
The Phase II/III Transition: Toward the Proof of Efficacy in Cancer Clinical Trials

Melissa Fazzari, MS, Glenn Heller, PhD,
and Howard I. Scher, MD

Department of Epidemiology and Biostatistics (M.F., G.H.), and Genitourinary Oncology Service, Division of Solid Tumor Oncology, Department of Medicine (H.I.S.), Memorial Sloan-Kettering Cancer Center, New York, New York

ABSTRACT: Few phase III investigations show a benefit for an experimental treatment when compared to a standard therapy or placebo. This illustrates the need for more reliable estimates of treatment effects from the phase II investigations used to design the more definitive phase III trials. In this manuscript, we examine four aspects of phase II clinical trial designs: (1) selecting endpoints; (2) defining the patient population for evaluation; (3) determining a level of activity that would justify a phase III trial; and (4) estimating sample sizes. In each area, problems with the conventional approaches are discussed and alternatives for the successful transition of phase II results to a phase III setting are suggested. An application of the design for patients with androgen-independent prostate cancer is illustrated. *Control Clin Trials* 2000;21:360–368 © Elsevier Science Inc. 2000

KEY WORDS: *Historical data, patient population, phase II, sample size, surrogate endpoint*

INTRODUCTION AND MOTIVATION

Cancer clinical research has reached a new intensity as the disease has moved from a drug-poor, target-poor arena to a drug-rich, target-rich arena. While the search for new agents continues, it remains important to evaluate continually the efficacy of available treatments. Classically, this evaluation takes the form of a sequence of studies beginning with a phase I dose-finding trial, continuing through phase II efficacy studies, and ending with a definitive randomized phase III trial. Each manuscript is designed to answer the question whether continued clinical development is warranted, based on the results obtained from the current manuscript.

Once the dose level has been established in the phase I trial, the manuscript of treatment efficacy begins. There are generally two types of phase II manuscript designs. The first is a screening trial, which determines whether the treatment exhibits any activity against the cancer. Such trials generally accrue between

Address reprint requests to Glenn Heller, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021; E-mail: heller@biosta.mskcc.org. Received August 9, 1999; accepted February 25, 2000.

15 and 25 patients; a prototype of this design is found in Gehan [1]. If treatment activity is demonstrated, the principal phase II trial is initiated. The primary objective of this phase II component is estimation of antitumor efficacy. Based on the observed level of efficacy, a decision is made to either continue exploration in a more definitive phase III comparative trial, or stop the investigation.

Of the different levels of evidence used to support clinical practice, data obtained from prospective randomized studies are considered the most definitive. The phase III randomized clinical trial is a comparative manuscript, contrasting the efficacy of the experimental treatment to the concurrent control. The randomization enables the investigator to attribute differences in clinically relevant endpoints to the difference in treatments.

Typically, the endpoints in randomized trials are survival-based [2]. Unfortunately, the literature is replete with phase III trials showing no benefit for the promising experimental approach relative to a control population (best available standard therapy or placebo). An example of this is the European Organization for Research and Treatment of Cancer trial of Mitomycin C in advanced prostate cancer where a small phase II manuscript instigated a phase III trial. The phase II trial, based on 31 patients, had a 29% response rate, which was considered sufficient evidence to bring the treatment to the randomized setting. Subsequently, the phase III trial of 171 patients showed no survival difference between the standard treatment and Mitomycin C arms [3]. Considering the difficulties in enrolling patients in such intensive trials, care must be taken so phase III trials are not initiated until there is a level of confidence that the results will be similar to phase II results.

In this manuscript we suggest modifications to the phase II trial design (Table 1). The proposed modifications of the phase II manuscript are: ensuring comparability between the phase II and phase III trials with respect to the patient population enrolled; accruing of a larger patient cohort than is typically accrued onto a phase II manuscript, possibly involving multiple institutions; estimating a treatment effect based on a survival endpoint; and demonstrating sufficient clinical benefit to justify proceeding to a large randomized trial. This design is primarily targeted toward a high-risk patient population where a survival endpoint in the phase II setting is reasonable.

