New designs for phase 2 clinical trials
Elihu H. Estey and Peter F. Thall

Conventional phase 2 clinical trials are typically single-arm experiments, with outcomes characterized by one binary “response” variable. Clinical investigators are poorly served by such conventional methodology. We contend that phase 2 trials are inherently comparative, with the results of the comparison determining whether to conduct a subsequent phase 3 trial. When different treatments are studied in separate single-arm trials, actual differences between response rates associated with the treatments, “treatment effects,” are confounded with differences between the trials, “trial effects.” Thus, it is impossible to estimate either effect separately. Consequently, when the results of separate single-arm trials of different treatments are compared, an apparent treatment difference may be due to a trial effect. Conversely, the apparent absence of a treatment effect may be due to an actual treatment effect being cancelled out by a trial effect. Because selection involves comparison, single-arm phase 2 trials thus fail to provide a reliable means for selecting which therapies to investigate in phase 3. Moreover, reducing complex clinical phenomena, including both adverse and desirable events, to a single outcome wastes important information. Consequently, conventional phase 2 designs are inefficient and unreliable. Given the limited number of patients available for phase 2 trials and the increasing number of new therapies that must be evaluated, it is critically important to conduct these trials efficiently. These concerns motivated the development of a general paradigm for randomized selection trials evaluating several therapies based on multiple outcomes. Three illustrative applications of trials using this approach are presented. (Blood. 2003;102:442-448)

Introduction

Recent years have seen the introduction of numerous new therapies for treatment of hematologic malignancies. This trend is likely to continue given the current intense activity aimed at identifying genetic abnormalities associated with specific types of cancer. Nonetheless most new agents have not, at least in the past, provided a substantive advance in therapeutic benefit, and it is not uncommon for a new agent or combination to be less effective, or more toxic, than existing treatments. Furthermore, despite the ever-increasing number of new agents, the number of patients available for clinical trials remains limited. Thus, there is a growing need for statistical methodologies that can reliably and efficiently evaluate the clinical efficacy and toxicity of new agents.

The purpose of this paper is to point out certain problems with current phase 2 trial designs and to propose some practical alternatives. We stress the importance of randomization in phase 2 to avoid biased comparisons because these trials are the basis for selecting which agents to investigate in larger trials. We argue that such selection trials should examine a wide variety of new agents, schedules, and combinations.

The designs that we describe are Bayesian, rather than frequentist (P-value based). Accordingly, we briefly describe the Bayesian paradigm. Each design is presented in the context of a specific clinical trial conducted at MD Anderson Cancer Center (MDA). To facilitate exposition, our initial illustration uses only one binary event (“response” or “no response”) as the basis for treatment selection. However, we will argue that characterizing patient outcome in this manner may oversimplify the complexity and reality of medical practice. Thus, we subsequently describe designs that, unlike those in common use, make decisions based on multiple, simultaneously monitored, outcomes. Finally, we present a practical approach to the problem of monitoring composite outcomes, defined in terms of one or possibly several different event times, that require a period of nontrivial length for their evaluation.

The need for randomization in phase 2 clinical trials

A problem inherent in the use of single-arm phase 2 trials to evaluate new treatments is “treatment-trial” confounding. This arises because latent (unobserved) variables that have a substantive impact on response to treatment usually vary a great deal between trials. Differences between the response rates of separate trials that are due to such latent variables are called “trial effects.” Latent variables may include supportive care, physicians, nurses, institutions, or unknown patient characteristics. Differences between response rates that are due to the treatments are called “treatment effects,” and these are the primary focus of clinical trials. The data from separate single-arm trials of different treatments provide a very unreliable basis for treatment comparison because the treatment effects and trial effects are confounded. That is, only the combined effect of the trials and the treatments can be estimated, and it is impossible to determine how much of this combined effect is due to actual treatment differences. Consequently, when the results of separate single-arm trials of different treatments are
compared, an apparent treatment difference may be due to a trial effect. Conversely, the apparent absence of a treatment effect may be due to an actual treatment effect being cancelled out by a trial effect.

