available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

The unclear zone in phase II clinical trials ☆

Sabrina Allegro [a], *Gregory R. Pond* [a,*], *Sébastien J. Hotte* [b]

[a] *Department of Oncology, McMaster University, Hamilton, Ontario, Canada*
[b] *Department of Oncology, McMaster University, Juravinski Cancer Centre, Hamilton, Ontario, Canada*

ARTICLE INFO

Article history:
Received 17 March 2010
Received in revised form 18 May 2010
Accepted 25 May 2010
Available online 26 June 2010

Keywords:
Phase II clinical trials
Statistical inference

ABSTRACT

Objective: There appears to be considerable confusion about the interpretation of phase II clinical trial conclusions as contrasted with the alternative (HA) and null (H0) hypotheses. This study was conducted to evaluate whether there is congruence with numerical results of phase II trials and their overall verbal conclusions.

Methods: A literature search of 2006 and 2007 phase II clinical trials was conducted. The alternative and null hypotheses were noted as were point estimates with confidence intervals (CIs). These were compared with the final conclusions and concordance and discordance rates were calculated.

Results: A total of 152 eligible analyses were reviewed. The point estimates were below H0 in 42 (27.6%) trials, above HA in 60 (39.4%) trials and between H0 and HA (i.e. the grey zone) in 50 (32.9%) trials. Thirty-three (21.7%) trials reported negative conclusions, 111 (73.0%) reported positive conclusions and 8 (5.3%) were ambiguous. All 60 trials in which the point estimate was greater than HA reported positive conclusions, as did 40/50 (80.0%) of trials with point estimates in the grey zone.

Conclusions: There exist inconsistencies and ambiguities in the conclusions drawn from phase II trials, particularly when results are in the grey zone (greater than H0 but less than HA). This may make the integration of phase II trials in phase III trial development strategies difficult and better understanding of the statistical properties of phase II clinical trials is required.

1. Introduction

The transition to clinical use of a new oncology drug is explicitly reliant upon the results of phase II trials. In a typical phase II trial, a cohort of patients is treated and the outcomes are related to a pre-specified target.[1] If the results meet or exceed the target, the treatment is declared worthy of further study; otherwise, further development is likely to be stopped. If the trial is deemed to have sufficient activity against a disease then the therapy is felt to warrant further testing in a phase III trial. This method is often referred to as the 'go/no-go' decision algorithm.

The design of most phase II oncology trials is based on the Simon–Fleming model.[1] This framework dates back to Fleming's 1982 and Simon's 1989 two-staged approach.[2,3] Using this model, there is one explicit value for the null hypothesis (H0), often defined as the maximum level of efficacy for which no further study of the new treatment would be warranted.[4] Efficacy is frequently measured by the shrinkage of the tumour, i.e. tumour response, to treatment.[5,6] If the true response rate of the new treatment is at H0 or less, development of the treatment in that indication should not proceed. There is also one explicit value for the alternative

☆ Presented in part at the 44th Annual Meeting of the American Society of Clinical Oncology, Chicago, IL, June 2008.

* *Corresponding author:* Address: 60(G) Wing, 1st Floor, Henderson Research Centres, 711 Concession St., Hamilton, ON, Canada L8V 1C3. Tel.: +1 905 527 2299x42616; fax: +1 905 575 2639.
E-mail address: gpond@mcmaster.ca (G.R. Pond).

0959-8049/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2010.05.027

hypothesis (HA), often defined as the minimum level of efficacy which would warrant further study of the new treatment.[7] If the true response rate of the new treatment is HA or greater, further study of the new agent in a phase III trial is warranted. However, the null and alternative hypotheses are separated by a gap (Fig. 1), which is a type of grey zone. If the true efficacy lies within this grey zone, by definition of the hypotheses, it would not be sufficiently active to warrant further study, but active enough that investigators might not be willing to stop the development of the treatment.