ISSUES IN THE INTERPRETATION OF PHASE II TRIALS

Patient Selection

Generally, phase II trials are not conducted with a phase III design in mind. In many cases, the patient population enrolled in a phase III trial is markedly different from the patient population enrolled in the phase II studies. This may reflect differences in the patient population treated at specialized centers where most phase II trials are conducted, relative to the global population of patients with a disease who are treated in the communities where they reside. The practice of enrolling a select group of patients onto the initial screening phase II trial is done to increase the chances of detecting treatment activity, while minimizing the chance of discarding a potentially useful therapy. However, these select patients may not be the most appropriate group on which to base treatment activity for phase III investigations.

Table 1 Phase II Design Issues

	Conventional Phase II	Phase III	Modified Phase II
Patient population	Strict patient eligibility criteria intended to provide purest interpretation of treatment activity	Relaxation of patient eligibility criteria in order to attain patient accrual	Similar to phase III study
Number of subjects	~25	~100–300	~75
Endpoint	Short-term response to treatment	Survival time	k-Year survival probability
Defining the effectiveness of the therapy	Relative to outcome of historical data; summary statistic treated as a parameter	Relative to concurrent control	Relative to outcome of historical data; summary statistic treated as a random variable

Our strategy is to broaden the pool of patients in the phase II trial so they more closely resemble the phase III population. To reduce the differences in the phase II and III patient populations, we propose two initiatives. First, reduce the number of eligibility restrictions for entry onto the manuscript, commensurate with the reduction typically employed in phase III trials. The relaxation of the eligibility criteria in the phase III setting is often borne out of the need to accrue a large number of patients in a reasonable time frame. However, it has the ancillary effect of providing a more realistic picture of the experimental treatment activity in the community. Second, increase the number of centers participating in the phase II manuscript, and attenuate the referral bias frequently observed in single institution studies. In the final analysis, enrolling a more heterogeneous mix of patients may not produce the activity levels seen in the best prognosis patients, but will provide a more realistic assessment of the true treatment effect.

Sample Sizes

The desire to execute large-scale phase III comparative trials is resolute in the clinical trial community. However, there is significantly less resolve to perform large phase II clinical trials. The rationale for smaller phase II studies traditionally revolves around cost, time, and effort. The trade-off for these savings is an imprecise estimate of the parameter of interest in the population. This imprecision is frequently referred to as a loss of information and may result in inappropriate, premature, and inconclusive phase III trials, or the aborted development of a clinically useful therapy.

We equate the confidence interval width with information. Larger clinical trials reduce the interval width and hence increase the information on the population parameter of interest. However, this information increases slowly as the size of the phase II manuscript increases. For example, a fourfold increase in the sample size results in a reduction in the interval width by one half. This suggests that small increments over traditional phase II sample sizes are not sufficient to improve the precision of the estimate.

If the observed parameter estimate from the new trial is close to the expectation, whether or not we reject the null hypothesis depends on the number of patients used in estimation. Too few patients will result in a negative phase II trial even if moderate activity is present. As a result, phase II trials are often underpowered to detect a treatment that only moderately surpasses our expectations. By tightening the interval of plausible population response rates, we obtain a clearer picture of treatment efficacy in the patient population and increase the probability that a good (but not a "home run") treatment will be detected. To do this, larger, possibly multicenter or cooperative group, phase II clinical trials are required.