The choice of which treatments to study in phase 3 is based on phase 2 data. Because many more new treatments are available than can be evaluated in phase 3, the most promising treatments must be selected based on phase 2 data. Because selection is inherently comparative, basing such selection on data from single-arm phase 2 trials unavoidably creates treatment-trial confounding.

The following example illustrates that, even if one accounts for observed patient covariates, a large trial effect may remain. A single-arm trial of fludarabine + ara-C + idarubicin + granulocyte colony-stimulating factor (G-CSF) + all-trans retinoic acid (FAIGA) for treatment of acute myeloid leukemia/myelodysplastic syndrome (AML/MDS) was conducted at MDA (1995). Subsequently (1996-1998), the FAIGA combination was studied again, as one arm of a randomized phase 2 trial. Because FAIGA was studied in 2 separate trials, the resulting data provide a basis for estimating the trial effect. We analyzed the data from the 2 trials using a regression model accounting for the trial effect and also for effects of observed patient prognostic covariates, including performance status, type of cytogenetic abnormality, and whether the patient was treated in a laminar airflow room. Figure 1 gives the relative risk (RR) of death with FAIGA in the first trial compared to the second. In general, a RR is the ratio of 2 death rates, in this case the rates associated with the 2 trials, after accounting for the effects of observed covariates. An RR = 1 corresponds to identical death rates. Figure 1 thus is a picture of an actual trial effect. It shows that patients in the first FAIGA trial had a much higher death rate. This example, and many others like it, strongly suggest that, given the widespread practice of conducting sequences of single-arm phase 2 trials of different treatments and then comparing their results, many of the “treatment effects” reported in the medical literature may be nothing more than trial effects. Furthermore, this analysis also shows that trial effects cannot be done away with by performing a statistical regression analysis to adjust for variables that may be unbalanced between the 2 trials. This example provides a strong motivation for randomizing in phase 2 because this is the only method that can effectively eliminate treatment-trial effect confounding. Indeed, it appears logically inconsistent that the need to avoid confounding trial and treatment effects is addressed by randomizing in phase 3, yet it is ignored in the evaluation of phase 2 data that determine whether the phase 3 trial will be conducted in the first place.

Rationale for investigating a large number of new agents

A typical single-arm phase 2 trial enters 14 to 60 patients. These sample sizes severely limit the number of new therapies that can be studied. The agents that are studied in phase 2 are thus those that are found to be particularly promising in preclinical studies. Rationale for investigating a large number of new agents

Selecting treatments: liposomal daunorubicin + topotecan or ara-C +/- thalidomide in AML

We used the randomized selection design5 for a trial in patients with untreated AML and abnormal karyotypes. We were interested in the possibility that complete remission (CR) rates in these patients would be higher if they received topotecan (T) + liposomal daunorubicin (LD), rather than ara-C (A) + LD. We also inquired whether addition of thalidomide (Thal) to either LDT or LDA would improve the CR probabilities. Patients were thus randomized among 4 experimental arms: LDA, LDA + Thal, LDT, and LDT + Thal. The design, and those that follow, use Bayesian statistics, which we have found to be especially useful in clinical trials. A Bayesian model has 2 main components. The first is its parameters, which we denote by $\theta$ for convenience. Parameters are theoretical quantities that, although they are not observed, characterize important aspects of the phenomenon under study. For example, $\theta$ may be median survival time, the probability of achieving CR with
a particular treatment, as in the 4-arm trial or, when comparing 2 different treatments, the difference between or ratio of 2 such values, such as the RR of death with 2 different treatments. In contrast with conventional statistical models, the Bayesian paradigm treats $\theta$ as a random quantity. Thus, a Bayesian model includes a prior probability distribution that characterizes one’s uncertainty or knowledge about $\theta$ before conducting an experiment and observing data. The second component of a Bayesian model is the observed data. The randomness in data is characterized by a likelihood function, which specifies the probability of observing any given data if $\theta$ is the parameter. Bayes theorem is used to combine the likelihood of the observed data with the prior, to obtain the posterior of $\theta$ given the data. Essentially, the posterior is a weighted average between one’s prior belief and the information in the data, and it quantifies how much the observed data have changed one’s uncertainty about $\theta$. The idea is that, by observing data and going from prior to posterior, one may learn about $\theta$, and hence about the phenomenon that it describes.

