Are There Circumstances in Which Phase 2 Study Results Should Be Practice-Changing?

Marcie R. Tomblyn¹ and J. Douglas Rizzo²

¹University of Minnesota, Minneapolis, MN; ²Medical College of Wisconsin, Milwaukee, WI

New pharmaceuticals, innovative combinations of approved agents, and novel treatment modalities have resulted in a marked increase in the need for clinical trials. Evidence for treatment efficacy is best derived from large phase 3 randomized, controlled clinical trials. However, phase 3 investigations are lengthy and expensive, and consume patient resources. Furthermore, some diseases and treatment indications are rare, and adequate numbers of patients for a definitive phase 3 trial do not exist. Consequently, it is imperative for clinicians to understand phase 2 trial design, since their interpretation is required to apply the findings in clinical practice appropriately. The complexity of phase 2 studies is explored, including unique designs, possible use of randomization, and other key elements necessary for interpretation of phase 2 trials. Specific examples and application of these concepts are discussed in this review.

Phase 3 randomized clinical trials are generally considered essential to evaluate efficacy of a new therapy and to compare outcomes between a novel therapy and the current standard. Selection of the experimental novel therapy arm for randomized studies is generally obtained from phase 2 investigations. The randomization process avoids selection bias and balances unknown confounding factors between treatment groups. Aside from meta-analysis of multiple randomized phase 3 studies, the single randomized controlled trial is considered the optimal design to determine treatment efficacy. Less conclusive evidence can be drawn from observational designs and from phase 2 clinical studies.

The primary goal of phase 2 trials is to identify therapies that warrant further investigation based upon acceptable toxicity and promising efficacy. Additional purposes include assessing feasibility (logistics and costs) and acquiring further pharmacologic information. This phase 2 exploration must be accomplished without exposing too many patients to potentially inferior or highly toxic therapy. In the past 5 years, several analyses have reviewed the quality of published phase 2 studies.¹,⁶ These analyses identified an increasing trend to publish phase 2 studies, especially those examining malignant diseases.⁴,⁵ Michaelis and colleagues found that 668 of 703 phase 2 studies claimed efficacy and more than 75% recommended further investigation.⁵ Despite increasing numbers of publications of phase 2 studies, these studies do not necessarily lead to better phase 3 trials. In a review of phase 2 publications, Zia and colleagues were unable to identify factors within phase 2 studies that were associated with positive results in the investigational arms of phase 3 studies.² Instead, they noted that, on average, the overall response rate (ORR) in the investigational arm was 12.9% lower in the subsequent phase 3 studies (n = 43) compared with the preceding phase 2 studies (n = 49). This indicates a publication bias for “positive” phase 2 clinical trials. Furthermore, it suggests that current designs of phase 2 studies may not adequately discriminate the experimental agents for subsequent phase 3 investigations.

Most phase 2 publications report insufficient detail on the statistical design of the trial. Information is often limited regarding rationale for expected outcomes, identification of comparison groups, and reporting of power and type I error.¹,⁵,⁶ An analysis by Thezenas et al found that in 393 phase 2 studies published in 6 oncology journals in 1995 and in 2000, only 123 (31%) satisfactorily reported the statistical design.⁶ An adequate statistical design was more likely to be reported in multicenter studies, more recent publications, and publication in a high-impact factor journal.

