Multiple Histology Phase II Trials
Michael LeBlanc,1,2 Cathryn Rankin,1,2 and John Crowley1,2

Abstract
Purpose: Testing agents in cancers with multiple disease subtypes, in which the activity of a new treatment may vary between subtypes, presents statistical and logistical challenges. We propose a flexible phase II strategy which includes both analyses for each histology or stratum and a combined analysis which borrows information from all the patients in the study. Sequential futility analyses are conducted once each subgroup or the overall group reaches a specified minimum accrual.

Experimental Design: Examples based on a soft tissue sarcoma phase II trial, which includes multiple histologies and simulation studies, are used to assess the statistical properties of the proposed strategy.

Results: The combined analyses in one phase II trial lead to smaller expected sample sizes when the drug is broadly inactive, and to greater statistical power if there is modest activity across multiple strata as compared with conducting several smaller phase II studies. In addition, by retaining the stratum-specific tests, the design allows the identification of subgroups for which the agents are most active.

Conclusion: To consider phase II testing with multiple biological subtypes, a strategy which combines both the individual subgroup tests and overall combined tests has promising statistical properties. Our results support the appropriate use of statistical borrowing of information in phase II studies in this setting. More broadly, this work fits the paradigm that phase II studies should include as broad a group of patients as scientifically reasonable, but incorporate design considerations for subsets of patients with potentially differing responses to therapy.

Background
Many cancers are complex and heterogeneous diseases with multiple subtypes diagnosed either histologically or by molecular methods. For example, soft tissue sarcomas have multiple subtypes, including angiosarcomas, hemangiosarcomas, hemangiopericytomas, leiomyosarcomas, and high-grade liposarcomas. Non-Hodgkin’s lymphomas also have a significant number of subtypes within either the categories of B-cell or T-cell neoplasms. In addition, genomic measurements such as gene expression or comparative genomic hybridization on these tumors can now lead to the refinement or regrouping of patients for clinical studies. Our interest is to acknowledge, in a simple fashion, this biological heterogeneity in phase II clinical trials.

Phase II trials are used to screen new regimens for activity and to decide which ones should be tested further. For targeted agents, there is the additional complexity that activity of the agent may depend on the tumor expressing the appropriate target, which may be expected to vary among known subtypes of the disease. If the suitable subgroup of patients is identified, standard phase II designs should typically restrict eligibility for the study to those patients. For instance, typical phase II hypotheses are $H_0: p = p_0$ versus $H_1: p = p_1$, with probabilities specified such that if the true probability of clinical response were $p_0$, the regimen would not be of interest, whereas a true probability $p_1$ would warrant further testing. Note that for many new targeted agents, the end points may need to be modified to include disease control rate or progression-free survival at some time point. Generally, designs are specified to have some significance level, e.g., 0.05 and power 0.8 to 0.9, and two (or more) stages of accrual for ethical reasons and to decrease expected sample size, with early stopping if insufficient activity is seen (1, 2). Acceptance and rejection bounds are often determined according to optimality criteria (such as minimization of sample size under the null hypothesis).

Yet, phase II studies should allow us to learn as much as possible for a potential phase III study. Therefore, if significant uncertainty in the drug target exists, we believe it is reasonable to use phase II strategies that are inclusive with respect to the patient population, but with appropriate subgroups acknowledged in the designed hypothesis testing and subsequent analyses. There are many possible testing and estimation strategies that could incorporate flexibility in phase II studies with respect to subgroups. To focus the discussion, we consider a specific case and statistical properties of using both a subgroup and overall testing for a soft tissue sarcoma phase II study that was coordinated by the Southwest Oncology Group and the Intergroup Coalition against Sarcoma. After showing the example, we describe the strategy in somewhat more general terms and

Authors’ Affiliations: 1Fred Hutchinson Cancer Research Center, Seattle, Washington and 2Cancer Research and Biostatistics, Seattle, Washington

Received 8/6/08; revised 3/24/09; accepted 4/15/09; published OnlineFirst 6/23/09.

Grant support: National Institutes of Health grants CA090998 and CA38926.

Requests for reprints: Michael LeBlanc, Fred Hutchinson Cancer Research Center, Southwest Oncology Group Statistical Center, 1100 Fairview Avenue, M3-C102, P.O. Box 19024, Seattle, WA 98109. Phone: 206–667–6087; Fax: 206–667–6089; E-mail: mleblanc@fhcrc.org.