In sequential data monitoring during a clinical trial, Bayes theorem may be applied repeatedly, so that each successive posterior is used as the prior for the next stage. Figure 2 provides an example of this process of “Bayesian learning,” in a setting where $\theta$ is the probability of achieving CR and patients are treated in sequence. The prior on $\theta$ is given by the dotted line, and it reflects almost no knowledge about $\theta$, beyond the average value 0.20. The dashed line is the posterior after observing 5 CRs in the first 10 patients. The solid line is the posterior after observing 7 CRs in the next 30 patients, and it is based on the combined data consisting of 12 CRs in 40 patients. This illustrates how the posteriors become successively more informative as the data accumulate and how they shift to reflect the overall average behavior of the data.

In the randomized setting, where each treatment arm has its own $\theta$, ethical considerations suggest that the priors on the different $\theta$s should be identical. In the 4-arm trial described, the prior for each $\theta$ was taken to be noninformative to reflect the lack of knowledge about the CR probabilities of the experimental treatments before the trial began (Figure 3A). We next specified a prior for the standard treatment $\theta$. This prior was derived from the historical MDA CR rate in patients with abnormal cytogenetics (291 of 591, 49%). Given this information, the prior for the standard treatment (S) was informative, with a shape (Figure 3A) dictated by the above CR rate. Because there was no standard treatment arm in the trial, the prior and posterior on S were the same. In contrast, as the outcomes (CR or not) with each experimental arm (E) became known, these data were combined with the prior and the updated posterior was compared to the prior for the standard treatment. In particular, we specified a minimum absolute improvement of 0.15 to be provided by each E compared to S. Cohorts of 5 patients were entered into each E. If after evaluating a cohort it appeared highly unlikely that the specified improvement would result for a given E, accrual on that arm was closed. The cut-off probability 0.10 was selected to quantify the phrase “highly unlikely.” This produced the rule that a given arm would be dropped if the (no. CRs)/(no. of patients evaluated) was less than or equal to 1 of 5, 4 of 10, or 7 of 15. Otherwise, each arm would accrue 20 patients, with this number chosen to ensure that 80% of the posterior probability for $\theta$ would be within $\pm 0.12$ of the mean observed CR rate. If accrual to an arm was stopped, all subsequent patients, up to a maximum total of 80, would be randomized among the remaining arms, again with evaluation after each cohort of 5 patients using stopping rules analogous to those described.

Accrual into the LDT and LDT + Thal arms was stopped after the first 6 patients in LDT and the first 5 in LDT + Thal failed to enter CR. Figure 3B depicts the resulting posterior probabilities. The figure shows that not only were LDT and LDT + Thal both unlikely to have the targeted 15% improvement over the historical CR rate, but they also were likely to be inferior to the LDA + Thal arms. The remaining patients were randomized between the LDA and LDA + Thal arms, which had observed CR rates of 18 of 40 and 19 of 40, respectively. The LDT arm had 6 rather than 5 patients, and the total sample size was 91 rather than 80, due to the 2-month time period required to evaluate each patient’s CR, an issue we discuss in “Continuous monitoring of delayed outcomes.” The posterior probabilities that the 2 LDA arms are improvements are also small (Figure 3B). Specifically, the posterior probability that LDA is on average 20% better than S is 0.001, and the posterior probability that it is 10% better is 0.039. The corresponding probabilities for LDA + Thal are 0.0004 and 0.019.
than one E met this criterion, the treatment with the highest CR rate would be selected. Including an active S arm requires only that the prior of S also be updated as new data on S are obtained. Given the preceding discussion about treatment-trial confounding, we would include such an arm were we doing the trial today.

