Picking the Winners in a Sea of Plenty

Howard I. Scher and Glenn Heller

Genitourinary Oncology Service, Division of Solid Tumor Oncology, Department of Medicine [H. I. S.] and Department of Epidemiology and Biostatistics [G. H.], Memorial Sloan-Kettering Cancer Center, New York, New York 10021

ABSTRACT

Purpose: Selecting an experimental arm for a Phase III trial is based on the results of Phase II investigations. Historical results show that this paradigm leads to the failure of many experimental therapies in the Phase III setting. This is the result of failures in the Phase II design that include differences in the patient populations and basing sample size determinations on levels of benefit derived from surrogate end points that do not accurately reflect the end point of interest in the Phase III study. An additional factor is how to ensure that the experimental therapy chosen was the best available at the time.

Experimental Design: We consider castrate metastatic prostate cancer, for which multiple regimens appear to have similar activity at this time. To assess superiority, we use a randomized Phase II/III design developed by Schaid et al. (D. J. Schaid et al., Biometrika, 77: 507–513, 1990) that allows multiple treatments to be tested at the same time and bases the determination to proceed from the Phase II study on the same clinical end point evaluated in the same population as the Phase III trial. A concurrent control group is also treated.

Results: We demonstrate the integrative Phase II/III clinical trial design to evaluate two, three, or four experimental treatments with a survival-based end point in the same patient population. It includes a concurrent control in both the Phase II and Phase III portions of the study. The sample sizes in the Phase II component of the trial are comparable with those found in conventional single-arm Phase II trials.

Conclusions: The proposed design is valuable in situations where multiple regimens are available that appear worthy of evaluation in the Phase III setting, and where there is no adequate short-term surrogate end point for survival. The design is also useful in the evaluation of cytostatic agents where traditional response parameters may not identify potentially active drugs or, as is the case in advanced prostate cancer, in the evaluation of therapies that have a direct effect on prostate-specific antigen with an uncertain effect on survival.

INTRODUCTION

Intensive research on the mechanisms of cancer development and progression is now being translated to clinical practice. Treatments are no longer restricted to conventional chemotherapy because we now have the ability to target specific biological events that contribute to malignant cell proliferation, survival, and death, based on the molecular phenotype of an individual patient’s cancer. Concurrently, preclinical and clinical studies are showing that approaches directed at the processes of angiogenesis, invasion, metastases, and other host-tumor interactions may provide additional therapeutic benefit. Developing these approaches is more complex than developing the traditional cytotoxic drugs because the optimal dose may not be the most toxic, the traditional parameters of tumor regression that are used to assess response do not apply, and reliable surrogates of true clinical benefit may simply not be available.

It has reached the point that for many cancers, there are too many options to explore, and in some situations, the recruitment of patients cannot keep pace with the myriad of treatments under study. The issue then becomes how to prioritize one approach versus another, to discard those that prove inactive, and to identify treatments that are beneficial, for whom, and to what degree as rapidly as possible. This is currently the case for the management of prostate cancer patients who have progressed after castration. Several approaches appear promising, but only a small number can or will be tested on a large scale.

To change treatment standards requires the demonstration in a randomized prospective Phase III trial that a new therapy is superior to an established standard or a placebo. These trials enroll large numbers of patients and require considerable resources because most advances in cancer therapeutics occur slowly and at modest increments. As a result, few institutions can go it alone to address a trial question. Given the limitations of resources including the number of patients eligible for trials and the complexities of monitoring, there are rarely more than two or at most three randomized Phase III studies running for a given disease state at any given time. This limits the number of approaches that can be tested. More disturbing is that few of the Phase III trials that are conducted show benefit after demonstrating clinical activity in Phase II. This suggests deficiencies in the methodology used to select which treatments to evaluate and determine which to move forward.