Anecdotally, there was an impression by the authors that the verbal conclusions of phase II trials often appear different than what the numerical results show. There are misconceptions about the statistical interpretation of the Simon–Fleming model due to the effect of this so-called 'grey zone'. If this is the case, adopting study conclusions de facto could lead to misrepresentation of the data. Moreover, these misinterpretations may result in erroneous clinical and drug development decisions. Ratain commented on this during his 2007 editorial in response to Vickers' review of phase II trials and the difficulty with the 'go/no-go' theory.[1,4]

Due to our anecdotal impression, we undertook the present study which examined the published phase II clinical trials to explore the effect of results which fall in the grey zone and the overall conclusive statements of these studies. It is hypothesised that there is a considerable incongruence between actual study results and study conclusions.

2. Methods

2.1. Search strategy

A review of phase II clinical trials was conducted. Medline and Web of Science were searched using the keywords 'phase 2 clinical trial' and having 'Fleming or Simon' in the title or abstract, or as a substance or subject heading word. Original reports of phase II oncology clinical trials published in 2006 and 2007 were then manually extracted. Inclusion criteria consisted of papers that documented an original report of a trial in cancer patients in either a single-arm study or a multi-arm study in which a hypothesis for each arm was tested independently. The primary end point had to be a binary measure of an anti-tumour effect and the sample size calculations had to be based on the Simon–Fleming methodology. Null and alternative hypotheses had to be specified, unless the rates could be uniquely inferred from the details of the study design.

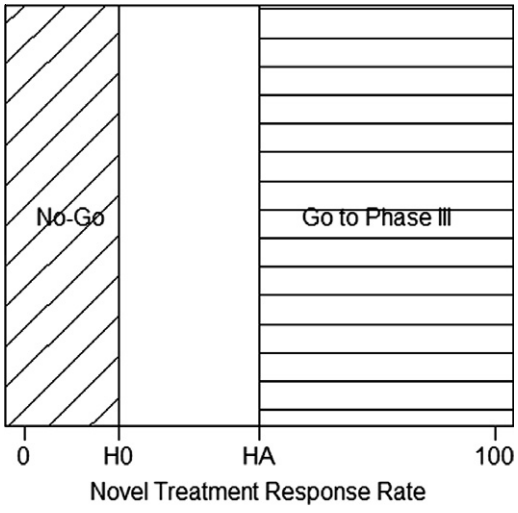

Fig. 1 – Phase II decision rules.

2.2. Data extraction

Data extraction was performed by the lead author (SA). The alternative and null hypotheses were noted. Point results for primary endpoints were collected, along with their confidence intervals (CI), and were defined as either positive, negative, or within the grey zone. Point results were defined as positive if the reported response rate (RR) was at HA or higher, negative if the reported RR was at H0 or lower and in the grey zone if the reported RR was between H0 and HA. For the CI analysis, a study was defined as positive if the lower bound of the CI exceeded the null hypothesis, negative if the upper bound of the CI was less than the alternative hypothesis and in the grey zone if the CI covered both the null and alternative hypotheses. Results were then compared to the final verbal conclusions, classified as either positive, negative or ambiguous (Table 1). Concordance and discordance rates were then calculated.

3. Results

3.1. Study eligibility

A total of 168 phase II trials were retrieved and reviewed. Six trials were multi-armed with hypotheses for each arm tested independently allowing for 174 separate analyses. Seven trials had insufficient data to analyse, one trial was incomplete and 14 trials did not have a clear null or alternative hypothesis, resulting in 152 eligible analyses. Fifty two trials provided no confidence intervals, while of the remaining 100, 97 reported 95% confidence intervals, while 3 trials reported 90% confidence intervals and all were two-sided. Sixty-eight articles were from the year 2006 and 84 articles were from the year 2007.

The median sample size of these studies was 40 (range 12–97) patients, and the majority of trials were performed in lung (36% or 23.7%), breast (17% or 11.2%), gastrointestinal (47% or 30.9%) or genitourinary (17% or 11.2%) cancer patients. Thirty-eight (25.0%) were industry-sponsored, 54 (35.5%) were funded by an academic source and 60 (39.5%) did not disclose their funding source.

