available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

Effect of a misspecification of response rates on type I and type II errors, in a phase II Simon design

Charlotte Baey [a], Marie-Cécile Le Deley [a,b,*]

[a] Service de biostatistique et d'épidémiologie, Institut Gustave Roussy, 114 rue Edouard Vaillant, 94805 Villejuif, France
[b] Université Paris-Sud 11, 63 rue Gabriel Péri, 94276 Le Kremlin-Bicêtre, France

ARTICLE INFO

Article history:
Received 24 February 2011
Accepted 10 March 2011
Available online 13 April 2011

Keywords:
Phase II trials
Simon design
Randomised phase II
Type I error
Type II error

ABSTRACT

Phase-II trials are a key stage in the clinical development of a new treatment. Their main objective is to provide the required information for a go/no-go decision regarding a subsequent phase-III trial. In single arm phase-II trials, widely used in oncology, this decision relies on the comparison of efficacy outcomes observed in the trial to historical controls. The false positive rate generally accepted in phase-II trials, around 10%, contrasts with the very high attrition rate of new compounds tested in phase-III trials, estimated at about 60%. We assumed that this gap could partly be explained by the misspecification of the response rate expected with standard treatment, leading to erroneous hypotheses tested in the phase-II trial.

We computed the false positive probability of a defined design under various hypotheses of expected efficacy probability. Similarly we calculated the power of the trial to detect the efficacy of a new compound for different expected efficacy rates. Calculations were done considering a binary outcome, such as the response rate, with a decision rule based on a Simon two-stage design.

When analysing a single-arm phase-II trial, based on a design with a pre-specified null hypothesis, a 5% absolute error in the expected response rate leads to a false positive rate of about 30% when it is supposed to be 10%. This inflation of type-I error varies only slightly according to the hypotheses of the initial design.

Single-arm phase-II trials poorly control for the false positive rate. Randomised phase-II trials should, therefore, be more often considered.

1. Introduction

Efficacy and safety of a new treatment are evaluated during the different phases of clinical development. Following phase I trials, the main objective of phase II trials in oncology is to provide information required for a go/no-go decision regarding subsequent phase III trials. Phase II trials are, therefore, a key stage in the development of a new treatment. While they were initially designed for the evaluation of new drugs given as single agents (phase IIA), they are now widely used to assess efficacy of various medical treatments including combination of several drugs or treatments including both chemotherapy and radiation therapy (phase IIB).

In the early 1980s, when there were a small number of new drugs available, the main consideration was to minimise the false-negative rate of phase II trials, to ensure that a high number of effective drugs would be assessed through phase III trials. In the past decade, the number of new anticancer

* *Corresponding author*: Address: Service de biostatistique et d'épidémiologie, 114 rue Edouard Vaillant, 94805 Villejuif, France. Tel.: +33 142 115 444.
E-mail address: marie-cecile.ledeley@igr.fr (M.-C. Le Deley).

0959-8049/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ejca.2011.03.013

agents has dramatically increased, and there is now a major concern about the very high attrition rate between phase II and phase III trials, estimated at about 60%,[1] raising the issue of false-positive rate of phase II trials.[2,3]

Even though several designs have been proposed[4–8] including randomised trials and bayesian approaches[9–11], non-randomised single-arm trials using a frequentist approach remain the most frequently used in oncology.[2] In such trials, the decision taken after completion of the trial is implicitly based on the comparison of observed data to historical data, formally addressed through the null hypothesis which is tested.

The Simon design is a sequential two-stage design specifying type I and type II errors (false-positive and false-negative rate, respectively) for defined response rates: the rate below which the treatment would be judged 'not promising', p_0, and the rate above which it would be worthy of further development, p_1. These two parameters are defined *a priori*, from historical data or experts' opinion, and their choice is in practice rarely explained. We study here the impact of a misspecification of one or both prespecified response rates on the type I and type II errors, using a Simon optimal two-stage design.

2. Methods

2.1. Decision rule in a Simon design

Five parameters need to be specified before the trial: p_0 and p_1, α and β the type I and II errors, and $N_{\max}$, the maximum number of patients to be included in the trial.[8]

p_0 is the baseline response rate expected to be observed with a standard treatment. If the true probability of response with the new treatment is equal to or lower than p_0 (null hypothesis H_0), we wish to avoid concluding promising efficacy of the new treatment. The probability α is set as the maximum probability of concluding efficacy of the new treatment when the true response rate is equal to or lower than p_0.

On the other hand, if the true response rate with the new treatment is equal to or greater than p_1, we wish a high $(1 - \beta)$ probability of concluding that the efficacy of the new treatment is promising. The difference $\Delta = p_1 - p_0$ is the effect size you hope to observe with the new treatment.

For given p_0, p_1 type I error α and type II error β, the Simon design is defined by searching all possible combinations of sample sizes and decisions rules and selecting those resulting in the specified type I and type II errors, subject to a maximal sample size of $N_{\max}$. Simon proposed either to minimise the expected sample size (Optimal design) or to minimise the maximum sample size (Minimax design) under the null hypothesis. The number of responses at the first stage s_1 (resp. at the second stage, s_2) follows a binomial distribution with parameters n_1, p (resp. n_2, p with $n_2 = N - n_1$), where p is the true response rate. The type I error α is calculated under the null hypothesis H_0: $p = p_0$ and the type II error β under a chosen alternative H_1: $p = p_1$. The Simon design does not plan to stop the trial at first stage for efficacy, however, in the calculation of type I error you have to take into account the case where you already observe the K responses at the first stage.

