Design of a Phase III Clinical Trial with Prospective Biomarker Validation: SWOG S0819

Mary W. Redman1, John J. Crowley2, Roy S. Herbst3, Fred R. Hirsch4,5, and David R. Gandara6

Abstract
The role of cetuximab in the treatment of advanced non–small cell lung cancer (NSCLC) is currently unclear. The molecular target of cetuximab, epidermal growth factor receptor (EGFR), as measured by FISH, has shown potential as a predictive biomarker for cetuximab efficacy in NSCLC. SWOG S0819 is a phase III trial evaluating both the value of cetuximab in this setting and EGFR FISH as a predictive biomarker. This work describes the decision process for determining the design and interim monitoring plan for S0819. Six possible designs were evaluated in terms of their properties and the hypotheses that can be addressed within the design constraints. A subgroup-focused, multiple-hypothesis design was selected for S0819 that incorporates coprimary endpoints to assess cetuximab in both the overall study population and among EGFR FISH-positive (FISH+ ) patients, with the sample size determined based on evaluation in the EGFR FISH+ group. The chosen interim monitoring plan specifies interim evaluations of both efficacy and futility in the EGFR FISH+ group alone. The futility-monitoring plan to determine early stopping in the EGFR FISH-nonpositive group is based on evaluation within the positive group, the entire study population, and the nonpositive group. SWOG S0819 uses a design that addresses both the biomarker-driven and general-efficacy objectives of this study. Clin Cancer Res; 18(15); 1–9. ©2012 AACR.

Introduction
It is well established that epidermal growth factor receptor (EGFR) is an important molecular target in the treatment of non–small cell lung cancer (NSCLC). Cetuximab, an EGFR-directed monoclonal antibody, blocks ligand-induced EGFR activation, stimulates receptor internalization, and is capable of inducing antibody-dependent cellular cytotoxicity. SWOG S0819, a currently active phase III trial, is evaluating whether the addition of cetuximab to chemotherapy improves efficacy in advanced NSCLC. The rationale for the trial was provided by the SWOG predecessor phase II trials S0342 and S0536. Trial S0342 evaluated the addition of cetuximab to the chemotherapy doublet carboplatin/paclitaxel, and S0536 added bevacizumab to the regimen (1, 2). Both studies met prespecified benchmarks, indicating that further study of cetuximab was warranted in the phase III setting and resulting in the development of S0819.

The development of trial S0819 was further motivated by 2 randomized phase III studies [BMS 099 and First Line Erbitux in Lung Cancer (FLEX)] that tested platinum-based doublet chemotherapy with and without cetuximab for first-line therapy of advanced NSCLC (3, 4). BMS 099 was conducted in an unselected population of patients with advanced-stage NSCLC who were not prospectively categorized for EGFR pathway activation. The primary outcome, progression-free survival (PFS) based on blinded, independent radiologic review, was not significantly different between the 2 treatment arms (HR = 0.90; P = 0.236), but PFS based on institutional review was significantly different in favor of the cetuximab arm (HR = 0.766; P = 0.0015). The median survival time (MST) was numerically higher in the cetuximab arm, but this difference was not statistically significant (HR = 0.89; P = 0.17). By comparison, the primary outcome in FLEX was overall survival with eligibility based on patients with tumors exhibiting at least 1 EGFR-positive (EGFR+) tumor cell measured by immunohistochemistry (IHC): 85% of screened patients met this criterion. In FLEX, cetuximab significantly prolonged survival (HR = 0.87; P = 0.044), with an MST of 11.3 months in the cetuximab-containing arm compared with 10.1 months in the control arm. However, because the benefit of cetuximab was viewed as modest, this trial has not yet resulted in regulatory approval for this combination. Further, these results bring into question whether selection based on minimal IHC criteria is an adequate predictive biomarker for cetuximab-based chemotherapy in this setting.
Alternatively, our group reported that EGFR measured by FISH shows promise as a predictive biomarker for cetuximab-based chemotherapy (5). Specifically, in a retrospective analysis of tumor tissue in S0342, the predictive role of increased gene copy number of EGFR measured by FISH was evaluated. Tumors with ≥4 copies of the EGFR gene in ≥40% of the cells (high polysomy) or tumors with EGFR gene amplification (gene-to-chromosome ratio ≥2 or presence of gene cluster or ≥15 gene copies in ≥10% of the cells) were considered to be EGFR FISH++, whereas all other tumors were considered to be EGFR FISH-negative (FISH-). The percentage of patients defined to be EGFR FISH++ in this study was 59%. This analysis showed a doubling of median PFS (6 months vs. 3 months) and MST (15 months vs. 7 months) among EGFR FISH++ patients relative to EGFR FISH- patients. Previous studies found that EGFR FISH has no prognostic significance (6, 7).

