Run-In Phase III Trial Design With Pharmacodynamics Predictive Biomarkers

Fangxin Hong, Richard Simon

Manuscript received January 2, 2013; revised July 31, 2013; accepted August 1, 2013.

Correspondence to: Richard Simon, PhD, Biometric Research Branch, National Cancer Institute, Bethesda, MD, 20892 (e-mail: rsimon@mail.nih.gov).

Background Developments in biotechnology have stimulated the use of predictive biomarkers to identify patients who are likely to benefit from a targeted therapy. Several randomized phase III designs have been introduced for development of a targeted therapy using a diagnostic test. Most such designs require biomarkers measured before treatment. In many cases, it has been very difficult to identify such biomarkers. Promising candidate biomarkers can sometimes be effectively measured after a short run-in period on the new treatment.

Methods We introduce a new design for phase III trials with a candidate predictive pharmacodynamic biomarker measured after a short run-in period. Depending on the therapy and the biomarker performance, the trial would either randomize all patients but perform a separate analysis on the biomarker-positive patients or only randomize marker-positive patients after the run-in period. We evaluate the proposed design compared with the conventional phase III design and discuss how to design a run-in trial based on phase II studies.

Results The proposed design achieves a major sample size reduction compared with the conventional randomized phase III design in many cases when the biomarker has good sensitivity (≥0.7) and specificity (≥0.7). This requires that the biomarker be measured accurately and be indicative of drug activity. However, the proposed design loses some of its advantage when the proportion of potential responders is large (>50%) or the effect on survival from run-in period is substantial.

Conclusions Incorporating a pharmacodynamic biomarker requires careful consideration but can expand the capacity of clinical trials to personalize treatment decisions and enhance therapeutics development.


Improved understanding of cancer biology has stimulated the development of molecularly targeted cancer treatments that will likely only benefit patients whose tumors are driven by deregulation of the drug targets. The standard phase III trial testing average drug effect across patients with broad eligibility criteria is often no longer efficient. Even when such trials result in statistical significance, a large proportion of the patients do not benefit from the new treatment.

A key component in developing targeted therapy is the identification of predictive biomarkers that can identify patients who are likely to benefit. Effective predictive biomarkers can benefit patients, control costs by personalizing treatment, and enhance the efficiency of clinical development. Statisticians are challenged to develop new designs and analysis strategies to incorporate predictive biomarkers. Several randomized phase III designs have been previously introduced for this purpose (1,2), including the marker strategy design, the enrichment design (3), and the marker-stratified design (4). All of these designs require pretreatment biomarker measurement. In many cases, it has been very difficult to identify such pretreatment biomarkers. Biomarkers measured after receiving the randomized treatment are generally not suitable because different treatment arms could have differential effects on biomarker values. Some studies use posttreatment biomarkers as surrogates of clinical outcome, but establishing an intermediate endpoint as a valid surrogate is quite difficult (5).

Run-in periods in which all patients receive the test drug for a short period of time have been used in some clinical trials to exclude or select patients for subsequent randomization (6). The earliest run-in designs were implemented to exclude patients with poor compliance to treatment (7,8). Run-in periods in which all patients receive placebo have been used to exclude placebo responders (9).

In this article, we explore the use of pharmacodynamic biomarkers measured after a short run-in period on the new treatment as a predictive biomarker. A wide variety of such biomarkers are potentially available. Immunologic response to a therapeutic cancer vaccine is one example. Dendritic cell–based cancer vaccines, although expensive, are very effective for inducing antitumor immunity in a variety of cancers (10,11). However, clinical responses are observed in only a subset of patients (12). Assessing early immunologic response may efficiently identify the subset of patients who will have a greater chance of eventually having clinical responses. A second area is the use of mechanistic markers.
of HIF-1 has been shown to be required for a positive response to EGFR-targeted therapies in triple-negative breast cancer (13). Unless there is a substantial downregulation of the target after a run-in period, it may be less likely that the tumor will be controlled by the drug. A third area is in resistance markers. Nuclear expression of EGFR, if detected after a short run-in period on cetuximab, might indicate potential resistance to cetuximab (14). If nuclear expression of EGFR could be measured accurately, it might serve as a predictive biomarker for clinical trials that examine the value of adding Src family kinases inhibitors to cetuximab for patients with potential cetuximab-resistant tumors. In this example, the run-in period would be on the control regimen, cetuximab alone. Finally, one might use an imaging measurement made after a short run-in. Imaging of apoptosis can facilitate early assessment of anticancer treatment before tumor shrinkage (15). The above examples are areas where pharmacodynamic biomarkers can potentially be useful. Incorporating posttreatment pharmacodynamic biomarkers in randomized designs has the potential to greatly expand the capacity of clinical trials to personalize treatment decisions.