© 2009 American Association for Cancer Research. doi:10.1158/1078-0432.CCR-08-2069

Clin Cancer Res 2009;15(13) July 1, 2009 4256 www.aacrjournals.org
investigate the statistical properties of this simple phase II method that both tests subgroups and borrows across groups.

**BAY 43-9006 in Soft Tissue Sarcoma**

The compound BAY 43-9006 has multiple targets that may be of relevance in soft tissue sarcomas. Sarcomas have been shown to express platelet-derived growth factor receptor (PDGFR) and vascular endothelial growth factor (VEGF) receptor (VEGFR), with mixed data on the correlation of VEGF and VEGFR expression as a predictor of clinical outcome. Human angiosarcomas, hemangiosarcomas, and hemangiopericytomas have all been shown to express VEGFR and the mRNA for VEGF. This suggests that there may be autocrine or paracrine growth stimulation in these tumors which might be inhibited by BAY 43-9006. In addition, there are reports of PDGFR expression in these tumor types as well.

In developing a phase II study, the Southwest Oncology Group and the Intergroup Coalition against Sarcoma selected the most common high-grade adult soft tissue sarcomas, leiomyosarcoma and liposarcoma, because of the known presence of VEGFR and PDGFR in these tumor types, as well as the potential importance of VEGF and VEGFR expression with prognosis. In addition, there was increasing preclinical data which suggested that targeting both VEGF and PDGFR may be very effective in causing the regression of tumor-associated vasculature.

The primary objective of the phase II study was to assess the clinical response probability (confirmed complete response and partial response) in patients with angiosarcoma, hemangiosarcoma, hemangiopericytoma, high-grade leiomyosarcoma, or high-grade liposarcoma. Our design-specified response would be of interest for the combined histologies or overall study. Within histologic subtypes, angiosarcoma, hemangiosarcoma, and hemangiopericytoma were grouped into a single stratum due to the expected low accrual within these subtypes.

The design actually chosen for the study had similarities to the strategy proposed in this report, except that the alternative hypothesis was only tested at the time of the first stage sample size and at final analysis. Instead, here we conduct repeated testing of the alternative hypothesis, \( H_A \), to close the study for futility after every five patients are accrued after reaching sufficient minimal accrual. We note that if accrual was rapid, any significant delay in response assessment would limit the reduction in sample size by using more frequent futility analyses.

We report on the properties of this study using the strategy developed and studied in this article. Within each of the three strata, a maximum of 25 patients was to be accrued. The first testing of the alternative hypothesis was to be conducted after the first 15 eligible patients registered. Once 30 eligible patients, including all strata, have been accrued, testing of the alternative hypothesis for the entire group of patients was to be conducted to determine if the study should be closed early because of overall minimal activity combined across subtypes of the disease. A maximum of 75 patients was to be accrued to the study.

At the time the study was designed, the following subtype frequencies were assumed: leiomyosarcoma (50%), liposarcoma (30%), and angiosarcoma/hemangiosarcoma/hemangiopericytoma (20%). It was determined that a response probability of \( \theta_{0} = 0.05 \) would not be of interest; however, within a stratum, a response probability of \( \theta_{A} = 0.25 \) would indeed be of interest. We considered that a lesser response probability of \( \theta_{0.15} = 0.15 \) would be of interest for the combined histologies or overall study data. Futility analyses based on testing of the alternative hypotheses were done with error specified by \( \alpha = 0.02 \), and testing of the null was type I error specified by \( \beta = 0.05 \).

The tables below give power and expected sample sizes. The following columns describe design properties for the given subgroup frequency (Accrual) and clinical response probability (Response). \( N(\text{Simp}) \) is the expected sample size for each subgroup at the end of the study using only subgroup or histology-specific testing. Power(\text{Simp}) is the power using only subgroup testing or histology-specific testing. Therefore, \( N(\text{Simp}) \) and Power(\text{Simp}) correspond to conducting individual phase II studies in each histology. The next two numbers, \( N(\text{Comb}) \) and Power(\text{Comb}), show the effect of adding a futility test based on