3.2. Estimates

Using the reported response rates, 42 (27.6%) of the studies were negative, 60 (39.4%) were positive and 50 (32.9%) were in the grey zone. Of the 100 studies which reported confidence intervals, 8 (8.0%) were positive and 92 (92.0%) were in the grey zone. Of the 52 studies which did not report confidence intervals, 28 (53.9%) had point estimates which were negative, 16 (30.8%) were positive and 8 (15.4%) were in the grey zone. None of the studies reported confidence intervals in which the upper bound was less than HA, which can be partially explained because of the two-stage procedure. Those studies with really poor results are stopped after the first stage, at

Table 1 – Examples of the study's classification of verbal conclusions of phase II oncology trials.

Positive	Drug X demonstrates high efficacy Drug X is an active agent against Y Drug X has a meaningful, durable response
Negative	Drug X did not produce clinical response Overall response rate was low Drug X inactive against Y
Ambiguous	Drug X is well tolerated Further investigation of drug X is warranted Drug X has action in range of other drugs

which time there are less patients on-study and confidence intervals are wider – and hence, more likely to cover HA.

Thirty-three (21.7%) trials were reported in their conclusions as negative, 111 (73.0%) were reported as positive and 8 (5.3%) were ambiguous or reported conclusions that were both positive and negative.

3.3. *Concordance and discordance rates*

Table 2 summarises the concordance/discordance between positive, negative or ambiguous study conclusions with the reported response rates and confidence intervals. If the point estimate or confidence interval was positive, the study conclusion was always reported as positive (60/60 and 8/8 studies, respectively), while 11 of 42 (26.2%) and 40 of 50 (80.0%) of studies with response rates at or below H0 or in the grey zone still reported the results as positive results.

Concordance and discordance between the point estimates and the study conclusion are illustrated in Fig. 2 (when the point estimate is negative) and Fig. 3 (when the point estimate is in the grey zone) for selected trial factors. Larger trials were more likely to report positive study conclusions even when the point estimates were less than H0 (Fig. 2A) or between H0 and HA (Fig. 3A). There was no observable difference between tumour sites and reporting of positive or ambiguous conclusions (Figs. 2B and 3B), however, industry-sponsored and undisclosed-sponsored studies were more likely to report positive trial conclusions when the point estimate was negative (Fig. 2C) or in the grey zone (Fig. 3C) than in non-industry-sponsored studies.

Table 2 – Concordance/discordance of verbal conclusions with point estimates and confidence intervals.

	N	Study conclusion	*n* (%)
Point estimates			
Negative	42	Negative	28 (66.7%)
		Positive	11 (26.2%)
		Ambiguous	3 (7.1%)
Positive	60	Negative	0 (0.0%)
		Positive	60 (100.0%)
		Ambiguous	0 (0.0%)
Grey zone	50	Negative	5 (10.0%)
		Positive	40 (80.0%)
		Ambiguous	5 (10.0%)
Confidence intervals			
Negative	0	Negative	0 (–)
		Positive	0 (–)
		Ambiguous	0 (–)
Positive	8	Negative	0 (0.0%)
		Positive	8 (100.0%)
		Ambiguous	0 (0.0%)
Grey zone	92	Negative	14 (15.2%)
		Positive	73 (79.4%)
		Ambiguous	5 (5.4%)
Not reported	52	Negative	19 (36.5%)
		Positive	30 (57.7%)
		Ambiguous	3 (5.8%)

4. Discussion

Inaccuracies due to statistical misconceptions are frequently observed in phase 2 oncology clinical trials based on the Simon–Fleming model.[4] This results from the mixing of different statistical theories which is not well understood by non-statisticians and remains an easy source of confusion. A number of reviews have highlighted the mixing of different theories[8–13] and this will be briefly summarised below.