Evaluating the probability of a false negative

The possibility that early closure, such as occurred with the LDT and LDT + Thal arms, will lead to rejection of an effective treatment prompts examination of the randomized selection design’s operating characteristics (OCs). The OCs, which may be obtained by computer simulation, describe under various clinical scenarios, the design’s average behavior, including the expected sample size, number of patients treated, probability of early termination (PET), and selection probabilities for each arm. Before the design is used, the OCs may be examined in light of the design’s parameterization, with these consisting of the targeted improvement (0.15), the per arm sample size (20), and the stopping criterion probability (0.10). In practice, we evaluate the OCs under several design parameterizations and several reasonable clinical scenarios, obtain the OCs in each case, and use these as a basis for choosing design parameters that have desirable OCs. For the 4-arm trial, first consider the OCs for each arm alone (intra-arm OCs; Table 1). These include the sample size, which is random because an arm may terminate early, and the PET. The values in this and similar tables are the means of 4000 to 10 000 simulated trials. In the scenario where the true CR rate is 39%, that is, 10% lower than historical rate and hence very undesirable, the PET is 0.85, and the median sample size is only 10 (Table 1). Thus the design is appropriately protective. Similarly, if the true CR rate equals the historical mean of 49%, the PET is also reasonably high, 0.59. In contrast, if the true CR rate is the desired 64%, the PET is only 0.18, equivalently, there is a 0.82 chance the arm will not be (incorrectly) terminated. The between-arm OCs consist of the probabilities of selecting an arm for future study under various scenarios (Table 2). For example, if the true CR rate in 3 of the arms is the historical mean 49% but equals the desired 64% in the fourth arm, then the probability of (incorrectly) rejecting the fourth arm for future study is 0.37. The correct selection probability (0.63) is the probability that the arm will accrue at least 20 patients and will have the highest observed CR rate. In contrast, if the true CR rate in all 4 arms is the historical 49%, the probability of (correctly) selecting none of the arms for further study is 0.32. Thus the false-positive rate is 0.68, with this rate reflecting the small sample size. The false-positive rate decreases to 0.16, however, if all 4 arms have true CR rate 39%.

If the OCs in Tables 1 and 2 are considered unacceptable, the design parameters may be altered, for example, by lowering the criterion probability from 0.10 to 0.05. However, although this sort of change decreases the false-negative rate, it also increases the false-positive rate, and in general, better values for both rates can be obtained only with larger sample sizes. Our nominally high false-negative rate of 37% must be compared with the effective rate that would be obtained if, based on a supposedly superior preclinical rationale, only one of the 4 arms was selected for a single-arm phase 2 trial in the absence of supporting clinical data such as would arise using the selection design. Assuming that preclinical rationale is an unreliable guide to clinical outcome and that, therefore, each of the 4 arms is equally likely to be successful, the effective false-negative rate is 75%, and even larger if the single-arm trial is run with early stopping rules. Although we acknowledge that preclinical rationale is not a completely unreliable guide, as our example suggests it is, we question whether it is sufficiently reliable to govern selection of new agents for large scale testing in the absence of clinical data.

In general, adaptive decision rules use the accumulated data at one or more interim times during an experiment to make decisions about what to do next. Possible adaptive decisions in a clinical trial include closing a treatment arm, stopping the trial, increasing the overall sample size, or modifying the randomization probabilities. A limitation of the selection design described in “Selecting treatments” is that it does not adaptively account for imbalances in prognostic covariates between patients on the different arms. Such an imbalance could occur as a result of the small sample sizes and could increase the nominal false-negative and false-positive rates described in “Selecting treatments.” This problem can be ameliorated by use of “dynamic allocation” when patients are randomized. This is an adaptive method that repeatedly modifies the randomization probabilities to maintain balance between the treatment arms with respect to patient covariates thought to be related to outcome. In the LDT/LDA trial, this was done on the basis of cytogenetics, age, performance status, and presence of an antecedent hematologic disorder. Finally, stopping after entry of a relatively small number of patients as in the LDT + Thal arms could lead to failure to detect a small, perhaps biologically unique, subset of patients who might respond to a treatment, even though the average patient is highly unlikely to respond. It is unclear, however, whether any trial should proceed with such a goal in mind.