### Table 1. Simple subgroup and combined strategy designs

<table>
<thead>
<tr>
<th>Accrual</th>
<th>Response</th>
<th>N(\text{Simp})</th>
<th>Power(\text{Simp})</th>
<th>N(\text{Comb})</th>
<th>Power(\text{Comb})</th>
<th>Cond Prob</th>
<th>Min Prob</th>
<th>Max Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Null case</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.05</td>
<td>20.01</td>
<td>0.04</td>
<td>18.77</td>
<td>0.03</td>
<td>0.38</td>
<td>0.31</td>
<td>0.46</td>
</tr>
<tr>
<td>0.30</td>
<td>0.05</td>
<td>20.20</td>
<td>0.03</td>
<td>16.61</td>
<td>0.03</td>
<td>0.31</td>
<td>0.41</td>
<td>0.39</td>
</tr>
<tr>
<td>0.20</td>
<td>0.05</td>
<td>20.77</td>
<td>0.04</td>
<td>14.89</td>
<td>0.03</td>
<td>0.32</td>
<td>0.38</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Experiment-wise Power(\text{Simp}) = 0.1042, Experiment-wise Power(\text{Comb}) = 0.0916

| (B) Efficacy in frequent histology |
| 0.50 | 0.25 | 24.86 | 0.90 | 24.68 | 0.90 | 0.99 | 0.00 | 0.99 |
| 0.30 | 0.05 | 20.00 | 0.03 | 19.96 | 0.04 | 0.05 | 0.56 | 0.01 |
| 0.20 | 0.05 | 20.44 | 0.03 | 20.22 | 0.04 | 0.05 | 0.59 | 0.01 |

Experiment-wise Power(\text{Simp}) = 0.9026, Experiment-wise Power(\text{Comb}) = 0.9054

| (C) Limited efficacy in all histologies |
| 0.50 | 0.15 | 24.01 | 0.53 | 23.96 | 0.53 | 0.61 | 0.37 | 0.40 |
| 0.30 | 0.15 | 23.93 | 0.52 | 23.80 | 0.52 | 0.60 | 0.38 | 0.38 |
| 0.20 | 0.15 | 24.01 | 0.52 | 23.83 | 0.52 | 0.60 | 0.37 | 0.39 |

Experiment-wise Power(\text{Simp}) = 0.8904, Experiment-wise Power(\text{Comb}) = 0.9042

NOTE: Min Stratum = 15, Max Stratum = 25, Min Combined = 30, Max Combined = 75.
combining all the histologies together and stopping if there is limited activity overall in the study. Therefore, N(Comb) is the sample size using subgroup efficacy testing but with combined or overall group futility testing and Power(Comb) is power for subgroup efficacy test with combined futility testing. The last three columns correspond to the addition of overall combined efficacy testing where the expected sample sizes remain as N(Comb) because early futility stopping is unchanged from the previous calculations. Cond Prob is an estimate of the conditional probability of rejecting the null hypothesis for the specific stratum, given that either the overall or the stratum-specific hypotheses have been rejected. Therefore, in the case that there are quite different response probabilities between strata, it gives an indication if one would also reject the null hypothesis for the stratum with the largest response rate. Finally, we present the estimates of the probability that the stratum response estimate is the largest or tied for the largest (Max Prob) or smallest (MinProb) estimate among all strata if the given stratum-specific or overall test is rejected. The title captions give the target accrual and the experiment-wise power across all subgroups in the phase II study.

The combined study design, allowing stopping for futility based on combining histologies, leads to a reduction in ~18% of patient accrual (50.27 compared with 60.98 estimated accrual) if BAY 43-9006 were ineffective in all subtypes, as shown in Table 1A. The largest reduction in accrual is achieved for the more slowly accruing or less frequent disease subgroups. Note that the type I error for each stratum is less than $\alpha_0 = 0.05$ due to the additional futility testing. We note that the last three columns are not informative in Table 1A because they are estimated conditional on rejecting the null hypothesis. The experiment-wise type I error is inflated due to the multiple strata, but are similar for simple separate stratum analyses 10.4% and with the combined strategy 9.1%.

In Table 1B, the case in which the compound was only effective in the common subtype, the caption indicates there is approximately the same experiment-wise power for the combined strategy compared with testing on all subgroups individually. In addition, the conditional probability of rejecting the stratum-specific test for the effective subgroup given that rejection for the combined strategy is 99%. Furthermore, the estimates obtained from each strata are in agreement with the truth, with 99% of simulated studies leading to the largest response rates in the active stratum, and for the other two strata estimated response rates are lowest (or tied for lowest) 56% and 59% of cases, respectively.

Finally, Table 1C shows whether BAY 43-9006 is approximately equally effective in all subtypes; the combined strategy yields a power of 90%; this is only slightly higher than the probability of rejecting at least one of the individual strata (89%). However, for a given stratum, there is only ~52% to 53% chance of rejecting the null hypothesis, potentially leading to falsely concluding negative results for one or more subgroups, if one were to use only stratum-specific testing. In summary, for the sarcoma study, the subgroup/combined phase II strategy leads to smaller sample sizes under the null and improved power when there is limited efficacy across subtypes.

