et al: ARCHER: Dacomitinib (D; PF-00299804) versus erlotinib (E) for advanced (adv) non-small cell lung cancer (NSCLC)—A randomized double-blind phase III study. *J Clin Oncol* 30:508s, 2012 (suppl; abstr TPS7615)

13. Johnston S, Pippin J Jr, Pivrot X, et al: Lapatinib combined with letrozole versus letrozole and placebo as first-line therapy for postmenopausal hormone receptor-positive metastatic breast cancer. *J Clin Oncol* 27:5538-5546, 2009

14. Freidlin B, Simon R: Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 11:7872-7878, 2005

15. Rothmann MD, Zhang JJ, Lu L, et al: Testing in a prespecified subgroup and the intent-to-treat population. *Drug Inf J* 46:175-179, 2012

16. Bruggen W, Triller N, Blasinska-Morawiec M, et al: Prospective molecular marker analyses of *EGFR* and *KRAS* from a randomized, placebo-controlled study of erlotinib maintenance therapy in advanced non-small-cell lung cancer. *J Clin Oncol* 29:4113-4120, 2011