The primary objective of a Phase II study is to determine whether the activity level of an experimental regimen warrants further investigation. Eligibility criteria are created so that the trial is targeted to a patient population with defined characteristics. Once a patient population is identified, a minimum level of clinical activity that is worth pursuing is specified, based on what agents or approaches may be available for the same population. The Phase II trial is then performed, and if the activity
rate observed in the trial does not exceed the minimum acceptable level, further study of the experimental treatment is discontinued. If it does exceed the minimum, a Phase III trial is then considered. Typically, multiple Phase II trials are ongoing at any point in time targeting a specific cohort of patients. Although the eligibility criteria may be similar, comparability cannot be assured in the absence of randomization. When completed and analyzed, the results are placed in the context of the other treatments that have been evaluated with the objective of determining which is most likely to succeed in the Phase III setting. An additional level of uncertainty when evaluating trials in advanced prostate cancer is introduced when trying to assess the superiority of one regimen versus another using PSA-based end points (1). Often, the Phase III trial design is developed semiautomatically from the outcomes of the Phase II trial in that it does not use the same primary end points.

To address a lack of coherence, we previously discussed the importance of Phase II study designs that have comparable patient populations and clinical end points comparable with those that may be planned for future Phase III trials (2). In this report, we address these issues when multiple treatment approaches are under consideration, using the example of patients with advanced prostate cancer who are progressing after castration. At present, two prospective randomized comparisons of an experimental treatment with mitoxantrone and prednisone, as the standard, are ongoing. The experimental treatments include two schedules of docetaxel in one, and docetaxel plus estramustine in the other. The planned enrollment of 804 in one and 620 in the other was designed to detect a 33% improvement in survival, respectively (3, 4). Approaches under consideration as experimental arms for future Phase III trials include the three-drug combination of estramustine and docetaxel or estramustine and paclitaxel each with carboplatin, a systemic chemotherapy that has shown encouraging response proportions with a favorable safety profile in single-institution Phase II investigations (5). In addition, the issue whether the taxane is best administered on an every three week or weekly schedule likewise remains. Can the regimens be prioritized using the data currently available? Are we limited to the evaluation of one or at most two experimental programs in a single three-arm study, or two separate two-arm randomized studies? Are there alternative designs that can be considered in the absence of survival data?

We use the randomized Phase II/III clinical trial design developed by Schaid et al. (6), which employs a two-stage design in which promising treatments in the Phase II component are carried forward into the Phase III portion of the trial. A concurrent control or standard arm is also included. This assures the integrity of the patient population under study and eliminates the uncertainties and bias associated with the reliance on a historical series using different end points to estimate the superiority of the experimental approaches. The goal of the first or screening phase of the study is to determine whether any of the experimental treatments show sufficient activity to warrant continuation into the second or confirmatory stage. At the end of the first phase, treatments that demonstrate sufficient clinical activity accrue additional patients, whereas those that do not no longer accrue patients. An overwhelmingly superior regimen can also be identified. Combining the Phase II and III components and requiring the same outcome measure for both phases lead to a consistency between the results of the individual phases. Furthermore, linkage of the Phase II and III study designs allow the Phase II patient data to be used in the principal Phase III trial analysis and, from an administrative standpoint, removes the time lag between the completion of the Phase II study and the start-up of the Phase III study. Finally, by using survival or some variant such as progression-free survival as the end point in both phases, we circumvent the issue of how to determine the true “clinical significance of a response,” which may be difficult to determine when evaluating using nontoxic approaches.

**MATERIALS AND METHODS**

**The First Stage (Phase II).** Assume there are a set of $k$ experimental treatments and one standard therapy under consideration. The individual treatment arms may include single agents and/or combinations or different doses, schedules, or sequences of the same combination. Our objective in the Phase II part of the trial is to reject any experimental therapies that demonstrate no clinical benefit relative to the standard therapy. Clinical benefit is based on the end point of survival, although time to progression can also be used.