We also investigated the effect of a hypothetical design in which we assumed that the most frequent and most infrequent subgroups were combined to make a targeted subgroup containing ~70% of patients. Such a design would be useful if it were known (or strongly believed) that those subgroups or histologies were most biologically suited to a particular treatment. For instance, tumors in those subgroups might express the target molecule most highly among all histologies. Under this design, given that we assume the efficacy will be greatest in the targeted group, evidence of a lack of efficacy for the targeted group implies overall futility, not just subgroup futility. Therefore, the entire study would be closed if the targeted subgroup alternative test is rejected. Testing of the alternative hypothesis within the subgroup was conducted as described below (Targeting a subgroup), with an alternative response probability of $\theta_0 = 0.20$. Under the null hypothesis, the targeted design further reduced the sample size compared with either the subgroup or combined analysis as shown in Table 2A. This result is due to the larger alternative parameter used in the targeted test compared with the overall combined test. In addition, in Table 2B, if the targeted strata had a response probability of 0.20, and the remaining strata had a response rate of 0.05, the simulation study showed that the targeted hypothesis significantly increased power (88%; type I error of 0.02) over individual subgroup analyses (~76%; type I errors of 0.03-04) for the two strata corresponding to promising subgroups. The experiment-wise power of rejecting at least one of the strata was modestly higher than the experiment-wise power of the strategy including targeting, but this was, in part, due to the larger type I error of the untargeted strategy compared with the targeted strategy (9.8% versus 6.1%). The reduction in type I error for the targeted method is due to the more direct futility testing that stops overall accrual if there is futility in the target subgroup, compared with considering each stratum separately in the untargeted strategy.

The General Case: Multiple Subtype Data and a Phase II Design

Now consider the more general case of a phase II study being planned in a disease with multiple subtypes, in which the activity of a new treatment may vary between subtypes. We will investigate a strategy that both tests within each subtype and combined across subtypes. Denote a model for the outcome of

<table>
<thead>
<tr>
<th>Table 2. Simple subgroup and targeted designs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accrual</strong></td>
</tr>
<tr>
<td>(A) Null case</td>
</tr>
<tr>
<td>0.50</td>
</tr>
<tr>
<td>0.30</td>
</tr>
<tr>
<td>0.20</td>
</tr>
<tr>
<td>Experiment-wise Power(Simp) = 0.0982</td>
</tr>
<tr>
<td>(B) Effective subgroup</td>
</tr>
<tr>
<td>0.50</td>
</tr>
<tr>
<td>0.30</td>
</tr>
<tr>
<td>0.20</td>
</tr>
<tr>
<td>Experiment-wise Power(Simp) = 0.9462</td>
</tr>
</tbody>
</table>

NOTE: Min Stratum = 15, Max Stratum = 25, Min Combined = 30, Max Combined = 75.
interest for K subgroups or disease subtypes \( R_k \); \( k = 1, \ldots, K \), where individuals in subgroup \( k \) have an outcome distribution model represented by \( F(\theta_k) \) and indexed by a parameter \( \theta_k \). We focus on the case of binary response data or disease control rate at some specific time point, but the same methods could easily be extended to time-to-event patient outcomes. The different subtypes of patients are assumed to register at a rate proportional to the expected frequencies of the subtypes, \( v_k \).

Consider a sequential design, motivated by two-stage phase II designs, and assume that at least \( n_k^1 \) patients will be accrued to any stratum before considering testing for futility against the alternative hypothesis, \( \theta_A \), and that a maximum number of patients, \( n_k \), can be accrued to any stratum. In addition to the subgroup analysis, the patients on the study, as a whole, are combined and the design tests for futility when a minimum \( n^1 \) patients has been accrued. The maximum size of the study is \( n \) patients, where \( n \) is less than or equal to the number of patients which would be accrued at full accrual for all strata, \( n_0 \).

Although one could limit tests of the alternative hypothesis to be conducted only at the time the study reaches the minimum number of patients in the stratum \( n_k^1 \) or the minimum number of patients including all groups reaches \( n^1 \), we choose to test the alternative hypothesis after each \( q \) (5 or 10) patients are registered to the study once the minimum sample size constraints are satisfied. We note that, in practice, one may test the hypothesis when each group of \( q \) patients are evaluated for response (e.g. 4 months after registration) rather than just based on accrual. In addition, the strategy would also work alternative end points (such as disease control at 2-4 months) or at alternative interim futility analysis times only when each stratum reaches \( n_k^1 \) or the overall number of patients reaches \( n^1 \).