Phase 2 Trial Design

There are multiple designs available for phase 2 investigations, including single stage, multistage, randomized control, and randomized selection designs. Each design is explained below. Within each of these designs, the basic rules are similar—simplicity, rapid accrual, and minimization of the number of patients receiving a potentially inferior or toxic therapy. Phase 2 studies are generally powered to minimize the risk of rejecting a potentially effective treatment (i.e., β ≤ 0.2 and usually 0.1 such that there is less than 20% probability of falsely rejecting a potentially useful therapy). Similarly, they are often designed to accept a higher than normal likelihood of chance observations (as indicated by setting the type I error (α) ≥ 0.05 and frequently 0.1). Phase 2 studies can be either single arm with an external comparator derived from the published literature or historic controls, or may have multiple arms assigned in random fashion. As with any study, early termination parameters (stopping rules) must be clearly defined and incorporated into the statistical design to avoid increasing type I error due to multiple testing. In general, stopping rules are created to ensure safety (thus avoiding exposure to toxicity). However, as described below, a study...
may be stopped early due to dramatic improvement in efficacy compared with the a priori estimate. Early study closure due to efficacy should be discouraged, however, since it almost invariably reduces the strengths of any conclusions regarding safety of an intervention.

**Single-stage phase 2 trial design**

Single-stage designs are the simplest with a defined sample size (n) and no specified interim analysis. All enrolled patients receive identical therapy, and the outcomes are assessed after completion of accrual. This design is appropriate when the goal is to estimate a response rate with a specified precision. The weakness of this design is that more patients may be treated with an inferior or particularly toxic therapy in the absence of formal interim review.7 This design is best for rapidly accruing studies or studies in which there is an expected delay for reaching the endpoint.

**Multi-stage phase 2 trial design**

Two-stage designs with an interim statistical analysis are commonly used and are based on the works of Gehan,8 Simon,9 and Ensign et al.10 These designs estimate efficacy after a small, defined accrual (the first stage). If a minimum level of efficacy is not reached, then no further patients are enrolled. However, if that minimum level is achieved, then accrual continues in the second stage to obtain a larger n to more precisely estimate efficacy.9 A recent adaptation of this design allows for examination of toxicity and efficacy within a two-stage design.11 These two-stage designs permit early termination of a trial for toxicity or lack of efficacy, but do not permit early closure due to marked success.12 In multistep designs (Fleming or the Triangular test), an interval including the minimum and maximum number of responses for each step is determined.13,14 The responses analyzed can be either for toxicity endpoints or efficacy endpoints. If the number of responses falls between the minimum and maximum values, then the study continues to the next stage. Conversely, if responses occur outside the predetermined range, the study is stopped for either success or failure. This approach permits early study closure of the investigational therapy, conserving patients and avoiding excessive exposure.13 Two-stage designs are useful when a primary endpoint can be rapidly evaluated (i.e. response at 3 months) such that an interim analysis could feasibly stop the trial if the treatment lacks efficacy or is particularly toxic. One disadvantage of sequential designs is that, depending upon the rapidity of accrual, enrollment may be suspended while an interim analysis is being conducted. An alternative to avoid temporary suspension involves a Bayesian analysis such that the probability of success is frequently monitored, taking into account the a priori probability and the observed data as accrual continues.7,15,16

**Randomized phase 2 trial design**

As described above, randomization allows for unknown patient confounders to be arbitrarily distributed between the groups. Stratification assigns patients to groups to balance the known prognostic factors. It is also feasible to perform randomization after stratification to improve balance between arms. In phase 2 clinical trials, randomization can be performed to compare a control arm and an experimental arm similar to a phase 3 randomized study.15 This is often done if there is limited historical control data or if other treatment factors, such as dramatic changes in supportive care, reduce the applicability of historical controls.17,18 However, unlike randomized phase 3 trials, these studies are not powered to determine true clinical improvement of the experimental therapy compared to the control arm.19 Instead, they reinforce that the experimental therapy is potentially efficacious and warrants further study in a larger, appropriately powered phase 3 clinical trial.

More commonly, phase 2 randomization is between two or more experimental therapies—a randomized selection design—to determine which therapy, if any, should move forward into the phase 3 trial.15,18,20 In general, this design is most appropriate when there is some level of efficacy that has been established for each of the arms. An adaptive method can be used to establish both efficacy and “pick the winner” using a randomized selection design.20 In this adaptive method, each arm is planned as a single arm Simon two-stage design with the type I and type II errors defined such that the type I error is relatively insensitive to success in the other arms. At completion of the first stage, if a minimum level of efficacy is not met (or excessive toxicity is found), the arm is stopped and subsequent patients are randomized to the remaining arms. While the power to select the “correct” winner may decline if all arms are efficacious, the power of the overall study is maintained to select a winner for further analysis.