Given the promising but inconclusive data from these phase II and phase III studies, SWOG sought to definitively assess the role of cetuximab in the therapy of advanced NSCLC and to evaluate the role of EGFR measured by FISH as a predictive biomarker. Here, we describe the process by which various study designs were compared in order to determine the "best fit" for addressing the coprimary end-points of efficacy in an overall study population, and to determine the predictive value of EGFR FISH.

Trial Designs to Address Biomarkers

Broadly speaking, clinical trial designs to evaluate the role of a treatment with a potentially predictive biomarker can be divided into 4 types: (i) an all-comers design, which includes secondary biomarker objectives; (ii) a targeted design, which restricts the study population to marker-positive patients; (iii) a strategy design, which randomizes patients to receive marker-based or nonmarker-based treatment; and (iv) a multiple-hypothesis design, which is a composite of the targeted and all-comers designs, and addresses multiple hypotheses as coprimary objectives (8, 9).

Each of these designs may have potential advantages and disadvantages depending on the study questions to be addressed. The targeted, strategy, and multiple-hypothesis designs prospectively address a potentially predictive biomarker, whereas the all-comers design requires independent validation of a biomarker in a separate trial. The targeted and multiple-hypothesis designs prospectively evaluate the efficacy of the experimental treatment within the marker-positive subset. In the strategy design discussed here, only marker-positive patients who have been randomized to receive marker-based therapy receive the experimental treatment, and all other patients are assigned to receive the standard of care. The primary objective of this design is to evaluate whether assignment to treatment based on a biomarker improves outcomes. Moreover, because all marker-negative patients are assigned to standard therapy, this design cannot assess the treatment effect in an unselected population or a marker-negative population. Because the targeted design restricts its patient population to marker-positive patients, it is also unable to address the role of the experimental treatment in unselected and marker-negative populations.

Multiple-hypothesis designs usually involve the specification of a hypothesis in the target population defined to be marker positive, as well as in either the entire study population or the marker-negative population. These hypotheses are treated as coprimary hypotheses, and to control the overall study false-positive rate, a fraction of the type I error is apportioned to each hypothesis.

Multiple-hypothesis designs can be subgroup or overall-population focused, or they can specify discrete hypotheses for multiple subgroups. In the case of a subgroup-focused design, the study design is based on the subgroup, and the properties of the design in the entire study population are based on the residual type I error and the number of patients who must be screened to achieve the sample size for the subgroup. An overall-population–focused design is just the reverse; however, in such a design, the marker (or set of markers) does not have to be prespecified as in the adaptive-signature design of Freidlin and Simon (10). Discrete-hypothesis designs are essentially parallel targeted designs that may include an additional hypothesis (such as an interaction hypothesis test) that joins the 2 designs together.

Examples of study designs for advanced NSCLC that were overall-population focused include the Sequential Tarceva in Unresectable NSCLC (SATURN) and INTEREST trials (11, 12). An example of a discrete-hypothesis design is the MARVEL trial, which specified separate hypotheses for marker-positive and -negative subgroups as defined on the basis of EGFR FISH (13). A common allocation of the type I error to the primary of the coprimary hypotheses is 80%, but other splits may be reasonable and will depend on the marker prevalence and specified hypotheses (10). To determine the testing level for the 2 hypotheses, one could just use a Bonferroni correction. However, for the subgroup-focused and overall-focused designs, the subgroup hypothesis is nested within the overall-population hypothesis, and therefore a Bonferroni split would be conservative. One can determine the exact split analytically, through simulations, or by using software for group sequential designs (14).

The choice of a trial design involving a biomarker depends greatly upon what is known about the marker specific to the treatment and disease setting. Hoering and colleagues (8) evaluated the relative merits of the all-comers, targeted, strategy, and overall-population–focused multiple-hypothesis trial designs. Their general recommendations are to use the targeted design only when there is good evidence that the marker-negative population will not benefit from the experimental treatment, and to use the strategy design only if the strategy hypothesis is truly of more interest and value than the efficacy hypothesis. Otherwise, they recommend using either the all-comers design with a marker-driven secondary hypothesis or a multiple-hypothesis design.
The choice of which multiple-hypothesis design to use should also depend on the specific scenario involved. The implication of a subgroup-focused design is that evidence indicates that the biomarker is predictive and the entire population may benefit. The implication of an overall-population-focused design is that the evidence around the biomarker is less strong, but a biomarker-based subgroup analysis is included as a backup plan should the overall evaluation fail. In a discrete-hypothesis design, separate hypotheses are specified for biomarker-defined groups. A natural consequence of this is that the separate hypotheses should be different. Should the general hypothesis be the same in both the marker-positive and -negative groups (at least in direction), the implication is that there is a hypothesized effect in the unselected population and either the subgroup-focused or overall-population-focused multiple-hypothesis design would address this setting more directly, and likely more efficiently.