In this article, we introduce a new class of clinical trial designs that enable pharmacodynamic biomarkers to be used as candidate predictive biomarkers for identifying the patients who are likely to benefit. Pharmacodynamic biomarkers are broadly defined as any posttreatment measurements indicating pharmacological effect of a drug. We do not propose to use the biomarkers as surrogates for clinical outcome. The pharmacodynamic biomarker serves only to permit separate evaluation of treatment effect in the subset of marker-positive patients. We evaluate the efficiency of these designs as a function of the sensitivity and specificity of the biomarker and the effect of the run-in period. The sensitivity and specificity incorporate both the accuracy of the assay used to measure the biomarker and the informativeness of the biomarker. We provide recommendations for application of the proposed design.

**Methods**

Pharmacodynamic biomarkers can be generally categorized into two groups: efficacy or resistance biomarkers. Efficacy biomarkers measure response to new treatment, and resistance biomarkers measure resistance to standard treatment. For simplicity, we used efficacy biomarkers to illustrate and evaluate the design, but the results can be easily generalized to designs with resistance biomarkers. We evaluated the proposed run-in design in the context of time-to-event endpoints (survival) and assumed a single prospective defined measurement can be effectively performed after a short run-in period and used as a candidate predictive biomarker.

**Design Framework**

The schema for the proposed design is illustrated in Figure 1 with three motivating examples. We assumed patients meeting broad eligibility criteria consist of a group of true responders (R+) who will benefit from the targeted therapy and a group of true nonresponders (R−) who will not. However, the true response class is, in general, nonobservable. We used this simple dichotomization for purposes of better understanding and quantifying the conditions when the run-in design is or is not beneficial but recognize that graded levels of sensitivity to therapy often exist. All patients receive the new treatment for a defined short run-in period. Biomarker status will be assessed afterward, and patients will be classified as either biomarker positive (M+) or negative (M−). The pharmacodynamic biomarker provides an (imperfect) estimate of the underlying responder status.

Depending on the strength of prior evidence, the design would either randomize only the M+ patients or all patients, stratified on biomarker status, at 1:1 ratio to receive the experimental treatment (new drug or new drug plus standard therapy) or control (standard therapy). The objective is to test whether the experimental treatment prolongs survival compared with the control. When all patients are randomized, two statistical tests will be performed, and the thresholds of significance will be adjusted to preserve the overall study-wise type I error. The first test (test all) includes all patients at a two-sided significance level of $\alpha_{1629}$. If the first test is not statistically significant, the second test (test positive) is performed for the M+ subset at a two-sided significance threshold of $\alpha_5$. If $\alpha_{1629} + \alpha_5 = 0.05$, then the study-wise type I error, which is the probability that either of the tests will be found statistically significant when the treatment is uniformly ineffective, will be no greater than 0.05. For our simulations, we used $\alpha_{1629} = 0.01$ and $\alpha_5 = 0.04$ but discuss later how to split the type I error. The probability of finding statistical significance for either test when the experimental treatment is beneficial is the power of the trial, which is referred to as trial-level power. When only M+ patients are randomized, one test will be performed for this subgroup at a two-sided significance level of 0.05.

**Design Parameters and Evaluation**

For a given biomarker, the sensitivity is the probability that a true responder is marker positive, and the specificity is the probability that a nonresponder is marker negative. Overall survival is measured as the time from randomization to death. We assumed exponential distributions for overall survival, a proportional hazard between experimental and control for true-responders with a hazard ratio $\theta = 0.6$, and a proportional hazard between experimental and control for nonresponders with a hazard ratio $\theta_{1629} = 1$. Patients were randomized upfront in the standard design and randomized after a fixed run-in period in the run-in design. We also evaluated the effect of the run-in period on the efficiency of the run-in design. For R+ patients who were randomized to the control arm, the potential survival benefit from the run-in period was characterized as a hazard ratio of $\theta_{1629}$, which is usually close to 1 because the duration of the run-in period is assumed short. We assumed no run-in benefit for control group patients who were R−. The overall survival distributions (Supplementary Materials, available online) do not follow a proportional hazard in this formulation.

The median survival for the entire population with control treatment was arbitrarily selected as 8 months. The efficiency results are not dependent on this assumption, however. We assumed no run-in effect with $\theta_{1629} = 1$, but also considered a run-in effect with $\theta_{1629} = 0.95$ (Supplementary Figures 1 and 2, available online). We considered three possible R+ prevalence values (25%, 50%, and 75%) and four possible levels of biomarker sensitivity and specificity (0.7, 0.8, 0.9, and 1.0). The ranges encompass those used with most potentially useful biomarkers (16).