4.1. *Statistical theory*

The Simon–Fleming model is based on the hypothesis testing framework, first described by Neyman and Pearson in 1928.[8] In this framework, one selects two hypotheses (H0 and HA) and at the end of the study, one rejects one of these hypotheses and accepts the other. The theory is designed as a way of choosing between one of the two competing hypotheses, and does not allow for any other possible hypothesis to be considered. Additionally, there is no measure of the weight of evidence (e.g. as reported by the *p*-value), one only concludes whether H0 or HA was accepted.[9] Two error rates, α and β, are defined to reflect the consequence of selecting the incorrect hypothesis at the study conclusion.

In practice, however, the hypothesis testing framework is used primarily for trial design and sample size calculations, but is rarely used for analysis. Most investigators provide information about the weight of evidence regarding the level of agreement between the data and H0, typically reported using the *p*-value. This measurement, described initially by Fisher, was to reflect the plausibility of H0, but the ultimate study conclusion was to be left to the investigators who could combine this information with all available data.[10] The relative merits of these statistical theories have been extensively debated, often with considerable vitriol, however in time they have become jumbled.[11] This is the source for many of the misunderstandings of phase II cancer clinical trial results.

4.2. *The Simon–Fleming model*

It is fairly well understood that the specification of the null and alternative hypotheses relates to 'uninteresting' and 'interesting' response rates. What is less commonly understood, however, is that in the hypothesis testing framework of the Simon–Fleming design, the study conclusion *must be* one of these two conclusions. Either the new treatment has a true, population-wide RR defined under H0 or it has a true,