In the context of the clinical trial setting for S0819, although there is evidence that the effect of cetuximab may be larger in the EGFR FISH+ population, it is not clear that there is no effect in the negative population. Moreover, the primary question is centered on the efficacy of cetuximab, rather than the effect of assigning treatment based on a biomarker. Therefore, the most applicable designs are the all-comers design and the subgroup-focused or overall-population-focused multiple-hypothesis designs.

Design Assumptions

For all of the designs evaluated in this work, the power for the primary hypothesis was specified to be 90% and the study-wide, 1-sided type I error rate was 2.5%. The antiangiogenic agent bevacizumab can be added to the treatment regimen in either arm, based on previously established clinical criteria, and randomization is stratified by bevacizumab-eligible status. To determine the null hypothesis for median PFS, it was assumed that ~50% of patients would be bevacizumab eligible, the median PFS with carboplatin/paclitaxel alone would be 4 months, and the median PFS with carboplatin/paclitaxel plus bevacizumab would be 6 months (15, 16). Based on previous study within SWOG, it was assumed that the accrual rate would be ~366 patients/year. Further, on the basis of a study by Hirsch and colleagues (5), it was assumed that 40% of all screened patients would be EGFR FISH+, 40% would be EGFR FISH−, and 20% would have unknown EGFR FISH status due to insufficient tissue or a failed assay. The EGFR FISH-nonpositive group is defined as the set of patients whose EGFR FISH status is unknown or who have been defined as EGFR FISH−.

On the basis of previous studies, it was determined that the target level of improvement in the EGFR FISH+ population would be a 33% improvement in median PFS (equivalent to HR = 0.75), and the target level of improvement in the unselected population was a 20% improvement (equivalent to HR = 0.83). PFS was chosen as the primary objective in the EGFR FISH+ group because of the potential benefit with an EGFR tyrosine kinase inhibitor in the second (or later) line setting. The actual trial was designed with overall survival as the primary outcome for the unselected population (for reasons not described here), which results in slightly less power for the overall-population assessment. However, to facilitate a fair comparison, we evaluate the candidate designs here using the same primary outcome (PFS).

If the data can be assumed to follow the exponential distribution, then the overall-population HR can be approximated by a weighted average of the HRs in the marker-positive and -negative groups. Moreover, if the marker is not prognostic, then the alternative HR is approximately a weighted average of the HRs for the marker-positive and -negative groups. Under these assumptions, a 20% improvement in median PFS overall and a 33% improvement in the EGFR FISH+ group translates into an 11% improvement in the EGFR FISH-negative group and a 7% improvement in the EGFR FISH− group. To address the strategy-design hypothesis, a 33% improvement in median PFS among EGFR FISH+ patients translates to a 13% improvement associated with assigning treatment based on marker status.

Allocating 80% of the type I error to one of the hypotheses in a multiple-hypothesis design results in a design with a 1-sided 0.02 type I error for the primary hypothesis. Using simulations, the exact residual type I error for this study was determined to be 0.008 to achieve a study-wide type I error rate of 2.5%. If 60% of the type I error is apportioned to one of the hypotheses in a multiple-hypothesis design, the levels are 0.015 for the primary group and ~0.013 for the other group.

Candidate Phase III Designs

Table 1 details the sample sizes and expected time to completion for the 6 designs considered. Throughout this work, the all-comers design is used as the reference design in terms of both sample size and time to completion. The sample size for the all-comers design is 1,352 patients with an expected 541 EGFR FISH+ patients. Evaluating the effect of cetuximab in the EGFR FISH+ population at the 2.5% level as a secondary objective has 89.6% power to detect a 33% improvement in median PFS.