Surrogate Endpoints

Phase II designs, due to time and sample size constraints, are usually constructed to screen for treatments that produce a good response in a surrogate endpoint. The quality of a surrogate endpoint for a given treatment is determined by its level of mediation between the biological mechanisms that affect

survival and survival time [4]. The strength of the surrogacy increases as the number (and importance) of the biological mechanisms affecting survival also affect the surrogate endpoint. The use of weak surrogates, in most cases, will serve only to confuse the interpretation of phase II trials in the context of phase III exploration. However, the use of these endpoints, especially when more definitive endpoints are not accessible, may provide the only possible assessment of treatment activity. In this case, it is possible that the results seen in the phase II trial are unable to be reproduced in the phase III trial.

In advanced-stage disease populations, it becomes possible to obtain a preliminary estimate of treatment efficacy on survival-based endpoints. We propose that the endpoint of the phase II trial be comparable to the endpoint that will eventually be utilized in the randomized trial. In randomized studies, endpoints typically include time to death, or some variant such as time to progression or time to relapse. In the phase II design for an advanced-stage disease population, we propose a k -year (progression-free) survival endpoint. In this design, the parameter of interest is the probability of k -year survival and each patient is followed for a minimum of k years to enable complete information to be collected. This allows investigators to understand how the treatment will affect the more significant survival type endpoint prior to initiation of a large-scale randomized trial.

Use of Historic Data to Define the Parameter of Interest

Once the parameter of interest and patient population have been defined, the next issue is to set the level of activity that would justify proceeding to a phase III trial. We term this level π_0 , or the maximum expectation under standard therapy. π_0 may be determined using prior data. If a historically treated cohort of patients with similar characteristics is available, an understanding of the efficacy of what is currently available can be made. The estimates can also be used to define what constitutes an inactive treatment, or one that is similar in terms of efficacy to what may already be in use. The set of values corresponding to levels considered insufficient is usually developed as if the prior data contained no variability. For example, let the parameter of interest be the probability of surviving 1 year. If the survival probability from a previous trial is estimated to be 0.40, the conclusion might be that the maximum expectation of the current level of activity should be set at $\pi_0 = 0.40$. In fact, the population parameter value from prior treatment lies somewhere in an interval around 0.40; the size of the interval is dependent on the sample size of the previous trial. This uncertainty should be considered when determining the region of interest for a new trial. Thall and Simon have outlined a method using historical data in planning a new phase II trial [5]. This method takes into account the variability of prior studies. A more conservative method that we propose is to determine the expectation from a 75% upper confidence bound. This will have the effect of shifting π_0 slightly to the right, making it harder to proceed to phase III trials and acknowledging some of the imprecision of the estimate.

Some attributes of this methodology are tied to the amount of information already available in prior studies. If the amount of information is small, the confidence interval around the calculated parameter value will be wide, and π_0 will be high. In this scenario, only an outcome of considerable impact will

Table 2 Maximum Expectation Under Standard Therapy π_0

π_{new}	0.41 (95% Upper Confidence Bound)	0.37 (85% Upper Confidence Bound)	0.35 (75% Upper Confidence Bound)
0.46	615	215	146
0.48	306	140	105
0.50	182	95	75
0.52	122	70	60
0.54	92	58	44

show sufficient activity to evaluate in a phase III trial. In contrast, if the amount of information available from prior studies is large, π_0 is lower, enabling more modest treatment outcomes to be declared active.

AN EXAMPLE

Consider the evaluation of a new approach for androgen-independent prostate cancer in the phase III setting. The primary objective is to determine whether a new approach will confer a survival benefit relative to current expectations of standard therapies (π_0). The endpoint is the proportion of patients who have died by a specified time. If it is determined that the new approach is sufficiently active, a more definitive phase III trial versus the best currently available standard or placebo will be undertaken. If it is not determined to be sufficiently active, development ceases.

To determine the appropriate upper bound for standard treatments (i.e., the maximum expectation π_0) a database of 93 patients treated at Memorial Sloan-Kettering Cancer Center (MSKCC) between 1987 and 1994 was utilized. A full description of these data can be found in Scher et al. [6]. These data represent the current distribution of death times for high-risk patients with androgen independent disease treated at MSKCC. The criteria used to define high risk will also be used as the entry criteria in the phase III trial, thus keeping the population consistent from phase II to phase III manuscript.