We used two approaches for evaluation. First, when all patients are randomized, we compared the power of the run-in design and the standard design with a common sample size. For each R+ prevalence,
Figure 1. The schema of run-in design with three motivating examples: 1) immunological biomarker in vaccine therapy, 2) imaging biomarkers for early response, and 3) mechanistic markers for drug resistance. PD = pharmacodynamics.

Figure 2. With sample sizes that give 80% power for the standard design, the trial-level power with the run-in design (solid lines) is shown when randomizing all patients, for a series of sensitivity and specificity of the biomarker, under 25%, 50%, and 75% prevalence of true responders, with no run-in effect ($\theta_0 = 1$).
the sample size was selected to give standard design 80% power with accrual rate of 20 patients/month and an additional 12 months follow-up. The trial-level power for the run-in design was calculated through simulations. When calculating the power (sample size) of the standard design, we used a numerical approximation implemented in a program, called powlgrnk. The method computes the power of a two-sample log-rank test with arbitrary survival distribution without a proportional hazard assumption. This numerical function is equivalent to the simulation method with a very large number of simulation replications and is used, where applicable, for efficiency (17).

We also compared the designs with regard to the number of events required to have 80% trial power. For simplicity of comparison, we assumed all patients had events in the trial with sufficient follow-up. The average simulated number of events was estimated through simulations for the run-in design when all patients are randomized. The same numerical approximation method described above was used to estimate number of events for the run-in design when only the M+ subset is randomized and for the standard design. The sample size for each scenario can be back-calculated from number of events under the specified assumptions.

### Results

#### Simulation Results

The standard design requires 1660, 560, and 300 case subjects to have 80% power under 25%, 50%, and 75% R+. The weighted average hazard ratio is approximately 0.9, 0.8, and 0.7, respectively. Figure 2 shows the trial-level power for the run-in design when randomizing all patients. The power increases with increasing values of sensitivity, specificity, and proportion R+. The run-in design generally has greater power than the standard design, especially when the R+ proportion is lower. In that case, specificity is more important than sensitivity because having nonresponders in the M+ group (low specificity) reduces power in the M+ subset that contributes most of the trial-level power. Nevertheless, the run-in design almost always has improved power for biomarkers with good sensitivities (0.8 or higher) regardless of R+ proportion. However, the run-in design loses some of its efficiency advantage when the prevalence of R+ is 75%.

Similar results were observed when evaluating the designs using the second approach (Figure 3). When only M+ patients are randomized (Figure 3, bottom panels), the run-in design requires substantially fewer events than the standard design, regardless of sensitivity, specificity, or R+ prevalence, although the advantage is reduced with high (75%) R+ proportion. For example (Supplementary Table 1, available online), a standard phase III design would require 1440 events when the R+ prevalence is 25%. A run-in design using a good biomarker with sensitivity 0.8 and specificity 0.9 would require about 835 events when all patients are randomized and 220 events when only M+ patients are randomized.

#### Designing a Run-In Trial

In evaluating whether to use a run-in design and to structure and size the trial, one will need data from previous studies on marker prevalence and analytical performance. Although we evaluated the run-in design as a function of the sensitivity and specificity of the biomarker, the true responder status of patients is not identifiable in many cases. Thus no direct estimate of sensitivity and specificity will be available. Instead, the treatment effect to be detected in M+ patients and overall can be specified for current clinical trials; the prevalence of M+ can be assessed from phase II trials and used to guide the trial planning. A large difference in treatment effect between M+ and M− groups in phase II trials provides promise that the pharmacodynamic marker warrants further investigation. A run-in trial is required, however, to distinguish whether the marker is predictive or merely prognostic. The strength of the biological basis that treatment benefit will be limited to M+ patients and preliminary evidence of treatment effect for M− patients would be used to decide whether to randomize all patients or only the M+ subset. In either case, development of a precise and reproducible assay for measuring the biomarker is a prerequisite, which is true also for pretreatment measured biomarkers. The sizing of the run-in phase III trial will depend on the prevalence of marker positivity, the size of the treatment effect to be detected in the M+ group and overall, and the split of the alpha between the two significance tests. When all patients are randomized, alpha is split, and proportional hazards are assumed within marker groups, a simulation might be used to size the study for a trial-level power of 80% or 90%.

### Discussion

We have introduced a design for phase III clinical trials with predictive biomarkers measured after a short run-in period. When fewer than half the patients are expected to benefit from the new treatment, the run-in design is generally much more efficient than the standard design regardless of whether all patients are randomized or only M+ patients. Run-in designs do not have clear advantages when proportion of patients expected to benefit is high (e.g., 75%), or when sensitivity/specificity are low (<0.7). We have provided the required number of events to achieve 80% power under a range of sensitivity/specificity and hazard ratios for R+ populations (Supplementary Table 2, available online). We presented the design in terms of an unobservable latent true-response variable and the sensitivity and specificity of the biomarker. However, one does not know those quantities to design a run-in trial. To evaluate drug activity in both M+ and M− groups, both groups need to be included in the study. In some cases with strong biological rational, it may be appropriate to do trials only in the M+ subset.