Relative to the all-comers design, the targeted design would require 15 (3%) more EGFR FISH+ patients and therefore necessitate screening 38 (3%) more patients overall, taking just >1 month longer to complete the study. The strategy design would require more than double the number of patients and would take 84% longer to complete. The subgroup-focused multiple-hypothesis designs would require 9% more patients (47 EGFR FISH+ and 110 overall) and 16% more patients (85 EGFR FISH− and 206 overall), and would take 5% and 11% longer to complete for the 80% and 60% splits, respectively. The overall-population-focused multiple-hypothesis design would require 5% more patients (23 EGFR FISH+ and 68 overall) and would take 4% longer to complete.
For the subgroup-focused multiple-hypothesis designs, the power to evaluate the secondary coprimary objective in the overall population is 84% for the 80% split design and 90% for the 60% split design. The power to evaluate the objective in the EGFR FISH$^+$ group in the overall-population–focused design is 81%.

S0819: The Design

For S0819, the design chosen to study the effect of cetuximab added to carboplatin and paclitaxel, with variable inclusion of bevacizumab, plus the role of EGFR FISH as a predictive biomarker, was the subgroup-focused multiple-hypothesis design with 80% allocation of type I error to the subgroup hypothesis. This design was chosen in part because the total sample size and time to completion were only slightly larger than those for the all-comers design, and the design was able to address the biomarker hypothesis as a primary objective.

The design of S0819 does not require that a patient’s tissue specimen be analyzed by the time he or she is registered in the study. The decision was based primarily on clinical pragmatism in the setting of advanced-stage cancer. Because this study is being conducted in an advanced-disease setting, both the patients and their physicians typically want to begin treatment as soon as possible. Without the data to support the need for a delay in treatment while waiting for marker determination, this approach was considered to be in the best interest of the patients and avoid a deterrent to accrual. The only detriment to this approach is that randomization cannot be stratified on biomarker status. However, given that the trial size is quite large, stratification is unlikely to be needed. We also determined that it is important to include patients with unknown marker values because the coprimary objective is in the overall unselected population.

Interim Monitoring

Once the design of the study was finalized, the next step was to determine the interim monitoring plan. Four possible decisions could be made at each of the interim time points: The study could continue in the entire study population, continue in the EGFR FISH$^+$ group alone, continue in the EGFR FISH-nonpositive group alone, or close to further accrual in both the positive and nonpositive groups.

To determine which monitoring plan to use, we had to consider the following questions: Should efficacy be monitored in both the entire study population and the marker-positive group or just in the marker-positive group? Should futility monitoring be carried out in both groups or in just one of the groups? Although the study is not designed around an evaluation of treatment efficacy in the nonpositive group, should this group be evaluated for efficacy or futility in the interim analyses?

The efficacy-monitoring plan chosen was to monitor the EGFR FISH$^+$ group alone and to recommend closure of the entire study if the null hypothesis is rejected in this group. This approach was chosen because the design is focused on the subgroup. Further, if the EGFR FISH$^+$ portion of the study is closed, the objective within the entire study population is not evaluable past this analysis. Therefore, continuation of the study in the nonpositive group alone would require a redesign of the trial. S0819 is not powered to evaluate a treatment effect in the EGFR FISH$^-$ group; moreover, it is unclear what level of testing should be used to test an effect in this group should the EGFR FISH$^+$ hypothesis be rejected.

Determination of the futility-monitoring plan was more complicated. We believed that although the design warranted a somewhat conservative monitoring plan to allow both hypotheses to be addressed, this approach should not be followed at the expense of any subgroup of patients. Because this design uses 1-sided testing, the evaluation of

### Table 1. Comparison of designs with 90% power for the primary hypothesis and 1-sided study-wide type I error of 2.5%

<table>
<thead>
<tr>
<th>Sample size/number screened</th>
<th>FISH$^+$</th>
<th>Overall</th>
<th>FISH$^+$ /unknown</th>
<th>Time to completion (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-hypothesis designs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-comers</td>
<td>541</td>
<td>1,352</td>
<td>811</td>
<td>57</td>
</tr>
<tr>
<td>Targeted design</td>
<td>556</td>
<td>1,390a</td>
<td>N/A</td>
<td>58</td>
</tr>
<tr>
<td>Strategy design</td>
<td>1,135</td>
<td>2,838</td>
<td>1,703</td>
<td>105</td>
</tr>
<tr>
<td><strong>Multiple-hypothesis designs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroup: 80%</td>
<td>588</td>
<td>1,462b</td>
<td>874</td>
<td>60</td>
</tr>
<tr>
<td>Subgroup: 60%</td>
<td>626</td>
<td>1,558c</td>
<td>932</td>
<td>63</td>
</tr>
<tr>
<td>Overall: 80%</td>
<td>564d</td>
<td>1,420</td>
<td>856</td>
<td>59</td>
</tr>
</tbody>
</table>

