

## Testing the Wrong Hypothesis in Phase II Oncology Trials: There Is a Better Alternative

□□ *Commentary on Vickers et al., p. 972*

Mark J. Ratain and Theodore G. Karrison

Although the Code of Federal Regulations (21 CFR 312.21) defines phase II studies as “controlled clinical studies,” the vast majority of phase II oncology trials have been single-arm investigations. Granted, a liberal definition of the word “controlled” would allow the use of historical controls; however, this is not the norm in other therapeutic areas (1).

The custom of single-arm, phase II oncology trials appears to date back to Gehan’s 1961 article (2), which proposed a method to calculate the minimum sample size required to estimate the response rate with a given degree of precision, under the caveat that the study would be terminated early if there was 95% confidence that the response rate was less than some target response rate of interest (usually 15-20%). In this classic design, 14 or 19 patients (corresponding to response rates of 20% or 15%) are initially treated, and if no responses are observed, the drug is considered inactive. On the other hand, if at least one response is observed, additional patients are enrolled and the response rate is estimated. Thus, the Gehan design combines elements of estimation and hypothesis testing.

In the 45 years subsequent to Gehan’s article, there have been a large number of articles about phase II design, most of which have focused on hypothesis testing, rather than estimation (3). It is important to emphasize for the average reader that hypothesis testing is generally organized around the concept of a null and alternative hypothesis. Usually, the alternative hypothesis is what the investigator hopes is true, and the null hypothesis is what the investigator hopes is not true. Furthermore, these two hypotheses are usually constructed such that one or the other must be true. In the simplest hypothesis testing framework, one is trying to compare a standard and investigational therapy in a randomized study. In this context, the alternative hypothesis is that the investigational therapy is different (and ideally better) than the standard therapy, and the null hypothesis is that the two therapies are the same. The data analysis focuses on testing the null hypothesis. If the null hypothesis is rejected (i.e.,  $P < 0.05$ ), then one concludes that the two therapies are different, and by definition, one can accept the alternative hypothesis.

In the Gehan design, the hypothesis being tested during the initial stage of the study is that the drug is active (response rate is *higher* than some minimal response rate of interest). Thus, this is the null hypothesis. If none of 14 (or 19) patients respond, this hypothesis is rejected. Therefore, the alternative hypothesis is that the drug is inactive. Thus, in this design, the investigators (and certainly the patients) would hope that the null hypothesis is true, which is an unusual scenario in clinical trials.

In 1982, Fleming (4) proposed the use of null and alternative hypotheses that do not imply that one or the other must be true, and this framework was also used in Simon’s 1989 design (5). The null hypothesis is that the response rate is less than or equal to some response rate ( $P_0$ ) that “does not warrant further investigation,” and the alternative hypothesis is that the response rate is greater than or equal to some response rate ( $P_A$ ) “which warrants further investigation,” with  $P_A > P_0$ . In this design, rejection of the null hypothesis is desirable and leads to the conclusion that the response rate is greater than  $P_0$  and that the data are *not inconsistent with* the alternative hypothesis. However, this should not lead to acceptance of the alternative hypothesis, and in many scenarios where the null hypothesis is rejected, the observed response rate is less than  $P_A$ . Thus, there are actually three possibilities: (a) the null hypothesis is true (response rate is less than  $P_0$ ); (b) the alternative hypothesis is true (response rate is greater than  $P_A$ ); and (c) neither the null nor the alternative hypothesis is true (response rate is between  $P_0$  and  $P_A$ ). Investigators need to recognize that rejection of the null hypothesis is consistent with either the explicit alternative hypothesis or the implicit third hypothesis (i.e., neither the null nor the alternative hypothesis is true).

This use of null and alternative hypotheses for uncontrolled phase II trials has been a frequent source of confusion for investigators. We generally expect to accept the alternative hypothesis if we can reject the null hypothesis, but this only applies if these are the only two possibilities. To better illustrate the opportunity for confusion, consider three possible implementations of the Simon design. In all three scenarios,  $P_0$  is 0.05 (5% response rate), but  $P_A$  is 0.20 (scenario A), 0.25 (scenario B), or 0.30 (scenario C). Using the National Cancer Institute calculator for the Simon design with  $\alpha$  (type I error) of 0.05 and  $\beta$  (type II error) of 0.10,<sup>1</sup> one can calculate the minimal response rate to reject  $P_0$  for each scenario, yielding 5/41 (12.2%) for A, 4/30 (13.3%) for B, and 3/17 (17.6%) for C. In no case does the minimal response rate to reject  $P_0$  exceed

**Authors’ Affiliation:** Departments of Medicine and Health Studies, Committee on Clinical Pharmacology and Pharmacogenomics and Cancer Research Center, The University of Chicago, Chicago, Illinois  
Received 10/19/06; accepted 10/27/06.

**Requests for reprints:** Mark J. Ratain, Department of Medicine, University of Chicago, MC2115, 5841 South Maryland Street, Chicago, IL 60637. Phone: 773-702-4400; Fax: 1-773-702-3969; E-mail: mratain@medicine.bsd.uchicago.edu.

© 2007 American Association for Cancer Research.  
doi:10.1158/1078-0432.CCR-06-2533

<sup>1</sup> <http://linus.nci.nih.gov/~simonr/otsd.html>

$P_A$ . If it is rejected, one can conclude that the response rate exceeds  $P_0$ , but rejection of  $P_0$  alone does not imply that a phase III trial should be initiated.