**Phase 2/3 trial design**

Phase 2/3 randomized designs are also reasonable to compare an experimental therapy to a standard therapy. In this design, the phase 2 and phase 3 segments should have similar eligibility criteria. An analysis is conducted to ensure efficacy without excess toxicity prior to continued accrual to the phase 3 piece. This allows patient data from the phase 2 portion to be included in the final analysis, which dramatically shortens the time to study completion and decreases the number of required patients.18 However, it does necessitate a pre-study commitment of both time and resources for design and study completion for both the phase 2 and phase 3 portions.

**Key Elements in Reviewing a Phase 2 Study**

Table 1 highlights the important factors required to assess the strength of a phase 2 study. Aside from study design, key elements that should be reported include the inclusion and exclusion criteria used, the endpoints to be assessed, a priori estimates with which the results will be compared, critical statistical elements, and the initial accrual goal plus the number actually enrolled and evaluated.
null hypothesis evaluating toxicity can also be stated. These estimates form the basis for A priori definition of endpoints

Response rate
Time to event analysis

A priori definition of endpoints

Response rate
Time to event analysis

Selection criteria

Inclusion
Disease
Disease status
Exclusion

Statistical considerations

Type I error (α)
Type II error (β)
Number of patients needed
Stopping rules

Other issues to be reported

Number of patients enrolled
Number of patients analyzed
Expected results for the endpoint
Method used to estimate expected results
Comparison group (if used)
Historical vs concurrent

Endpoint definitions in phase 2 designs are complicated. Response rate estimates are typical but are more subjective than survival endpoints. However, the increased use of biologic agents that delay time to progression without objective response (reduced tumor volume) may lead to false dismissal of a potentially efficacious therapy. While the specific endpoint for the study is generally described, the clinical value of the endpoint must be assessed by the reader. For example, in myelodysplasia, clinically meaningful endpoints such as decrease in the number of transfusions may have more importance than a complete cytogenetic response. In addition, when using a time-to-progression endpoint, surveillance bias must be considered. For instance, if the study assessed disease status monthly and the clinical standard is every 3 months, early recognition of progression while on the experimental therapy may actually be due to more frequent assessment compared with historic controls rather than inferiority of the experimental treatment to standard therapy.

Phase 2 studies generally have an implied comparator arm even in the absence of a randomized trial. Often this is based on an expected rate of response from preliminary data or results from historic controls. These estimates need to be clearly defined and stated in the study design before the study is initiated. These estimates form the basis for validating the study endpoint. Determination of the alternate hypothesis for efficacy is based on the change from expected historical outcomes (null hypothesis). A second null hypothesis evaluating toxicity can also be stated. This allows the investigator to evaluate a trade-off between the efficacy and the toxicity of a new therapy.

As previously stated, the goal of a phase 2 trial is to identify the promising interventions that warrant further evaluation in a large, randomized controlled phase 3 study. However, most investigators need assurance that potentially effective therapy is not abandoned (low false-negative probability) and, thus, the type II error (β) must be more stringent than in a phase 3 study. At the same time, the investigators are willing to reject the null hypothesis (high false-positive likelihood) when the therapy is actually no better than the control—a higher type I error (α)—because subsequent phase 3 studies will more stringently test rejection of the null hypothesis. The values of α and β must be clearly stated to assess level of certainty of the results and the power of the study.