aNumber to be screened.
bPower is 84%.
cPower is 90%.
dPower is 81%.
futility is framed in terms of rejecting or failing to reject the alternative hypothesis. For futility monitoring in the EGFR FISH\(^+\) population, the question was, should decisions be based on this group alone, or should information from the overall-population assessment be included? That is, should futility for assessing the marker-driven hypothesis be determined if the alternative hypothesis is rejected in the marker-positive group alone, or should futility be determined only if both the marker-positive and the overall-population alternatives are rejected?

The next issue to address was how to determine stopping for futility related to the overall study population hypothesis. Using the same logic employed to determine the efficacy-monitoring approach, we determined that the entire study should be closed for futility should futility be established in the EGFR\(^+\) group. However, failing to stop for either efficacy or futility in the EGFR FISH\(^+\) group should not necessarily mean that the study should continue in the nonpositive group. But how should futility be determined in the nonpositive subgroup? Should the evaluation be based on an evaluation in the overall study population, the EGFR FISH-nonpositive population, or some combination of the two?

Consistent with the plan to use a moderately conservative approach for monitoring, it was decided that all monitoring plans should use the standard SWOG approach for interim monitoring. Boundaries for testing are defined on a fixed-sample, \(P\)-value scale, and hypotheses are tested at one tenth of the overall level. The same level was used to evaluate both the nonpositive and positive groups. Therefore, interim testing for efficacy and futility in the EGFR FISH\(^+\) group and futility alone in the nonpositive group was specified at the 0.002 level, and for the overall population interim futility testing was specified at the 0.0008 level. To account for the effect of efficacy testing, the level for testing the null hypothesis in the EGFR FISH\(^+\) group in the final analysis was 0.018. Levels in the overall and nonpositive groups were not adjusted because they were only evaluated for futility. To determine efficacy in the overall population, the null hypothesis will be tested at the 0.008 level in the final analysis, whether this is an interim analysis time with stopping based on the EGFR FISH\(^+\) group or at full completion of the trial. Figure 1 depicts the stopping boundaries on the HR scale with values \(>1\) representing benefit and \(<1\) representing harm. The upper line represents the efficacy boundary for the EGFR FISH\(^+\) group. The lower lines represent the futility boundaries for the EGFR FISH\(^+\) group (thin line), the entire study (thick line), and the EGFR FISH-nonpositive group (dashed line).

Figure 1. Stopping boundaries on the HR scale for S0819, a subgroup-focused, multiple-hypothesis design.
Simulation Study to Evaluate Monitoring Plans

To address the above questions and determine the rules for stopping for futility using these boundaries, we simulated the study design under 5 scenarios. For each scenario, PFS data were simulated under an exponential distribution using the SWOG S0819 parameters, with 588 EGFR+ patients and 1,462 total patients included, and equal probability of randomization to the experimental versus control arm. Registration times were generated from a uniform distribution over a 48-month accrual period. For each scenario, 10,000 sets of data were generated. The HRs specified in terms of the percentage of improvement in median PFS used to generate the experimental arm data for the EGFR FISH+ and EGFR FISH-nonpositive groups and the associated overall population HRs are presented in Table 2.

In the first scenario, both alternative hypotheses are true: There is a significant effect in both the EGFR FISH+ group and overall, indicating a modest effect in the nonpositive group. In the second and third scenarios, the treatment effect is restricted to the positive group and there is no effect in the negative group. The effect is modest in the overall population for the second scenario, and equal to the design effect of a 20% improvement in the third scenario. In the fourth scenario, there is an effect overall, but EGFR FISH is not predictive. The fifth scenario specifies no effect in any group.

Interim analyses were performed at 40%, 60%, and 80% of the expected progression events. For each of the scenarios, the frequency and properties of stopping the EGFR FISH-nonpositive group were determined based on an evaluation of the overall-population hypothesis alone, the associated hypothesis within the EGFR FISH-nonpositive group alone, and stopping in the nonpositive group if either the overall-population or the nonpositive-group alternative hypothesis was rejected. Table 3 presents the frequency of rejecting the null hypothesis, stopping early for futility, and the average sample size, percentage of information, and study time under the different scenarios and stopping rules only for the EGFR FISH+ group.