Comparison to a conventional design

It is instructive to compare the randomized selection design as applied in the 4-arm trial to what would have obtained using a more conventional phase 2 design. Suppose, for example, that a Simon “optimal” 2-stage design had been used within each of the 4 arms to distinguish between response probabilities of 49% (the historical rate with S) and 64% (the desired 15% improvement) with type I (false-positive) and type II (false-negative) error rates both 10%. This design requires 37 patients in stage 1 and, if 20 or more responses are observed, an additional 51 patients in stage 2. Thus, because 74 patients would have received LDT or LDT + Thal,

<table>
<thead>
<tr>
<th>Clinical scenario</th>
<th>Arm 1</th>
<th>Arm 2</th>
<th>Arm 3</th>
<th>Arm 4</th>
<th>No arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arm 4 has true CR rate 64%, arms 1-3</td>
<td>0.63</td>
<td>0.10</td>
<td>0.17</td>
<td>0.32</td>
<td>0.04</td>
</tr>
<tr>
<td>All arms have true CR rate 49%</td>
<td>0.09</td>
<td>0.09</td>
<td>0.17</td>
<td>0.17</td>
<td>0.04</td>
</tr>
<tr>
<td>All arms have true CR rate 39%</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>

*Mean historical CR rate is 49%.
†For example with a true CR rate of 49%, in 25% of the 10,000 simulated trials, the sample size was 0 to 10, the median sample size was 15, and in another 25% of the trials the sample size was 20, the maximum.
rather than 11, use of the randomized selection design reduced exposure to these ineffective combinations. Because all 4 arms would have been terminated by the Simon design, a total of 148 patients would have been treated, compared to the 91 treated here. Although the increase in overall sample size would have led to lower nominal false-positive and false-negative rates, the price would have been a much less protective and much longer trial. The contrast between the selection design and more conventional designs in part reflects their different goals. The selection design aims to choose the best treatment, among those not dropped due to lack of efficacy, regardless of the degree of difference between the best and the second best experimental treatment. In contrast, hypothesis test–based (frequentist) designs, such as the Simon design, aim to decide whether the best treatment provides a specified degree of improvement over the others, and the power of such a test is the probability that it correctly detects such an improvement. This type of goal generally requires a larger sample size.

**Monitoring multiple outcomes: the double induction trial**

Like the great majority of clinical trial designs, the 4-arm selection design arrives at conclusions about the value of a therapy based on evaluation of a single outcome. In reality, however, a given trial usually has not one but several goals. The following “double induction” trial provides an example. Untreated patients under age 50 with AML received double induction consisting of a course of idarubicin and “high-dose” ara-C, followed by a second course of this combination 14 days after beginning the first course, regardless of the status of their bone marrow at that time. The second course was to be given only to patients judged “eligible,” that is, recovered from first-course toxicity, by their attending physicians. It was thus possible not only that this approach would result in excess early mortality, but that the number of patients eligible for the second course would be so small that the results of the trial would be of little practical significance. Our design thus formally monitored 3 outcomes: the course 2 eligibility rate and, among eligible patients, both the CR rate and the death rate within 90 days after the start of course 1. The 90-day window was selected because the risks of treatment failure or death within this time frame are high. It was decided that the requisite course 2 eligibility rate was 67% or higher. A 4% increase in 90-day mortality rate (“death”), from the historical 14% to 18%, was considered an acceptable trade-off for a 15% increase in the 90-day CR rate (“success”), from 67% to 82%. The trial was to stop if, after evaluating each cohort of 5 patients, the probability was unacceptably high that (1) the eligibility rate was less than 67%, or (2) the death rate among eligible patients was increased by more than 4%, or (3) the success rate was below the targeted 15% increase over the historical rate. Criteria probabilities of 95%, 90%, and 95% for these events quantified the terms “unacceptably high.” These 3 criteria led to 3 sets of stopping boundaries, one for each rate being monitored. If early termination did not occur, 50 patients would be entered, which would ensure that 90% of the posterior probability for the 90-day success rate would be within ± 0.12 of the mean.