We randomly assign $n_i$ patients to each of the $k + 1$ treatment groups. At the completion of the first stage, a log-rank test statistic is computed for each of the $k$ pairwise survival time comparisons between an experimental regimen and the standard therapy. We express the log-rank statistic for the comparison between the $j^{th}$ experimental treatment and the standard therapy at the completion of the first stage by $S_j (n_j)$. Under the assumption of exponential survival time, the hazard rates, the fundamental components of the log-rank parameter, can then be computed from the more easily understood median survival times $m_j (j = 0, 1, \ldots, k)$, where $m_0$ represents the median survival time for the standard therapy. The log-rank statistic is oriented so that a positive value denotes a larger median survival time for the subjects on the experimental treatment. Similarly, a pairwise log-rank statistic $S_j (n_j)$ less than 0 is an indication that the experimental therapy is performing at a level that is no higher level than the standard therapy and is thus dropped from the Phase III portion of the trial. Note that the log-rank statistic $S_j (n_j)$ for the Phase II component here is not powered to determine the superiority of the $j^{th}$ experimental treatment over the standard regimen; rather, the intention is to provide evidence that this treatment has sufficient promise to move into the confirmatory Phase III study.

Determining which experimental treatment(s) move forward to the Phase III component of the study is based on the outcome of the Phase II component. Operationally, the $k$ pairwise comparisons are evaluated using the log-rank statistics once the $n_i$ patients are accrued to each treatment. If all $k$ log-rank test statistics $S_j (n_j) (j = 1, 2, \ldots, k)$ are less than 0,
then none of the experimental therapies are considered promising, and the study is terminated. Similarly, if a log-rank statistic produces a strongly significant \( P \)-value (\( P^* \)), the study is terminated, and a treatment effect is indicated for the \( j \)th treatment. Otherwise, \( n_2 \) patients are accrued to the standard therapy and the experimental treatments that produce a positive log-rank statistic. Sample sizes \( n_k \) in each treatment group for the Phase II component of the study, and \( n_2 \) additional patients per group moving to the Phase III study) are determined by a numerical optimization program. This algorithm chooses the combination \( (n_1, n_2, P^*) \) to produce the smallest total sample size, on average, for a specified test size and power under the hypothesis that the survival distributions are equal for all \( k + 1 \) groups. Thus, the design minimizes the expected number of patients accrued onto experimental therapies that are no better than the standard therapy. The average total sample size is derived as

\[
(k + 1)n_1\pi_0 + \sum_{j=1}^{k} \{n_2(j + 1) + n_1(k - j)\} \pi_j
\]

where \( \pi_0 \) is the probability of stopping accrual at the first stage, and \( \pi_j \) \( (j = 1, \ldots, k) \) is the probability that accrual will continue for the standard treatment and \( j \) experimental treatments. Additional details regarding the computations leading to this algorithm are found in Schaid et al. (6). A FORTRAN program that provides the samples sizes is available on request to the authors.

The Second Stage (Phase III). This second stage is analogous to the conventional Phase III comparative design. It is expected that only a subset of the experimental therapies from the screening phase will be studied in the second stage. Subjects are randomly assigned to the standard therapy and the subset of \( k' \) experimental regimens accepted into the second stage based on the decision rule proposed above. Definitive pairwise comparisons, based on \( n_1 + n_2 \) patients, are performed in the second stage of the design. At the conclusion of this stage, \( k' \) log-rank test statistics \( S'(n_1 + n_2) \) \( (j' = 1, 2, \ldots, k') \) are computed to determine the superiority of these experimental treatments over the standard therapy. Due to the multiple comparisons undertaken in this design, the size of each test is adjusted using the Bonferroni correction to 0.05/\( k' \). Thus, if we begin the study with four experimental regimens, the size of each pairwise comparison is 0.0125.

### RESULTS

**A Clinical Example.** Suppose four experimental treatments are to be screened versus a single control, and the projected median survival for the control arm is 18 months. We anticipate an accrual rate of 125 patients/year and an additional 18 months of planned follow-up after accrual is completed. For each of the four comparisons, a power of 80% is desired to detect a 33% median survival improvement offered by an experimental treatment. The global type 1 error rate is set at 0.05, i.e., there is a 5% chance of claiming at least one of the four experimental treatments is better than the control when all of the treatments are equivalent. The two-stage design includes a decision rule to stop accrual onto an experimental treatment if it is no better than the control at the first stage or if it is significantly better than the control at the first stage.