The proposed monitoring plan within the EGFR FISH+ group performed as expected. In the first and second scenarios, which specified the same alternative for the positive group, the study stopped early for efficacy 59% of the time, and retained the power to reject the null hypothesis and stopped for futility 0.6% of the time. For the scenario with the larger-than-specified effect (a 50% improvement), the trials were stopped early for efficacy and futility 74.5% and 0.5% of the time, respectively. Under the smaller effect size of a 20% improvement (scenario 4), the percentage of early stops was 20.6% for efficacy and 5.7% for futility. Under the null hypothesis (scenario 5), the percentage of early stops was 0.4% for efficacy and 61.4% for futility.

The properties of the 3 candidate futility-monitoring plans for the nonpositive group did not vary significantly across the scenarios (Table 4). In the first scenario, monitoring the overall population alone was stopped for futility 0.8% of the time, 0.2% more often than monitoring in the EGFR FISH+ group alone. Monitoring the EGFR FISH+ group alone was stopped for futility 1% of the time, and monitoring both groups was stopped 1.1% of the time. There was a significant decrease in power from 82.5% to 74.3% as a result of early stopping for efficacy in the EGFR FISH+ group, and there was no difference across the 3 candidate plans. When the benefit was restricted to the positive group, the percentage of early stops for futility based on monitoring the overall group alone, the nonpositive group alone, and the combined evaluation was 3.2%, 7.5%, and 8.3% in scenario 2 and 0.6%, 2.9%, and 3.0% for

### Table 2. HRs for simulation scenarios to evaluate interim monitoring for the SO819 design

<table>
<thead>
<tr>
<th>Scenario</th>
<th>EGFR FISH+</th>
<th>Overall population</th>
<th>EGFR FISH-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>1.33</td>
<td>1.2</td>
<td>1.11</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>1.33</td>
<td>1.13</td>
<td>1.0</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>1.5</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>1.2</td>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Table 3. Properties of the interim monitoring plan in the EGFR FISH+ group for the subgroup-focused multiple-hypothesis design with 90% power to detect a 33% improvement in median PFS and 1-sided 2% level type I error

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Design powera</th>
<th>Rejectb H0 (%)</th>
<th>Early stops</th>
<th>Stops for efficacy</th>
<th>Stops for futility</th>
<th>Sample size</th>
<th>Events (%)</th>
<th>Study time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 and 2</td>
<td>90.1</td>
<td>89.5</td>
<td>59.6</td>
<td>59.1</td>
<td>0.6</td>
<td>498</td>
<td>75</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>99.6</td>
<td>99.5</td>
<td>75.0</td>
<td>74.5</td>
<td>0.5</td>
<td>483</td>
<td>70</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>53.6</td>
<td>52.6</td>
<td>26.4</td>
<td>20.6</td>
<td>5.7</td>
<td>553</td>
<td>90</td>
<td>54</td>
</tr>
<tr>
<td>5</td>
<td>2.0</td>
<td>1.9</td>
<td>61.8</td>
<td>0.4</td>
<td>61.4</td>
<td>487</td>
<td>74</td>
<td>44</td>
</tr>
</tbody>
</table>

aThe proportion of times the null hypothesis was rejected accounting for interim monitoring.

bProportion of time the null hypothesis was rejected if the study went to complete accrual with no interim stops.
scenario 3, respectively. The candidate plans were essentially identical when the effect was consistent across the subgroups (scenario 4), stopping for futility 5.8% of the time, and had only a modest reduction in power from 84% to 81%. In the completely null situation (scenario 5), the trial was stopped based on assessment in the EGFR FISH+ group alone 61.8% of the time, and it was this evaluation that accounted for the majority of early stops for futility in the EGFR-nonpositive group. Monitoring in the overall population alone was stopped for futility 9.3% more often, monitoring the nonpositive group alone was stopped 3.1% more often, and using the combined evaluation was stopped 9.5% more often than stopping based purely on EGFR FISH+ futility monitoring.

Given that the choice of monitoring plan did not seem to affect the power for the overall hypothesis substantially across the 3 futility-monitoring plans in the nonpositive group, the final design included the plan that performed the best across all scenarios. Therefore, the final plan defines futility based on the combined evaluation in the overall and nonpositive groups.