The binomial tail probabilities are calculated under the alternative distribution

\[
\Pr(X \leq r_k | n_k, \theta_A) = \Pr(X \leq r_k | n_k, \theta_A)
\]

where \( r_k \) is the number of successes when \( n_k \) patients are assessed for response at analysis time \( j \), where \( n_k \geq n_k^1 \). If the subgroup test is rejected at an analysis, \( T_k^A < x_0 \), testing at level \( x_0 \), then accrual to that stratum is discontinued, but continues for the remaining strata. To improve the probability of stopping early in the event of limited (or no) improved activity, the data are combined across the subtypes. The combined test statistic is calculated using binomial tail probabilities

\[
T_A = \Pr(X \leq r_j | n_j, \theta_A)
\]

where \( r_j \) is the number of successes when \( n_j \) patients are assessed for response at analysis time \( j \). As in the subgroup tests, the alternative overall test is calculated only after a prespecified number of patients are accrued, \( n^1 \). Therefore, all overall tests of the alternative are conducted with sample sizes, \( n_j \geq n^1 \). If the overall test of the alternative is rejected, \( T_A < x_0 \) then accrual to the entire study is terminated. Note that typically one would set the overall alternative to be less than the subgroup alternative, \( \theta_A < \theta_j \). For instance, although a subgroup alternative value for response probability could be \( \theta_j = 0.3 \) (for a null response probability \( \theta_0 = 0.1 \)), a smaller treatment activity may be of interest for the study as a whole; hence, in this case, an alternative of \( \theta_A = 0.2 \) may be appropriate.

After completing accrual, tests of the null hypothesis

\[
T_k^0 = \Pr(X \geq r_k | n_k, \theta_0) \quad \text{for each stratum and} \quad T^0 = \Pr(X \geq r | n, \theta_0)
\]

for the combined group are conducted. The subgroup null hypothesis would be rejected if \( T_k^0 < x_{0k} \) or the overall patient group rejected if \( T^0 < x_0 \); we think it will usually be sufficient to keep the subgroup and overall type I specification equal, \( x_{0k} = x_0 \). One could set \( x_0 = 0.05 \) leading to a larger overall experiment-wise type I error rate or calibrate \( x_0 \) so that

\[
\Pr(T_k^0 < x_0) \quad \text{or} \quad (T_k^0 < x_0) \quad \text{or} \quad (T^0 < x_0 | \theta_0) = 0.05.
\]

An important aspect of the design is the selection of sample sizes per stratum \( n_k^1 \) and \( n_k \), corresponding to a possible first test of the alternative hypothesis and maximum stratum sample size for each stratum \( s \), and similar sample sizes for the entire study, \( n \). We propose using a standard two-stage design to pick \( n_k^1 \) and \( n_k \), such as given in refs. (2, 3). A similar strategy can be used for the overall sample sizes based on \( \theta_0 \) and \( \theta_A \).

We note that in some cases, the assumption of a global null hypothesis may not be reasonable. One may know that baseline tumor response or disease control rates vary between histologic subgroups, so it would be reasonable to specify a null probability, \( \theta_{0k} \), for stratum \( k \). In that case, stratum-specific tests would be based on tests of the hypothesis \( T_k^0 = \Pr(X \geq r_k | n_k, \theta_k) \) for each stratum and an overall test of \( T^0 = \Pr(X \geq r | n, \theta_0) \) where the overall null is an average of stratum-specific values weighted by the fraction of cases in each stratum, \( \theta_k = \sum_k \theta_{0k} \). At the time of analysis, using a subgroup weighted null would have similarities to adjustments proposed in ref. (4). The overall testing strategy would remain unchanged.

A property of the proposed subgroup and combination strategy is that a single stratum is stopped only for futility related to that stratum. Negative results in one stratum lead to stopping the accrual of patients to other strata only if the combined results cause rejection of the alternative hypothesis. Alternatively, one could borrow information across strata in a more complex fashion to make subgroup decisions. A Bayesian approach using a hierarchical model for borrowing between strata was implemented in ref. (5). For each stratum, that method evaluates the posterior probability of the response parameter being less than the alternative hypothesis, and stops accrual to that stratum if that probability becomes very small, 0.005.