In the first stage, we will accrue 104 patients/treatment (for a total of 520 patients). After the accrual of the 104 patients/group, we compute the pairwise log-rank statistics (experimental therapy versus standard therapy). Each experimental treatment for which the pairwise log-rank statistic relative to the control is between 0 and 2.62 (\( P^* = 0.009 \)) continues to accrue to a total of 282 patients. The control arm also continues to accrue. If the log-rank statistic is less than 0, accrual to that arm stops. Accrual can terminate altogether at the first stage if either \( a \) all four log-rank statistics are less than 0 or \( b \) at least one treatment-control comparison produces a \( P \)-value* less than or equal to 0.009. If all treatments continue to the second stage, the maximum accrual will be 1410 patients. All pairwise comparisons at the second stage will use 0.05/4 = 0.0125 as the significance level.

The sample sizes generated in Table 1 are governed by the ratio of median survival times between the \( j \)th experimental treatment and the standard therapy \( m_j/m_0 \). As is observed in conventional Phase III clinical trial design, the sample size is inversely related to this ratio. A reduction in the projected median survival improvement from 50% to 25% results in an approximately 2.5-fold increase in the sample size. For example, with two experimental groups, the sample size increases from 155 to 392. The number of experimental groups in the study has a small incremental affect on the total sample size. Interestingly, sample sizes for the Phase II segment show a marginal decrease as the number of experimental groups under study increase.

To evaluate the sample size requirements in the Phase II

<table>
<thead>
<tr>
<th>No. of experimental groups</th>
<th>Integrative Phase II/III</th>
<th>Conventional Phase II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>172</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>357</td>
<td>120</td>
</tr>
<tr>
<td>3</td>
<td>128</td>
<td>111</td>
</tr>
<tr>
<td>4</td>
<td>245</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>087</td>
<td>079</td>
</tr>
<tr>
<td></td>
<td>145</td>
<td>155</td>
</tr>
</tbody>
</table>

Table 1 Sample size calculations for the integrative Phase II/III and conventional phase II designs

For the integrative design, the first row in each cell represents the number of patients/group at the first stage. The second row is the total number of patients for each of the treatment arms that continues to the Phase III portion of the trial. For the single-arm Phase II tabulations, the median survival times are translated into 2-year survival probabilities (\( \pi \)) to perform the sample size calculations.
component of the study, we convert the median survival time from the exponential distribution into a survival probability and use this parameter in the sample size evaluation of a conventional single-arm Phase II study. Specifically, a median survival time of 1.5 years on the control arm is equated to a 2-year survival probability of 0.40. Similarly, a 33% improvement in this median survival as a result of an experimental therapy translates into a 2-year survival probability of 0.50. Thus, using the null hypothesis \( \pi = 0.40 \) and the alternative hypothesis \( \pi = 0.50 \), a trial with 145 patients has power 0.80 for a one-sided test with significance level of 0.05. This sample size calculation is found in the final column of Table 1. An examination of this final column indicates that the sample sizes computed for the Phase II portion of the proposed integrative Phase II/III design are comparable with the conventional Phase II design sample sizes. In particular, for a 33% improvement in median survival, the Phase II sample sizes range from 104–128 in our proposed design.

DISCUSSION

The fruits of modern science have created a new problem in clinical drug development: how to determine which of a series of “promising” therapies should undergo more intensive Phase II testing and ultimately undergo definitive Phase III evaluations. The problem is complicated further with the addition of cytostatic agents and agents that modify host-tumor interactions to the therapeutic armamentarium. The decision to begin this level of clinical investigation, and in particular a definitive Phase III study, is not taken lightly because the cost to all participants is high. For patients, it can represent a decision that may or may not prolong their life; for physicians, it is ability to provide more effective therapies for their patients, and for sponsors, it can be millions of dollars with no return. Equally costly is time; the longer a treatment is not available, the fewer who benefit. Worse still is the possibility that an effective therapy may never see the light of day.