## Table 4. Properties of the 3 candidate monitoring plans for the overall-population component of the subgroup-focused, multiple-hypothesis design with 80% type I error allocation to the subgroup

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Design power</th>
<th>Reject H₀(%)</th>
<th>Early stops</th>
<th>Result at early stop</th>
<th>Average Stopping</th>
<th>% Events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>% Early</td>
<td>% Futility</td>
<td>Reject H₀</td>
<td>Inconclusive</td>
</tr>
<tr>
<td>Scenario 1</td>
<td>82.5</td>
<td>74.3</td>
<td>59.8</td>
<td>0.8</td>
<td>0.3</td>
<td>12.5</td>
</tr>
<tr>
<td>Overall only</td>
<td>74.3</td>
<td>FISH+ only</td>
<td>59.8</td>
<td>1.0</td>
<td>0.1</td>
<td>12.7</td>
</tr>
<tr>
<td>Either</td>
<td>74.3</td>
<td>59.9</td>
<td>1.1</td>
<td>0.2</td>
<td>12.7</td>
<td>47.0</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>44.3</td>
<td>32.9</td>
<td>61.9</td>
<td>3.2</td>
<td>2.9</td>
<td>34.6</td>
</tr>
<tr>
<td>Overall only</td>
<td>32.9</td>
<td>FISH+ only</td>
<td>63.0</td>
<td>7.5</td>
<td>1.7</td>
<td>37.0</td>
</tr>
<tr>
<td>Either</td>
<td>32.9</td>
<td>63.7</td>
<td>8.3</td>
<td>2.5</td>
<td>36.9</td>
<td>24.3</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>80.4</td>
<td>58.4</td>
<td>75.0</td>
<td>0.6</td>
<td>0.1</td>
<td>36.7</td>
</tr>
<tr>
<td>Overall only</td>
<td>58.4</td>
<td>FISH+ only</td>
<td>75.0</td>
<td>2.9</td>
<td>0.1</td>
<td>36.7</td>
</tr>
<tr>
<td>Either</td>
<td>58.4</td>
<td>75.0</td>
<td>3.0</td>
<td>0.1</td>
<td>36.7</td>
<td>38.3</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>83.9</td>
<td>80.8</td>
<td>26.5</td>
<td>5.8</td>
<td>0.2</td>
<td>6.4</td>
</tr>
<tr>
<td>Overall only</td>
<td>80.8</td>
<td>FISH+ only</td>
<td>26.4</td>
<td>5.8</td>
<td>0.1</td>
<td>6.4</td>
</tr>
<tr>
<td>Either</td>
<td>80.8</td>
<td>26.5</td>
<td>5.9</td>
<td>0.2</td>
<td>6.4</td>
<td>19.9</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>1.0</td>
<td>1.0</td>
<td>71.6</td>
<td>71.1</td>
<td>39.5</td>
<td>32.0</td>
</tr>
<tr>
<td>Overall only</td>
<td>1.0</td>
<td>FISH+ only</td>
<td>65.3</td>
<td>64.9</td>
<td>32.3</td>
<td>32.9</td>
</tr>
<tr>
<td>Either</td>
<td>1.0</td>
<td>71.7</td>
<td>71.3</td>
<td>39.1</td>
<td>32.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*The proportion of times the null hypothesis was rejected if the study went to complete accrual with no interim stops.

bProportion of time the null hypothesis was rejected accounting for interim monitoring.

## Table 5. Probability of rejecting either null hypothesis in the subgroup-focused, multiple-hypothesis design with 90% power in the subgroup and 80% type I error allocation in the subgroup for the 3 candidate futility-monitoring plans

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Minimal power</th>
<th>Overall alone</th>
<th>FISH alone</th>
<th>Either</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.0</td>
<td>93.7</td>
<td>92.5</td>
<td>92.4</td>
</tr>
<tr>
<td>2</td>
<td>90.3</td>
<td>89.9</td>
<td>89.7</td>
<td>89.7</td>
</tr>
<tr>
<td>3</td>
<td>99.6</td>
<td>99.5</td>
<td>99.5</td>
<td>99.5</td>
</tr>
<tr>
<td>4</td>
<td>86.1</td>
<td>83.9</td>
<td>82.5</td>
<td>80.6</td>
</tr>
<tr>
<td>5</td>
<td>2.7</td>
<td>2.6</td>
<td>2.9</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Discussion

Clearly, there is no one "right" design for addressing the variety of issues involved in assessing a potential predictive biomarker of cancer therapy, but the choice of trial design in each case must account for what is known and what remains unknown in specific clinical scenarios (17). Each of the designs considered for S0819 has associated benefits and statistical costs. Although the chosen subgroup-focused multiple-hypothesis design is almost as efficient as the all-comers design (i.e., similar sample size and time to completion), the cost of the multiple-hypothesis design is a reduction in power in the overall population.