In this issue, Vickers et al. (6) appropriately identify that the use of prespecified null and alternative response rates can be problematic, and they focus on the justification for the choice of these rates. Another important problem not addressed in this study is misinterpretation of results of studies that use the framework of null and alternative response rates, as alluded to above. In reviewing recent articles that cite Simon's 1989 article (of 656 citations as of October 6, 2006), one readily finds examples of misinterpretation.

As one example, Utkan et al. (7) concluded that the combination of cisplatin and gemcitabine was active in mesothelioma because they rejected their null hypothesis ( $P_0$  of 0.10), despite the fact that their  $P_A$  was 0.30 and their actual response rate was 23%. In a second example, Philip (ref. 8; in an National Cancer Institute-sponsored study) used the Simon design "to detect a 24-week progression-free rate of at least 20%" for erlotinib in advanced biliary cancer. They implemented this design using  $\alpha$  of 0.09 and  $\beta$  of 0.08, and determined that "promising activity" would be defined as 4 or more of 35 evaluable patients who were progression-free at 24 weeks. (The authors did not explicitly indicate  $P_0$ , but using 0.05 and  $P_A$  of 0.20, the optimum design was recapitulated.) They actually enrolled 42 patients and observed that 7 (17%) were progression-free at 6 months. This would allow the rejection of the null hypothesis and the conclusion that the progression-free survival rate at 24 weeks most likely exceeds 5%, a conclusion that is not of any medical significance. It certainly does not allow the acceptance of the alternative hypothesis, especially because the identified progression-free survival rate was less than  $P_A$ . [Although the authors' conclusion that erlotinib has "modest activity" in this disease is probably valid, this conclusion is more appropriately based on the identification of 3 (7%) patients with partial responses than on the rejection of their implicit null hypothesis.] Thus, as

Vickers et al. point out, where one "sets the bar" for  $P_0$  and  $P_A$  is indeed crucial, if a single-arm trial is the only option. If  $P_0$  is set close to the "response rate" achievable with the current standard of care, then rejection of  $P_0$  might be considered sufficient evidence for mounting a phase III, randomized trial (although it should not be considered as strong evidence that such a trial will be positive). On the other hand, if  $P_0$  is set lower, one should carefully assess what one has actually "proved."

In any case, even if one is careful to document how the null and alternative hypotheses were determined, and even if these are based on a fair degree of historical experience, the well-known problems associated with the use of historical controls will remain. Thus, rather than questioning how we define  $P_0$  and  $P_A$  (6), maybe we should reconsider what we learn from this design. As previously articulated in this journal, the use of uncontrolled phase II trials has been very good for uncovering inactivity but did not have a high positive predictive value for identifying approvable compounds (9).

However, the desired goal of high positive predictive value can be obtained with a randomized controlled trial. Although this raises concerns about the requisite sample size, this problem can be ameliorated by using one-sided tests and a more liberal  $\alpha$  level than typically used in phase III trials. If one takes a common scenario, the addition of a new drug to an established therapy with a 20% response rate, one might hope to achieve a 40% response rate. Using the Simon design, one would require 54 patients (with  $\alpha$  of 0.05 and  $\beta$  of 0.1). If a randomized trial was conducted with a one-sided  $\alpha$  of 0.15 and  $\beta$  of 0.2 (80% power), 74 patients would be required (37 per group). A positive randomized trial (i.e.,  $P < 0.15$ ) is strongly suggestive that the new combination is better.

It should be no surprise that phase III trials fail without predictive phase II studies. We believe that proceeding to phase III with unbiased evidence, and therefore a higher level of confidence that the alternative hypothesis is true, will improve our dismal success rate in oncology phase III trials (10, 11).

## References

1. Temple R. Current definitions of phases of investigation and the role of the FDA in the conduct of clinical trials. *Am Heart J* 2000;139:S133-5.
2. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis* 1961; 13:346-53.
3. Ratain MJ, Mick R, Schilsky RL, Siegler M. Statistical and ethical issues in the design and conduct of phase I and II clinical trials of new anticancer agents. *J Natl Cancer Inst* 1993;85:1637-43.
4. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982;38:143-51.
5. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989;10:1-10.
6. Vickers AJ, Ballen V, Scher HI. Setting the bar in phase II trials: the use of historical data for determining "go/no go" decision for definitive phase III testing. *Clin Cancer Res* 2007;13:972-6.
7. Utkan G, Buyukcelik A, Yalcin B, et al. Divided dose of cisplatin combined with gemcitabine in malignant mesothelioma. *Lung Cancer* 2006;53:367-74.
8. Philip PA, Mahoney MR, Allmer C, et al. Phase II study of erlotinib in patients with advanced biliary cancer. *J Clin Oncol* 2006;24:3069-74.
9. Ratain MJ. Phase II oncology trials: let's be positive. *Clin Cancer Res* 2005;11:5661-2.
10. Michaelis LC, Ratain MJ. Measuring response in a post-RECIST world: from black and white to shades of grey. *Nat Rev Cancer* 2006;6: 409-14.
11. Booth B, Glassman R, Ma P. Oncology's trials. *Nat Rev Drug Discov* 2003;2:609-10.