The 1-year survival probability for this high-risk population was estimated to be 0.32. Acknowledging the variability inherent in this estimate, a conservative expectation may be obtained by setting π_0 equal to the 75% upper confidence bound for the 1-year survival probability derived from the historical data. This means that our maximum expectation under a standard treatment regimen of 1-year survival probability is equal to 0.35, creating a stricter decision rule (Table 2). Often this variability is not accounted for, resulting in an expectation set at 0.32, the observed value. This methodology allows the practitioner to be even more (or less) conservative in accounting for the variability by employing a larger (or smaller) upper confidence bound to the historical point estimate in determining the insufficient activity region. For example, a 95% upper confidence bound would produce a maximum 1-year survival probability of 0.41 for the insufficient activity region (Table 2). A higher expectation provides more confidence that a positive trial does reflect a superior population parameter value, but at the cost of an increase in sample size.

Table 3 Sample Sizes Necessary for 80% Power

π_{new}	75% Upper Confidence Bound = 0.35 ($n = 90$)	75% Upper Confidence Bound = 0.36 ($n = 60$)	75% Upper Confidence Bound = 0.37 ($n = 30$)
0.46	146	160	215
0.50	75	78	95
0.54	44	48	58

n = number of patients in historic data set.

Table 2 also illustrates the approximate sample sizes required to demonstrate treatment activity based on our historical data set and a maximum 1-year survival probability of 0.35 for the insufficient activity region. For example, if we expect a new treatment to have an estimated 1-year survival probability of 0.50, 75 patients will be required. To compute these sample sizes we assume a true 1-year survival probability π_{true} greater than π_0 , and employ Monte Carlo simulation to determine the sample size such that the asymptotic 95% lower confidence bound excludes π_0 with probability 0.80.

Although the sample sizes calculated in Table 2 are larger than those found in typical phase II trial designs, they are consistent with our view of the importance of providing greater information on the parameter of interest prior to entering the phase III setting. It also allows us to estimate the parameter in a way that we are confident that the true parameter value from the current population lies in the sufficient activity region determined by the standard care population. The more active we anticipate a new therapy to be the smaller the number of patients needed to show that the estimated survival probability exceeds π_0 . As illustrated, if the anticipated observed probability estimate is equal to 0.54, a sample size of 44 patients is needed to derive intervals that are not in the inactive region.

The size of the previous trial from which we obtain our expectation level for the new trial is important. Smaller sample sizes result in larger upper confidence bounds, which has the effect of shifting the active region to the right. This makes it harder to declare the current treatment active. To investigate the degree to which this region is affected by the sample size of our previous data, bootstrapping methods were utilized [7]. This method generates B samples of size 93 from our historical data, and treats the set of 93 patients as if it were the entire theoretical population. In each of the B bootstrap samples, we estimate a 1-year survival probability and then examine the distribution of these 1-year survival estimates. The 75th percentile of this distribution will serve as the demarcation of the active region based on a sample of size n . If we set $n = 93$, the size of our prior data set, then this percentile will be very close to the estimate of the upper bound π_0 . By changing the sample sizes we are taking from our prior data, we will get an idea of how the sample size affects the upper bound of the confidence interval, and ultimately our expectation level.

Table 3 provides the bootstrap calculations of π_0 based on different sample sizes taken from the historical data set. If we used bootstrap samples of size 30 then the resulting calculation of π_0 would be set at 0.37 (shifted to the right from 0.35). The uncertainty of our estimate based on 30 patients (as opposed

to the entire group of 93 patients) makes it more difficult to declare a new treatment efficacious.