This design’s OCs are summarized in Table 3. For example, under the scenario where the true death rate is the unacceptably high value 33%, the PET is 96% and 50% of the simulated trials stopped after 15 patients were entered. For higher true death rates, the median sample size is lower. In contrast, if the minimum study goals were met, the trial ran to completion in 79% of the simulated trials, with this percent increasing for higher eligibility rate, higher success rate, or lower death rate. In fact, the trial closed after 14 of the first 25 patients were ineligible (Figure 4), although the observed numbers of deaths (2 of 11) and successes (9 of 11) within 90 days among eligible patients were both acceptable.

Multiple outcome designs allow the investigator to explicitly specify a trade-off between “efficacy” and “toxicity,” as was done with the 90-day success death rates in the double induction trial. Different investigators might have different trade-offs because the latter are inherently subjective. For example, in the double induction trial, one might consider a 4% increase in death rate acceptable only given a 20% increase in success rate, rather than the 15% that was used. In the scenario where the true death rate is 33% (row 3 in Table 3), on average, 5 of the median sample size of 15 would die compared to 1 of 15 in the historical situation (4% death rate). If the investigator believes this is unacceptable, the cut-off probability may be lowered, say from 90% to 85%. However, making it easier to stop the trial would also increase the PET if the true success rate is the desired 82%. In our experience, a dialogue between clinical investigators and statisticians is necessary when specifying such trade-offs.

Many types of multiplicities arise in the phase 2 setting. For example, it is naive to assume that the effect of one treatment can be evaluated independently of the effect of the preceding treatment. For example, because a new “targeted therapy” may affect multiple targets, a therapy directed at one target may also “down (up) regulate” a second target, thereby influencing response to a future therapy aimed at the second target. The question of whether young patients with CML receive so imatinib mesylate prior to allogeneic transplantation, or vice versa, is another example of the issue of treatment sequence. Conventional statistical designs, however, regard each therapy as a distinct entity, ignoring the issue of the sequence in which 2 or more therapies are administered. Elsewhere we describe a design that evaluates multicourse treatment strategies, rather than a single treatment. As with other multiple outcome designs, this design encourages the use of a wider range of data in therapeutic decision-making.

**Randomized multiple-outcome designs: the clofarabine versus clofarabine + ara-C versus clofarabine + idarubicin trial**

The need to randomize and the desirability of monitoring multiple outcomes led logically to designs that do both. We applied these ideas to design a trial in relapsed/refractory AML. The aim was to select among clofarabine (a new nucleoside analog), clofarabine + ara-C, and clofarabine + idarubicin for use in a subsequent phase 3 trial to include idarubicin + ara-C (IA) as the standard treatment arm (S). Patients are randomized among these 3 experimental treatments,
with each arm evaluated in terms of the trinary outcome [CR, death, neither CR nor death], in successive cohorts of 4 patients. Accrual to an arm terminates if, compared to the historical rates with S, (1) the induction death rate is highly likely to increase by 5% or more or (2) the CR rate is highly unlikely to increase by at least 15%. Thus, a 5% increase in the probability of death (ie, from the 15% historical rate with IA to 20%) is considered an acceptable increase in probability of CR (from the historical 11% to 26%). The criterion probabilities of 95% and 3% quantify the terms “highly likely” and “highly unlikely.” These probabilities led to 2 separate sets of early stopping rules, one for CR and one for death. If accrual on an arm is not terminated early, then a maximum of 24 patients per arm are entered. If accrual to an arm is stopped, all subsequent patients, up to a total maximum of 72, are randomized among the remaining arms, again with evaluation after each cohort of 4 patients. These parameters led to the within-arm OCs in Table 4 and the selection probabilities in Table 5. Under the scenario where 2 of the 3 arms are truly equivalent to IA but the third arm meets the minimum targeted CR rate (26%), with an acceptable increase in death (5%), the probability of correctly selecting the third arm is 0.78 (Table 5). When all 3 arms are truly equivalent to IA, the probability of correctly selecting no arm is 0.46 (Table 5). Thus, the design is more protective against a false-negative than a false-positive result. Again both error rates could be reduced by increasing the sample size, but this would make it more difficult to evaluate as many treatments in as short a time.