We used a simple method to split the alpha between test all and test positive, although other methods could be used. The split may depend on the a priori evidence that the treatment effect will be limited to M+ patients and the prevalence of M+ patients (18,19). The 0.01/0.04 (overall/subset) split is useful when the proportion of M+ patients is small and the marker has sufficiently strong credentials. Then the sample size will be driven by powering the treatment evaluation in the M+ subset, and using a small alpha for the subset will increase the sample size unnecessarily. When the biomarker has weaker credentials, one may wish to use a split of 0.04/0.01 (overall/subset). Only a modest increase in the overall sample size is needed to preserve the power of the overall analysis relative to a standard clinical trial with 0.05 significance level.

The potential advantages of the run-in design can be easily generalized to biomarkers measured on a continuous or graded scale. In such cases, the run-in design would randomize all patients.
Tests of treatment effect in all patients and in the subset with biomarker greater than an optimized threshold can be performed using strategies previously introduced for the adaptive threshold design (19). Such a design enables the threshold to be optimized using phase III data while fully controlling the study-wise type I error, providing rigorous statistical testing of treatment effects overall and in a subset using an optimized cut point for marker positivity.

In some cases, there might be a minimal survival benefit from the run-in period on R+ patients randomized to control. The hazard ratio between control and experimental arms would be reduced, which leads to efficiency loss for the run-in design (Supplementary Figures 1 and 2, available online), particularly with high R+ prevalence. Therefore, we do not recommend a run-in design when a large run-in effect and large R+ proportion (>50%) might be expected. One might also be concerned about a detrimental effect of the run-in for R− patients because it might delay their starting the control regimen. When there is highly effective standard treatment, however, the test regimen is likely to involve the standard treatment plus new drug. When the control regimen is not highly effective, the delay is of less concern and the run-in might use the new drug alone if that provided the most informative biomarker assessment.

Although run-in designs have rarely been used for phase III oncology trials, the randomized discontinuation design is a phase II design that has been used (16). In randomized discontinuation design, all patients receive the new drug, and patients with early objective response or early progressive disease are excluded from randomization. The design can be more or less efficient than the standard design depending on the mechanism of action of the drug (20). The run-in design uses an unvalidated marker measurement and either randomizes all patients or only marker-positive patients; thus, it can potentially provide a much greater improvement in power than randomized discontinuation design. The proposed run-in designs use an intermediate (posttreatment) measurement as a predictive biomarker. However, the marker will not be used as a surrogate endpoint nor necessarily be a response marker; the biomarker is only being used to focus the treatment comparison on the subset of M+ patients. Although early response is one type of posttreatment predictive biomarker, it is not used as a surrogate for making claims about treatment effectiveness. For phase III run-in trials used for drug registration, if the trial demonstrates benefit of the new regimen only in M+ patients or if the M− patients are not randomized, the intended use of the drug would be for patients who were determined to be M+ based on the prespecified cut point after the defined run-in period. And the label should indicate that the estimated improvement in survival time is measured from the end of the short run-in period.

Our proposed design is not without limitations. It adds substantial complexity and cost to the clinical trial and requires substantial credentials for the marker. It depends on the development of an analytically validated assay for accurate measurement of the candidate marker before the initiation of the phase III clinical trial. It loses some efficacy when a large run-in effect and large R+ proportion (>50%) are expected. It would involve novel drug labeling when used for drug registration.

Advances in molecular and imaging technology provide substantial opportunities for development of biomarkers that measure early treatment effect. Early immunologic response,
mechanistic measures of treatment effect, early imaging assessments of patient response, and treatment of autoimmune diseases on the basis of early response from mediators of inflammation are but a few examples. Efficient design strategies, prospectively specified analysis plans, and analytically validated assays are, however, essential to successfully establishing the clinical utility of pharmacodynamic biomarkers. Further research on the design of clinical trials with pharmacodynamic predictive biomarkers is warranted.

References


Funding

This work was supported by the Department of Biostatistics and Computational Biology Research Scientist Fund (to FH).

Note

All authors of this research paper have directly participated in the planning, execution, or analysis of the study. All authors of this paper have read and approved the final version submitted. The study sponsor did not have a role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

We thank the referees for helpful suggestions.

Affiliations of authors: Departments of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, MA (FH); Biometric Research Branch, National Cancer Institute, Bethesda, MD (RS).