**Targeting a subgroup.** A variation on subgroup testing versus combined strategy testing can be used if there is sufficient biological information that treatment will be most active within one or more prespecified subtypes. For instance, certain subtypes of disease may be known to have the greatest gene expression of the target. An alternative test can be based on the strata defined to be the corresponding target cohort. Consider the target test of the alternative hypothesis

\[
T_A = \Pr(X \leq r_j | n_j, \theta_A)
\]

where \( r_j = \sum_k r_k \) and \( n_j = \sum_k n_k \) are the sums of strata indicating the target or most promising biological group. The parameter \( \theta_A \) is the alternative hypothesis appropriate for the target group. Once a sufficient number of cases are entered into the target group, if the target test is rejected, then the entire,

```
study is closed. Therefore, targeted futility testing is different from other subgroup tests in that lack of efficacy for the targeted subgroup implies overall futility, not just subgroup futility. Such a design can lead to smaller expected sample sizes if there is no improved activity in the target group (as seen in the sarcoma example). In addition, there is a potential for improved power when there is modest activity in the target group and none or/but little in the other strata.

### A Simulation Study

A simulation study was conducted to evaluate the operating characteristics of the proposed phase II strategy. Each hypothetical phase II multi-subtype clinical trial was generated with subtype frequencies of \( v_1 = 0.5, v_2 = 0.3, \) and \( v_3 = 0.2. \) The null response probability was \( \theta_0 = 0.1 \) and the subgroup alternative response probability was \( \theta_A = 0.3. \) The overall alternative hypothesis response probability was set at \( \theta_A = 0.2. \) For each stratum, the first stage sample size was \( n^1 = 20 \) and the minimum possible stratum size was \( n_s = 35. \) For the combined study, the first stage sample size was \( n^1 = 40 \) and the maximum sample size was \( n = 105. \) Testing of the alternative was done with \( x_A = 0.02, \) and testing of the null was set at \( x_0 = 0.05. \) Futility analyses were conducted after every five patients were entered once each stratum and the overall study satisfied the minimum sample size requirements given above. We generated 5,000 hypothetical trials to evaluate type I error, power, and expected sample sizes for strata (subtypes) and for the overall study. Four scenarios were considered:

- **Case A. Null case:** \( \theta_1 = \theta_2 = \theta_3 = 0.1. \)
- **Case B. Improved activity in frequent subtype:** \( \theta_1 = 0.3, \theta_2 = \theta_3 = 0.1. \)
- **Case C. Improved activity in infrequent subtype:** \( \theta_1 = \theta_2 = 0.1, \theta_3 = 0.3. \)
- **Case D. Limited activity in all subtypes:** \( \theta_1 = \theta_2 = \theta_3 = 0.2. \)

In the tables below, we give power and expected sample sizes. The following columns describe design properties for the given subgroup frequency (Accrual) and clinical response probability (Response): \( N(\text{Simp}), \) expected sample size only using subgroup testing; \( N(\text{Comb}), \) sample size using combined futility testing; \( \text{Power}(\text{Comb}), \) power for subgroup efficacy test with combined futility testing; \( \text{Cond Prob}, \) an estimate of the conditional probability of rejecting the null hypothesis for the specific stratum, given either the subgroup or combined efficacy tests have been rejected. In addition, we present estimates in which the stratum response estimate is the smallest (or largest) estimate among all strata if the given stratum-specific or overall test is rejected. Target accrual and experiment-wise power across all subgroups in the phase II study are recorded in the title captions.

#### Results

**Null case.** Substantial reductions in the expected sample size were obtained using the combination strategy as shown in Table 3A. The largest reduction was obtained for the lower frequency (more slowly accruing subtypes). For instance, in the strata with 30% and 20% subtype frequency, the relative sample sizes for the strata for the combination strategy compared with only using subgroup testing were 0.79 and 0.69, respectively. As seen in the sarcoma example, the type I error for each stratum is \( x_0 < 0.05 \) due to the additional futility testing. As expected, the experiment-wise error is \( >0.05; \) however, the combination analysis does not seem to substantially increase the error on a

### Table 3. Simple subgroup and combined designs

<table>
<thead>
<tr>
<th>Accrual</th>
<th>Rate</th>
<th>( N(\text{Simp}) )</th>
<th>Power(\text{Simp})</th>
<th>( N(\text{Comb}) )</th>
<th>Power(\text{Comb})</th>
<th>Cond Prob</th>
<th>Min Prob</th>
<th>Max Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Null case</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>0.10</td>
<td>26.45</td>
<td>0.02</td>
<td>24.49</td>
<td>0.02</td>
<td>0.23</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>0.30</td>
<td>0.10</td>