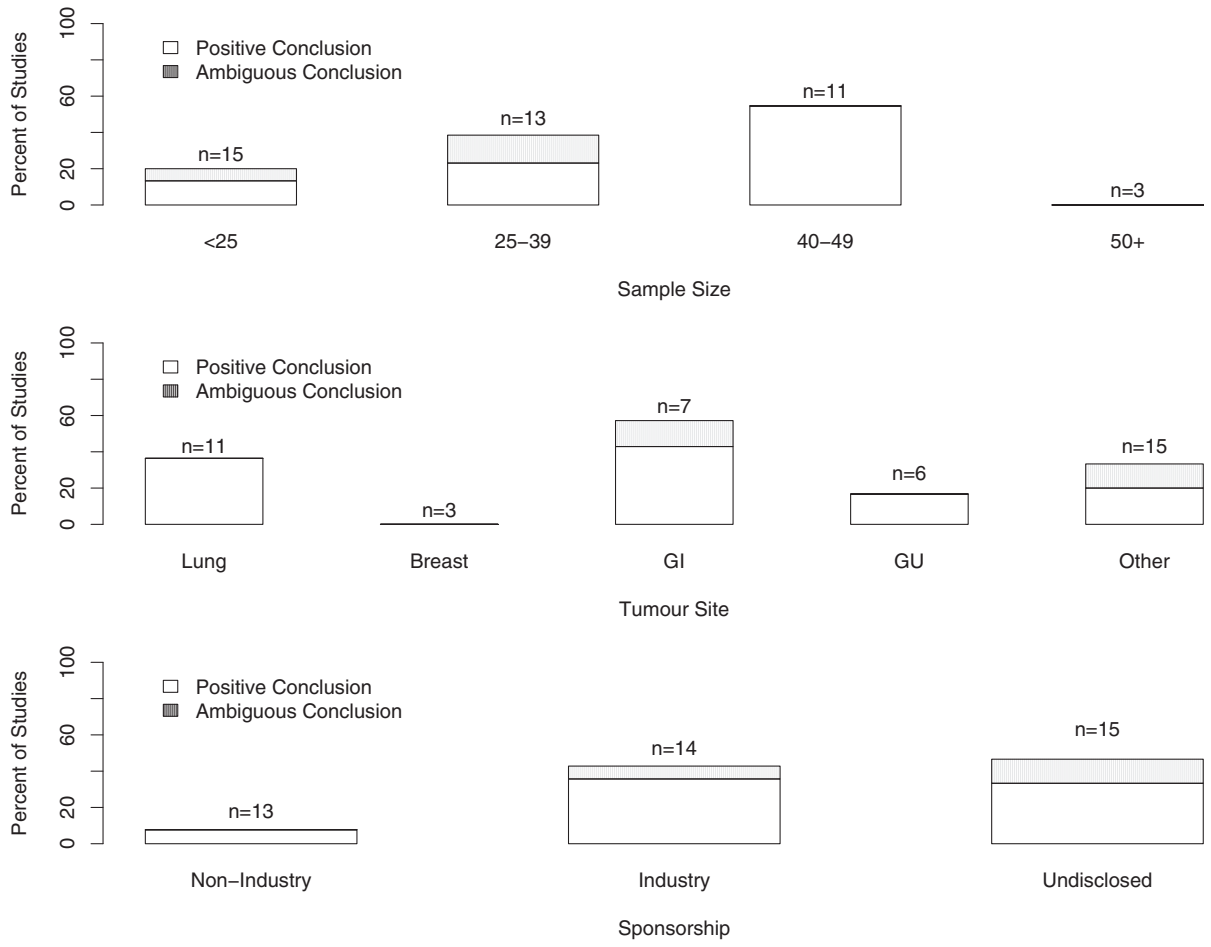

Fig. 2 – Percent of studies with a *negative point estimate* in which the study conclusion is positive or ambiguous by (a) sample size, (b) tumour site and (c) sponsorship.

population-wide RR defined under HA. It is impossible to conclude that the new treatment has a true RR which lies between H0 and HA. For example, one common phase II Fleming design is based on H0: RR = 5% versus HA: RR = 20%. At the end of the second stage, if 3/30 or less patients respond, one would accept H0, however, if 4/30 or more patients respond one would accept HA. For this design, if the observed response rate in a trial of 30 patients is 10% or less, the study conclusion is that H0 is true (i.e. the true RR is at most 5%). Alternatively, if one more patient had a response and the observed response rate is 13.3% or more, the study conclusion is that HA is true (i.e. the true response rate is at minimum 20%).

At the end of a study, however, investigators typically do not report the study conclusion based solely on the hypothesis test decision rule. Typically, investigators report *p*-values and confidence intervals based on the Fisher paradigm. By doing so, investigators are no longer using the Neyman–Pearson theory of selecting between two possible hypotheses, but are using a theory which describes the plausibility of one hypothesis (H0). This is noted by observing that the *p*-value gives evidence about H0, however, the *p*-value is not associated in any way to HA. For example, if one calculates a *p*-value at the end of a study, the *p*-value is the same regardless of whether one designed the trial using H0: RR = 5% versus HA: RR = 20%, H0: RR = 5% versus HA: RR = 15%, or H0: RR = 5% versus HA: RR = 25%. A low *p*-value indicates investigators have some evidence to reject H0. As Ratain and Karrison point out,[4] however, rejecting H0 based on a low *p*-value does not imply that investigators can therefore accept HA, because there is the possibility that neither H0 nor HA is correct. Confidence intervals, which show a range of values which would be consistent with the observed results, similarly do not imply that HA is correct.

Finally, some common analytical practices compound the problem. Confidence intervals and *p*-values are typically reported as two-sided at the 0.05 level of significance, whereas the Simon–Fleming design is a one-sided hypothesis test, and often designed with $\alpha \neq 0.05$. To be consistent with the original Simon–Fleming design, analyses should be one-sided, as these designs are based on one-sided testing procedures, and confidence intervals should be set to the $(1 - \alpha) * 100\%$ level. Many journals require two-sided analyses. In these situations, the rationale for using two-sided analyses should be discussed and confidence intervals set at the $(1–2 * \alpha) * 100\%$ level. Additionally, confidence intervals and *p*-values are regularly reported without adjustment for the performed interim analysis, resulting in a bias.[13] The statistical theory underlying non-adjusted calculations include the possibility of 0 responses after stage II, however, in practice, this never occurs because this trial would be stopped after the first stage