Phase 2 studies may be subject to bias in interpretation of analyses, particularly when subset or post-hoc analysis is performed. These should be interpreted with caution by the user, since the general study design is subject to bias in the selection of patients that may affect such subset analyses. Similarly, as in any study it is important that users carefully consider any differences between patients enrolled to the trial and those evaluated for results. Patients may be removed from analyses as “not evaluable” for reasons that are both outcomes and predictors of response, such as failing to complete therapy, which will bias the results that are reported in favor of the treatment, or away from toxicity.

Application of Phase 2 Data in Clinical Practice

Using phase 2 data in clinical practice can be misguided or even dangerous. Phase 2 studies are exploratory. They are not powered to provide a definitive answer regarding improvement in efficacy compared with a standard. In fact, response rates in phase 2 studies have been reported to be nearly 13% higher than in subsequent phase 3 studies. Furthermore, phase 2 studies are more subject to various types of bias such as selection bias (since patients are likely to be highly selected), surveillance bias, and publication bias. These issues must be weighed against the reality, however, that in certain diseases or therapies a definitive phase 3 study is either unlikely to be done, or cannot be done due to limited numbers of patients. In these instances, the clinician is left only with phase 2 data. If the limitations are properly understood and acknowledged, phase 2 data can be meaningfully applied to clinical decision-making.

Hematopoietic cell transplantation (HCT) is used for a wide variety of diseases but in small numbers of patients at multiple institutions. Consequently, much of the data and scientific advances in the field are based on phase 2 data. In 2003, approximately 17,700 patients received HCT in North America (http://www.cibmtr.org/SERVICES/summary_slides.html). This total includes both autologous and allogeneic HCT for a broad range of diseases. Transplantation outcomes vary based on several characteristics such as patient variables, disease variables, conditioning
regimen, and supportive care. Despite the need, phase 3 trials in this field remain difficult to conduct due to limited numbers of patients with specified eligibility criteria. Consequently, phase 2 studies, both single-institution and multicenter, remain the norm, and transplantation physicians must rely on these data to guide clinical care. Fortunately, large HCT observational registries such as the Center for International Blood and Marrow Transplant Research (CIBMTR) and the European Blood and Marrow Transplant Group (EBMT) can provide outcome estimates useful for the proper design of phase 2 HCT trials.

Phase 2 data may also be clinically useful in rare diseases, for which insufficient numbers of patients exist for randomized clinical trials. For example, specific treatments for T-cell chronic lymphocytic leukemia are unlikely to be tested in a randomized controlled phase 3 design. Instead, clinicians must rely on case reports, case series, and, if available, a well-designed phase 2 study. Conversely, it would be inappropriate to accept as a new treatment standard data obtained from phase 2 investigations of novel therapy for the early management of a common malignancy such as B-cell chronic lymphocytic leukemia. Instead, the high incidence and prevalence of this disease makes sufficient patient resources available to conduct adequately powered comparisons of the promising experimental therapies in randomized, controlled phase 3 trials. This does not imply that phase 2 data cannot be used to guide treatment recommendations, but that these data should not be considered definitive and are likely to change should data from phase 3 studies become available.

Summary
Phase 2 clinical trials are increasingly being performed and reported. However, these investigative designs are not an adequate replacement for randomized, controlled, phase 3 studies. However, with well-designed studies that have complete reporting of trial design, patient eligibility, study endpoints, and statistical analyses, data from phase 2 trials may be reliable and, thus, applicable in certain clinical situations where “better” evidence is lacking. The clinician must judge the merit of the study and its relevance to a particular clinical setting. Whenever relevant equipoise exists in clinical decision-making, enrollment of patients on a well-designed phase 2 or, more importantly, a randomized controlled phase 3 trial is preferred.

Correspondence
Marcie Tomblyn, MD, MS, Dept. of Hematology, Oncology & Transplantation, University of Minnesota, 420 Delaware St., SE, MMC 480, Minneapolis MN 55455; phone (612) 624-5109; fax (612) 625-6919; tomb001@umn.edu

References