Based on these shifts in π_0 due to less precision in our historical estimate, the sample size required to declare a new treatment efficacious increases. For example, if the expectation under standard therapy is based on a previous trial of size 30 and the new treatment is such that π_0 is equal to 0.46, the sample size required to have 80% power to declare the new treatment efficacious would be raised to 215 patients. This is much larger than the 146 patients required when the expectation is based on a previous sample of size 90. The differences in sample size required diminish as the effect of the new therapy increases.

DISCUSSION

Many treatments are evaluated in the phase III setting without a reliable estimate of the survival benefits, resulting in a clinical trial landscape that is filled with promising but failed treatments. The inability to detect this failure before the phase III investigation is particularly harmful because these trials are resource-intensive, requiring hundreds of patients and vast sums of money, and often take many years to complete.

A number of factors in the phase II manuscript design can contribute to the failure of a subsequent phase III trial. These include but are not limited to: the estimation of a treatment effect on a surrogate endpoint in the phase II trial that does not accurately reflect the endpoint of interest in the phase III manuscript; differences in the populations enrolled in the respective studies; an overestimation of the level of benefit of the therapy through misinterpretation of the outcomes observed in prior studies; and a failure to identify clinically beneficial approaches through underpowered trials. Other issues that merit evaluation include stratifying the phase II (and phase III) population into more homogeneous groups, and assessing the risks of the proposed treatment relative to its potential benefits.

The current design is intended as an intermediate trial between a screening phase II manuscript, based on an initial treatment response endpoint and the phase III randomized comparative survival manuscript. If the experimental therapy is inactive relative to expectations, the design requires a greater number of patients than the conventional screening phase II manuscript. However, even when the two phase II studies are combined, they require less than half the number of patients needed typically needed for a phase III manuscript. Implementation of this proposal increases the type II error, which is the probability of removing a therapy from the process that has an incremental benefit to the patient population. The inflation in the type II error is counterbalanced by an increase in the precision of the survival estimate, providing the investigator with more reliable information on the potential benefit of the new therapy. For example, if at the conclusion of the trial the lower confidence bound for the survival parameter lies above the historical control estimate but below π_0 , the investigator may choose to further pursue this therapy in combination with other therapies. This issue highlights the conservative nature of this approach; it requires new experimental therapies to demonstrate clear effectiveness in a clinically relevant endpoint prior to its introduction into a phase III trial. However, if we proceed under the assumption that the great majority of experimental

therapies for advanced-stage disease do not provide sufficiently enhanced activity relative to the current therapies, then the proposed design provides a mechanism to accelerate the testing process by reducing the number of approaches that are tested in the phase III setting. The result is a lower number of patients enrolled in trials to identify those that are indeed active.

This manuscript was supported by NCI Grants CA-05826 and CA-09207, the Tarnapol Foundation, and the PepsiCo Foundation.

REFERENCES

1. Gehan EA. The determination of the number of patients required in a preliminary and follow-up trial of a new chemotherapeutic agent. *J Chronic Dis* 1961;13:346–353.
2. Simon RM. Design and analysis of clinical trials. In: DeVita VT, Hellman S, Rosenberg SA, eds. *Cancer: Principles and Practices of Oncology*. Philadelphia: Lippincott-Raven; 1997.
3. Newling DW, Fossa SD, Tunn UW, et al. Mitomycin C versus Estramustine in the treatment of hormone resistant metastatic prostate cancer: The final analysis of the European Organization for Research and Treatment of Cancer, genitourinary group prospective randomized phase III manuscript. *J Urol* 1993;150:1840–1844.
4. Fleming TR, Demets DL. Surrogate endpoints in clinical trials: Are we being misled? *Ann Intern Med* 1996;125:605–613.
5. Thall P, Simon R. Incorporating historical control data in planning phase II clinical trials. *Stat Med* 1990;9:215–228.
6. Scher HI, Kelly WK, Zhang ZF, et al. Post-therapy serum prostate specific antigen level and survival in patients with androgen-independent prostate cancer. *J Natl Cancer Inst* 1999;91:244–251.
7. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.