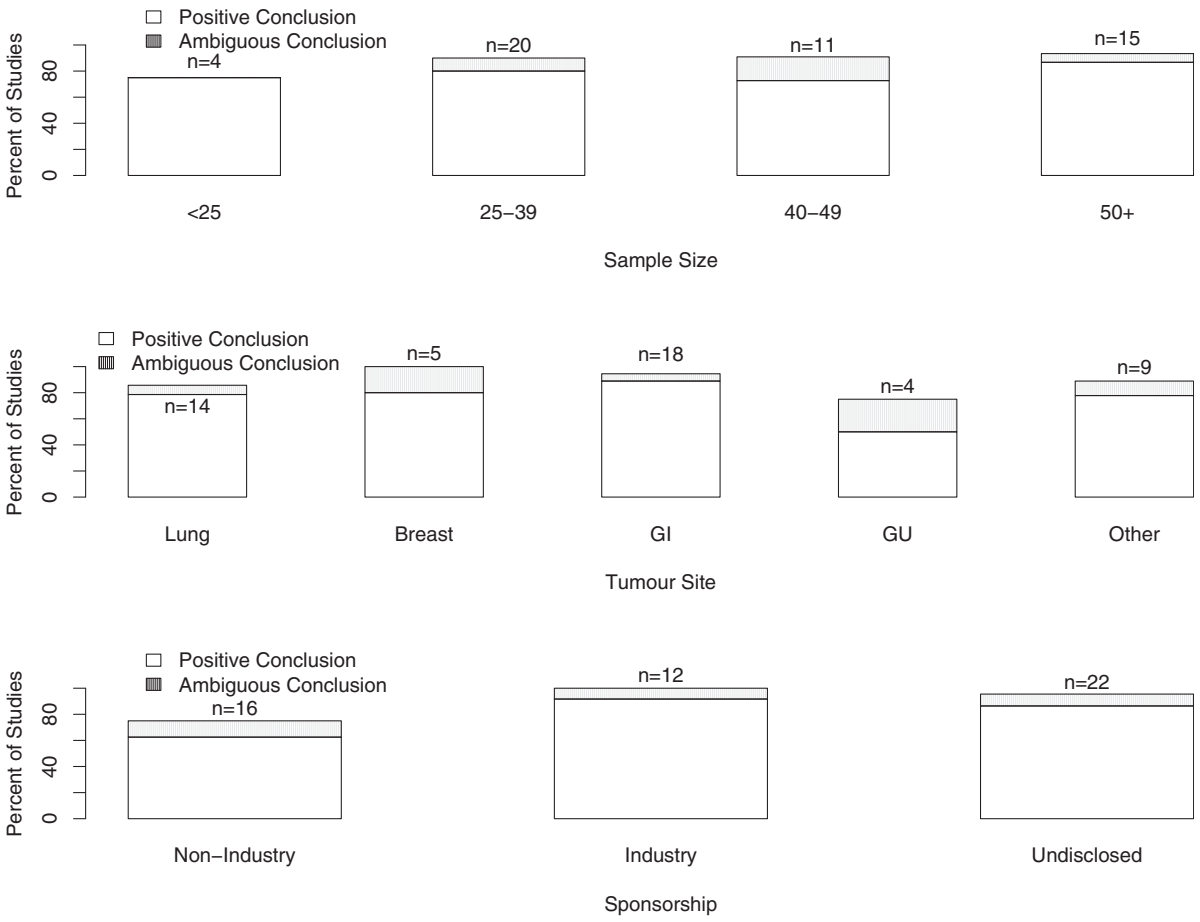

Fig. 3 – Percent of studies with a point estimate in the *grey zone* in which the study conclusion is positive or ambiguous by (a) sample size, (b) tumour site and (c) sponsorship.

of accrual. The result is a jumble of statistical theories which ultimately causes confusion and unclear conclusions.