<td>26.70</td>
<td>0.02</td>
<td>21.22</td>
<td>0.02</td>
<td>0.19</td>
<td>0.34</td>
<td>0.40</td>
</tr>
<tr>
<td>0.20</td>
<td>0.10</td>
<td>27.08</td>
<td>0.02</td>
<td>18.80</td>
<td>0.02</td>
<td>0.17</td>
<td>0.38</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Experiment-wise Power(\text{Simp}) = 0.0538, Experiment-wise Power(\text{Comb}) = 0.0636

| (B) Efficacy in frequent histology |
| 0.50 | 0.30 | 34.78 | 0.87 | 34.32 | 0.86 | 0.98 | 0.00 | 0.99 |
| 0.30 | 0.10 | 26.48 | 0.02 | 26.05 | 0.02 | 0.03 | 0.52 | 0.01 |
| 0.20 | 0.10 | 26.82 | 0.02 | 26.32 | 0.02 | 0.03 | 0.53 | 0.01 |

Experiment-wise Power(\text{Simp}) = 0.8736, Experiment-wise Power(\text{Comb}) = 0.8718

| (C) Efficacy in infrequent histology |
| 0.50 | 0.10 | 26.50 | 0.02 | 25.93 | 0.02 | 0.03 | 0.53 | 0.01 |
| 0.30 | 0.10 | 26.49 | 0.02 | 25.35 | 0.02 | 0.03 | 0.53 | 0.01 |
| 0.20 | 0.30 | 34.80 | 0.87 | 31.22 | 0.77 | 0.98 | 0.00 | 0.99 |

Experiment-wise Power(\text{Simp}) = 0.8714, Experiment-wise Power(\text{Comb}) = 0.7862

| (D) Limited efficacy in all histologies |
| 0.50 | 0.20 | 32.92 | 0.40 | 32.74 | 0.40 | 0.47 | 0.37 | 0.38 |
| 0.30 | 0.20 | 33.01 | 0.40 | 32.61 | 0.40 | 0.47 | 0.38 | 0.39 |
| 0.20 | 0.20 | 33.07 | 0.39 | 32.48 | 0.40 | 0.47 | 0.36 | 0.38 |

Experiment-wise Power(\text{Simp}) = 0.7804, Experiment-wise Power(\text{Comb}) = 0.8606

NOTE: Min Stratum = 20, Max Stratum = 35, Min Combined = 40, Max Combined = 105.
subtype-specific basis; this is likely, in part, due to additional early stopping corresponding to the overall group testing.

**Improved activity in frequent subtype.** In this case, both the simple and the combined analyses yield good power to detect a difference (Table 3B). The average sample sizes and power for both techniques are similar. The conditional probability of rejecting the stratum-specific test for the effective subgroup, given rejection for the combined strategy is 98%. In addition, the estimates obtained from each stratum match the underlying generating response rates in 99% of simulated phase II studies leading to the largest response rates in the active stratum, and for the other two strata, estimated response rates are lowest (or tied for lowest) with 53% of cases. Figure 1 shows response estimates for the three strata for this case, and the case in which there is limited activity across strata (Table 3D).

**Improved activity in infrequent subtype.** There is a more significant reduction in experiment-wise power for the combined strategy (79%) compared with the subgroup strategy in this case (87%). This is due to the increased number of simulated studies in which accrual is stopped early due to a test of futility using the combined analysis (Table 3C). Similarly, the overall sample size of the subtype with the best outcome is smaller, again reflecting the greater chance of stopping accrual to the entire study early due to futility tests on the combined data.

**Limited activity in all subtypes.** There is a substantial increase in power for the combined strategy in this case. The experiment-wise power for at least one of the subgroup tests is 78% compared with 86% for the combined strategy (Table 3D). In addition, the simple stratum-specific testing leads to a greater chance of falsely concluding there is no activity in individual subgroups with POWER(Simp) approximately 0.39–0.40. Note that although there was some modest inflation of type I error shown in the null case (Table 3A), the results between the individual stratum tests and the combined strategy were similar. Therefore, comparisons of power between the two methods seems appropriate.

Other combinations of accrual frequencies and response probabilities yielded similar trends in sample size and power. In the simulations, the inflation of the type I error was noted due to the multiple strata testing (not increased substantively to the addition of combination testing). If desired, testing could also easily be done with the experiment-wise error rate calibrated to 0.05. Because efficacy testing is only conducted at the end of the accrual, the assessing effect of modulating the error rates for null hypothesis does not require repeating the simulations.

**Discussion**

For testing agents in complex diseases with multiple biological subtypes, practical choices for a phase II design typically include (a) combining biological subtypes and conducting a single test or (b) designing one or more separate phase II studies for the most promising or most frequent subtypes. The former has the drawback of not acknowledging...
differences in activity in disease subtypes, which are likely to occur, whereas the latter strategy may be logistically difficult if diseases are rare and the relative frequencies in subgroups are not as expected.