In addition to the scientific advantages detailed, a logistic advantage of the randomized phase 2 trial is that only one trial must be organized, rather than several consecutive trials. One also may include an additional rule, not included in the clofarabine trial, that accelerates the selection process by stopping the entire trial early if, based on the interim data, one of the treatments is greatly superior to all of the others.

Table 4. Operating characteristics of clofarabine versus clofarabine + ara-C versus clofarabine + idarubicin trial: within-arm OCs

<table>
<thead>
<tr>
<th>Clinical scenario</th>
<th>Probability of early termination (percentiles)</th>
<th>Achieved sample size (percentiles)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>25th</td>
</tr>
<tr>
<td>Same as rate with IA (11% CR, 15% death, 74% neither)</td>
<td>0.68</td>
<td>8</td>
</tr>
<tr>
<td>26% CR rate, 20% death rate, 54% have neither</td>
<td>0.17</td>
<td>24</td>
</tr>
<tr>
<td>11% CR rate, 35% death rate, 54% have neither</td>
<td>0.78</td>
<td>8</td>
</tr>
</tbody>
</table>

Continuous monitoring of delayed outcomes

A limitation of the designs discussed here, as with nearly all phase 2 designs, is that it may take a nontrivial amount of time to observe the patient outcomes. If there is a delay in observing the outcomes that are the basis for the adaptive decision rules, then repeated application of such rules may be impractical. Trials with delayed outcomes arise in numerous ways. In many trials, “response” is actually a composite outcome defined in terms of 2 or more event times. For example, in AML, response may be the composite event that the patient is alive and in CR 6 weeks from the start of therapy, with CR defined to include hematologic recovery. This requires that the times to CR and death both be monitored and, unless the accrual rate is very low, the 6-week evaluation period for response may render the use of interim decision rules unrealistic. For example, if a patient is alive and in CR after 5 weeks, then that patient’s outcome cannot yet be scored because he may die during the sixth week of therapy. This also is the case for a patient who is alive after 1 week but has not achieved CR, because she may either die or fail to achieve CR during the next 5 weeks. Clearly, the available information on these 2 patients is very different, because the first patient is much more likely to end up as a responder. More complex examples of this sort of problem, involving 3 or more different events times, arise in other clinical settings. Two common ways to deal with incompletely evaluated patients are to simply ignore them or to suspend accrual until their outcomes have been evaluated. The former approach wastes information, and the latter is usually impractical. Recently, a practical solution to this problem has been provided by Cheung and Thall. They propose an adaptive Bayesian method that monitors all patients’ events continuously throughout the trial. The method, “CMap,” allows response to be defined as a composite event in terms of whatever time-to-event variables are clinically relevant. CMAP uses all available information each time a new patient is accrued to decide whether the trial should continued or stopped due to an unacceptably low response rate. Because CMAP is computationally sophisticated, implementation requires a user interface, which looks like a patient log that asks for clinical information. This software, as well as computer programs that implement the other methods described in this paper, are freely available at biostatistics.mdanderson.org.

Conclusions

The fundamental tenet of this paper is that clinical investigators are poorly served by conventional statistical methodologies for phase 2 clinical trial design. In particular, such designs ignore the issues of treatment-trial confounding, selection of therapies for large-scale trials, and the artificiality introduced by the reduction of complex clinical phenomena to a single outcome. As a consequence, such designs are inefficient, wasting both patient resources and data. These problems motivated the development of a general paradigm.
randomized multiple outcome selection design. Logistical difficulties in implementing such designs arising due to delayed outcomes may be dealt with by monitoring one or more time-to-event outcomes continuously.

Regardless of the readers’ opinion of our specific designs, we hope that we have convinced them that the problems we discuss are acute and that new statistical methodologies may provide more desirable designs. These considerations will hopefully expand the range of dialogue between clinical investigators and statisticians.

Acknowledgments

The authors thank Drs Steven Goodman (Division of Biostatistics, Johns Hopkins Oncology Center, Baltimore, MD) and Peter Mueller (Department of Biostatistics, MD Anderson Cancer Center) for reading an original draft of the paper and providing many valuable suggestions, which we have incorporated. We also thank Angela Culler for expert secretarial assistance.

References