The results of this review highlight the confusion caused by this mix of different statistical theories. It is believed that this is the first published research evaluating the issue of concordance between results which fall in the grey zone and study conclusions in a systematic fashion. When reviewing point estimate results of phase II trials an almost equal division among positive, negative and grey zone values was observed. Yet, there are a disproportionate number of trials described with positive conclusions. All trials in which the observed point estimate for the trial response rate was equal to the alternative hypothesis led investigators to give positive study conclusions. However, when results were negative, the study conclusion was similarly negative in only two-thirds of the trials. Further, when the observed response rate was between H0 and HA, the study conclusion was positive 80% of the time. This indicates that investigators appear to base study conclusions on whether the observed results are better than the standard of care results (i.e. H0) and do not account for any random variation which occurs. Alternatively, investigators may be accepting HA – a hypothesis testing notion – based on *p*-values or confidence intervals which are incompatible with hypothesis testing. Ultimately, this raises questions regarding the authenticity of the natural development of a therapeutic agent through clinical trials. If a phase II conclusion is erroneously factored into the development of a drug, then potential scientific, clinical and economic losses can occur.

5. Conclusions

There exist inconsistencies and ambiguities in the conclusions drawn from phase II trials, resulting, at least in part, from misconceptions and mixing up of different statistical concepts. This results in poor reliability of phase II trial results[1,4] and in turn, is likely to be a large contributor to the poor success rate of cancer therapies in phase III.[14,15] Alternative trial designs, such as randomised comparison phase II trials,[16,17] three-outcome designs[18] or Bayesian designs,[19] might eventually improve the success rate of cancer therapies in phase III trials. However, improved efficiency in anti-cancer drug development will not transpire regardless of trial design improvements if misconceptions about the statistical theory continue to occur.

Conflict of interest statement

None declared.

Role of funding source

There are no funding sources to disclose for this research.

REFERENCES

1. 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**(3):972–6.
2. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982;**38**:143–51.
3. Simon R. Optimal two-stage design for phase II clinical trials. *Control Clin Trials* 1989;**10**:1–10.
4. Ratain MJ, Karrison TG. Testing the wrong hypothesis in phase II oncology trials: there is a better alternative. *Clin Cancer Res* 2007;**13**(3):781–2.
5. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumours. *J Natl Cancer Inst* 2000;**92**:205–16.
6. WHO handbook for reporting results of cancer treatment. Geneva (Switzerland): World Health Organization Offset Publication No. 48; 1979.
7. Roberts Jr TG, Lynch Jr TJ, Chabner BA. The phase III trial in the era of targeted therapy: unraveling the "go or no go" decision. *J Clin Oncol* 2003;**21**:3683–95.
8. Goodman SN. *P* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate (plus commentary). *Am J Epidemiol* 1993;**137**(5):485–501.
9. Berger JO. Could Fisher, Jeffreys and Neyman have agreed on testing? (plus commentary). *Stat Sci* 2003;**18**(1):1–32.
10. Blau DJ, Jolles BM, Porcher R. *P* value and the theory of hypothesis testing. *Clin Orthop Relat Res* 2010;**468**:885–92.
11. Lehmann EL. The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc* 1993;**88**(424):1242–9.
12. Marden JI. Hypothesis testing: from *p* values to Bayes factors. *J Am Stat Assoc* 2000;**95**(452):1316–20.
13. Chang MN, O'Brien PC. Confidence intervals following group sequential tests. *Control Clin Trials* 1986;**7**(1):18–26.
14. Michaelis LC, Ratain M. : Measuring response in a post-RECIST world: from black and white to shades of grey. *Nat Rev Cancer* 2006;**6**:409–14.
15. Booth B, Glassman R, Ma P. Oncology's trials. *Nat Rev Drug Discov* 2003;**2**:609–10.
16. Rubinstein VL, Korn EL, Freidlin B, et al. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol* 2005;**23**(28):7199–206.
17. Steinberg SM, Venzon DJ. Early selection in a randomized phase II clinical trial. *Stat Med* 2002;**21**:1711–21.
18. Sargent DJ, Chan V, Goldberg RM. A three-outcome design for phase II clinical trials. *Control Clin Trials* 2001;**22**(2):117–25.
19. Thall PF, Simon R. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Control Clin Trials* 1994;**15**(6):463–81.