The straightforward strategy presented in this article is an obvious compromise; it does both the individual subgroup tests and the overall combined tests all the while acknowledging the multiple testing properties of the design. The combination strategy yields smaller sample sizes when the drug is inactive across all strata, and more power in cases when there is some activity across all strata compared with conducting individual phase II studies in the subgroups. In addition, by retaining the stratum-specific tests, the design allows active subgroups to be identified. We believe that our results support the general proposal of appropriate "borrowing" of information in the phase II setting in which there are multiple biological or histologic subgroups.

To facilitate the study of this general proposal, we have made some specific choices. For instance, we have suggested futility testing for every 5 to 10 patients accrued to the entire study but without stopping accrual at those times unless full accrual to a stratum has been achieved. Clearly, such a strategy needs rapid response assessment and forwarding of data to the data operations center for statistical analysis. In practice, one would likely analyze when each additional 5 to 10 patients become evaluable for response (not just accrued) as the data arrive at the statistical center. Additional problems exist if clinical response is defined (as "best response") so it can be achieved over several to many cycles, which would mean longer periods of time of accruing patients without having information on efficacy. If accrual was quite rapid, this would eliminate much of the potential reduction in sample size under the null hypothesis of doing more frequent testing. Response at a fixed time (2-4 months or the disease control at some fixed time) may be better suited to our proposed designs. Yet, we note that repeatedly stopping accrual is also problematic in the multi-institutional setting, when each stoppage in accrual is often associated with significantly decreased enthusiasm of the investigators for placing patients on a study.

Alternative designs with combined stratum-specific and overall analyses are possible. For instance, one could consider futility testing only at the time of a single interim analysis for each stratum and for the combined group (e.g., stratum size \(n_i\) and total \(n\)). That would give at most \(K + 1\) interim looks at the data where \(K\) is the number of strata. The combined futility testing would continue to lead to smaller sample sizes compared with only conducting stratum-specific testing, if there were no overall effect of the treatment. With a smaller number of analysis times, potentially halting accrual at each interim analysis could also be considered. This strategy would most closely parallel what is often currently done in two stage phase II trials in oncology (3).

It has been noted that there is inflation of type I error due to the subgroup and overall testing. The major component in type I error inflation is due to acknowledging multiple strata (not the addition of the combined testing); therefore, this would still be an issue in running several parallel trials if one views them as a single experiment. Although it is straightforward to modestly reduce the type I error for each to the stratum-specific and overall tests, we think it is best to keep the individual designs as familiar as possible and just acknowledge any increased error due to multiple testing. We believe this is also true for the design choices of minimal sample sizes, \(n_i\) for each of the strata and combined subgroups \(n\). Although one could obtain better properties (such as expected sample sizes) by modulating these sample sizes, we think there is some advantage in communication to keeping the subgroup rules as familiar as possible. However, the exploration of multiple designs can be useful; therefore, we have written an R-function that allows the study of the effect of differing design parameters.

Extensions of this phase II strategy to time-to-event data can be implemented provided it is feasible to accrue the larger number of patients needed for survival studies until the time-to-event is short enough to make the interim analysis seem appropriate. Large sample results, for both subgroup and overlapping overall test statistics, can be developed paralleling work in ref. (6), or properties can be evaluated in smaller sample settings directly via simulation. Furthermore, we focused on hypothesis testing rather than estimation in each subgroup or overall groups of patients. With any adaptive sequential procedure, one could view the estimators as subject to some amount of selection bias; however, the magnitude of the bias can also be estimated in simulations as given above.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

References
Clinical Cancer Research

Multiple Histology Phase II Trials
Michael LeBlanc, Cathryn Rankin and John Crowley


Updated version
Access the most recent version of this article at:
http://clincancerres.aacrjournals.org/content/15/13/4256

Cited articles
This article cites 5 articles, 1 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/15/13/4256.full.html#ref-list-1

Citing articles
This article has been cited by 2 HighWire-hosted articles. Access the articles at:
/content/15/13/4256.full.html#related-urls

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.