This study addresses the issue of prioritizing therapies by giving many promising approaches a formal evaluation in two defined stages. It circumvents the inherent difficulties in trying to select the “optimal” regimen for a Phase III trial on the basis of small, independent Phase II investigations that differed grossly or subtlety in patient selection, in the regimen used, and in the assessment of outcome. An overwhelmingly superior regimen can also be identified early (6). Issues of selection bias and imbalance of known and unknown prognostic factors are addressed through randomization. Reliance on an intermediate end point of uncertain significance, such as a posttherapy change in PSA, is likewise eliminated by using a survival-based end point.

The design is applicable in situations in which there are a number of approaches that have completed preliminary Phase II evaluations and shown promise, but resources are few. Such is the current situation in selecting treatment for a patient with advanced prostate cancer who has progressed after castration. It is important because it provides an additional safeguard to ensure that a potentially promising treatment is not discarded prematurely or not evaluated at all. The level of superiority selected in the example presented, 33%, is identical to that being sought in a three-arm randomized Phase III trial of mitoxantrone/prednisone versus two dose schedules of Taxotere in which the accrual of 840 patients is planned (3). At the same time, a second two-arm study is comparing mitoxantrone/prednisone to Taxotere/estramusine with a planned accrual of 620 patients to detect a 25% difference. The question of dose and schedule is important for an elderly population from the standpoint of efficacy and safety. In this case, if the Taxotere arms produce comparable survival results, the less toxic regimen would be preferred. Using the current design, an evaluation of all three experimental arms using two dose schedules of Taxotere alone, and the Taxotere combination in comparison with mitoxantrone/prednisone, would require approximately 1648 patients (412 patients/arm) to assess a 25% improvement in survival or 1100 patients (275 patients/arm) to detect a 33% improvement if both stages accrue fully. The proposed multiple-arm design can also be used to compare different doses and schedules of the components of an active regimen. In most cases, we are looking to show an improvement in outcome, but there are situations, as noted, where a comparable outcome would also be important.

From a patient standpoint, the proposed design increases the chance of receiving a promising experimental therapy in the context of a clinical trial and circumvents the need to obtain definitive data for each of the control regimens. In this way, the number of patients treated on a potentially inferior control regimen is reduced. In addition, sponsors with candidate regimens for the experimental arms in Phase III trials could share the costs of development, whereas independent monitoring insures the integrity of the results. This too has the advantage of allowing more promising approaches to be evaluated definitively.

To allow as many promising approaches as possible to be evaluated, “play-the-winner” strategies have been proposed (7). In one report, patients are randomized to one of four regimens as initial therapy. At fixed time intervals, a response assessment is made. Patients deemed to have benefited from treatment are continued on the same regimen; those who have not benefited are changed to one of the alternatives. After four sequential assessments have been completed, the regimens are ranked on the basis of the number of cycles administered across all patients. However, although this does allow multiple approaches to be studied, the end point of this randomized Phase II design is a short-term measure of tumor regression. It would therefore be difficult to apply for therapies or approaches that would be predicted to prolong life on the basis of static as opposed to cytotoxic effects, which slow tumor growth without causing tumor regressions per se, or that modify tumor growth through effects on the interaction between the tumor and the host. The proposed Phase II/III design avoids this difficulty by using the same survival type outcome measure in the Phase II and III portions of the study. This assures the relevance of the end point to the question at hand. By using the end point of survival or time to progression, we avoid the use of short-term measures of response that may be difficult to assess and also may not be the appropriate intermediate or surrogate measure that predicts for a change in the natural history of a disease. For example, a range of posttherapy changes in PSA have been used as measures of “response” in Phase II trials and associated with a survival...
benefit. However, in one trial of hormonal therapy, although a higher proportion of patients treated on the experimental arm showed a normalization of PSA after treatment, this did not translate into an improvement in survival (8). As such, it is questionable whether a posttherapy change in PSA should be used as a surrogate end point for survival. Finally, through randomization, we also eliminate the need to try and correct for potential prognostic factor imbalances between reported Phase II investigations in an attempt to assess superiority (9).

REFERENCES
Picking the Winners in a Sea of Plenty

Howard I. Scher and Glenn Heller

*Clin Cancer Res* 2002;8:400-404.

Updated version

Access the most recent version of this article at:

http://clincancerres.aacrjournals.org/content/8/2/400

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.