A possible complication of the S0819 design is that we must decide what to conclude about EGFR FISH as a predictive biomarker if both hypotheses are rejected in favor of the alternative. Because the power to detect differences in the HRs specified in this design is only 50% using a 1-sided 0.05 level test, it is highly unlikely that an interaction will be detected, even if it exists. However, this does not reduce the value of significant findings. In this case, the focus should be on whether the statistically significant effects are clinically meaningful effects. The sample size for both subgroups based on biomarker values (positive, negative, and unknown) are relatively large and will provide good measures of precision of the estimates. These data can be used to evaluate whether the EGFR FISH group derives a greater benefit from cetuximab.

The multiple-hypothesis design chosen for this study retains flexibility to evaluate additional biomarkers, perhaps even some that are discovered or become important after the trial is initiated. Subsequent to activation of S0819, both the BMS099 and FLEX studies reported on the potentially predictive effect of EGFR as measured by FISH, IHC, and mutational status (18–20). Although the majority of biomarkers evaluated were not found to be associated with cetuximab efficacy, in FLEX, high tumor EGFR expression (≥200 on a scale of 0–300) was found to be significantly associated with cetuximab efficacy.

Another key aspect of the trade-offs among the various designs is the prevalence of the biomarker (21, 22). For the subgroup-focused, multiple-hypothesis designs, the power to detect a specific HR in the overall study population increases with decreasing prevalence as a result of the need to enroll more patients. That said, the specified HR in the overall population should likely be decreased with decreasing marker prevalence, or the treatment effect in the marker-positive group should be increased. For example, if the prevalence is only 20%, then a 33% improvement in the positive group and an 11% improvement in the nonpositive group would translate to a 15% improvement overall. Thus, a subgroup-focused, multiple-hypothesis design would require the accrual of 2,846 patients to accrue 572 EGFR FISH patients over a 94-month period. The residual type I error with 20% marker prevalence is 0.6%, and therefore with 2,846 patients, the study would have 88% power to detect a 15% improvement in median PFS. Alternatively, assuming that the prevalence is 20% and retaining the assumed 20% improvement overall and 11% improvement in the nonpositive group would result in a target of 56% improvement in the marker-positive group. With the larger effect size for EGFR FISH patients, a total of 1,212 patient accruals are needed to accrue 248 marker-positive patients. However, the design now has only 72% power to detect a 20% improvement in median PFS.

Conclusions

In conclusion, we have discussed possible study designs specific to our disease setting and trial-specific assumptions. After carefully evaluating the properties of various designs, we conclude that the multiple-hypothesis design selected for SWOG S0819 is well suited to robustly address the variety of possible scenarios in this clinical setting and to provide meaningful answers to the S0819 study questions.

Disclosure of Potential Conflicts of Interest

F.R. Hirsch is a member of the consultative/advisory boards of BMS, Genentech/ROCHE, Merck Serono, and Imclone/Lilly. No other potential conflicts of interest were disclosed.

Authors’ Contributions

Conception and design: M.W. Redman, J.J. Crowley, R.S. Herbst, F.R. Hirsch, D.R. Gandara

Development of methodology: M.W. Redman, R.S. Herbst, F.R. Hirsch

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M.W. Redman, R.S. Herbst, F.R. Hirsch, D.R. Gandara

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M.W. Redman, R.S. Herbst, F.R. Hirsch

Writing, review, and/or revision of the manuscript: M.W. Redman, J.J. Crowley, R.S. Herbst, F.R. Hirsch, D.R. Gandara

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M.W. Redman

Study supervision: F.R. Hirsch

Acknowledgments

We thank Michael LeBlanc for his helpful comments on this manuscript and discussions during the design of this trial. We also thank James Moon for his input on the trial design.

Grant Support

National Cancer Institute, National Institutes of Health (PHS Cooperative Agreement/DHHS grants CA32102, CA38926, CA46441, CA105409, CA42777, and NIH CA090998).

Received January 17, 2012; revised April 5, 2012; accepted May 2, 2012; published OnlineFirst May 16, 2012.

References


Design of a Phase III Clinical Trial with Prospective Biomarker Validation: SWOG S0819


Clin Cancer Res  Published OnlineFirst May 16, 2012.

Updated version  Access the most recent version of this article at: doi:10.1158/1078-0432.CCR-12-0167
Supplementary Material  Access the most recent supplemental material at: http://clincancerres.aacrjournals.org/content/suppl/2012/07/30/1078-0432.CCR-12-0167.DC1

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.
Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.
Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.