As an example, let $p_0 = 0.20$, $p_1 = 0.40$, $\alpha = 0.10$ and $\beta = 0.10$. The corresponding optimal and minimax designs are ($n_1 = 17$, $k_1 = 3$, $N = 37$, $K = 10$) and ($n_1 = 19$, $k_1 = 3$, $N = 36$, $K = 10$), respectively. Considering the optimal design, the trial is stopped for inefficacy after the analysis of $n_1 = 17$ patients, if the number of responses is equal to or less than $k_1 = 3$. Conversely, recruitment continues to reach a total of $N = 37$ patients if more than 3 responses are observed among the 17 first patients. At the end of second step, the treatment is deemed unpromising if the total number of responses is equal to or less than $K = 10$ among the 37 patients, and promising otherwise. Using this rule, the risk of falsely concluding efficacy when the true response rate is equal to or lower than $p_0 = 0.20$ is $\alpha = 0.0948$. If the true response rate is equal to or greater than $p_1 = 0.40$, then the risk of wrongly concluding inefficacy is $\beta = 0.0967$.

2.2. Calculation of the errors associated with a specified design

In our previous example, the new treatment was deemed unpromising if the true response rate was equal to 0.20. Let us consider, as an example, what occurs when using the decision rule ($n_1 = 17$, $k_1 = 3$, $N = 37$, $K = 10$) if a shift in outcome (due for example to a change in standard of care, staging or to improved surgical or radiation techniques[12]) has lead the true response rate with the standard treatment to 0.25. Similarly, what happens if we expect a 0.45 response rate with the new treatment? Type I and II errors can be re-calculated under these new hypotheses, applying decision rule defined according to the initial hypothesis.

Let ($n_{1,\mathrm{h}}$, $k_{1,\mathrm{h}}$, N_h, K_h) be the decision rule of an optimal Simon plan corresponding to prespecified $p_{0,\mathrm{h}}$, $p_{1,\mathrm{h}}$, α_h and β_h. Based on the decision rule defined in each situation, we calculated type I and II errors if the true response rates varied from $p_0 = p_{0,\mathrm{h}} - 0.10$ to $p_0 = p_{0,\mathrm{h}} + 0.10$ and from $p_1 = p_{1,\mathrm{h}} - 0.10$ to $p_1 = p_{1,\mathrm{h}} + 0.10$. We first calculated errors using an optimal

Table 1 – Description of the seven optimal Simon design scenarii considered. The effect size $p_{1,\mathrm{h}} - p_{0,\mathrm{h}}$ remains equal to 0.20, and the type I and II errors are fixed to 0.10.

$p_{0,\mathrm{h}}$	$p_{1,\mathrm{h}}$	α_h	β_h	$n_{1,\mathrm{h}}$	$k_{1,\mathrm{h}}$	N_h	K_h	Calculated α	Calculated β
0.10	0.30	0.10	0.10	12	1	35	5	0.0977	0.0985
0.15	0.35	0.10	0.10	19	3	33	7	0.0962	0.0957
0.20	0.40	0.10	0.10	17	3	37	10	0.0948	0.0967
0.25	0.45	0.10	0.10	14	3	44	14	0.0967	0.0986
0.30	0.50	0.10	0.10	22	7	46	17	0.0973	0.0950
0.35	0.55	0.10	0.10	20	7	47	20	0.0938	0.0950
0.40	0.60	0.10	0.10	18	7	46	22	0.0952	0.0996

design with fixed $\alpha = 0.10$ and $\beta = 0.10$. Seven cases were considered with various p_0 but a fixed expected effect size $\Delta = 0.20$ (see Table 1).

We extended this approach by computing type I error for various designs with: (i) $\Delta = 0.20$ or $\Delta = 0.15$, (ii) optimal or minimax design, (iii) $\beta = 0.10$ or $\beta = 0.05$ and (iv) $\alpha = 0.10$ or $\alpha = 0.05$. For each combination of these parameters, seven designs have been defined, for an initial $p_{0,h}$ varying from 0.10 to 0.40, all other parameters remaining unchanged. All these plans are detailed in Webtable 4, given in appendix. The maximal sample size varies from 33 (optimal design, $p_{0,h} = 0.15$, $\Delta = 0.20$, $\alpha_h = 0.10$, $\beta_h = 0.10$) to 137 (optimal design, $p_{0,h} = 0.30$, $\Delta = 0.15$, $\alpha_h = 0.05$, $\beta_h = 0.05$). We then calculated the mean α of the seven plans for each combination of the different parameters, and for a misspecification of the $p_{0,h}$ rate varying from 0 to +0.10.

3. Results

3.1. *Example*

Applying the decision rule of the previous example, type I error equals 0.0948 if $p_0 = 0.20$ but reaches 0.284 if $p_0 = 0.25$, meaning that there is more than 28% risk of concluding promising efficacy if the true response rate with the new treatment is lower than or equal to 0.25, which is much higher than the accepted type I error of 0.10.

Conversely, if the true response rate p_1 is equal to 0.45 instead of 0.40, type II error equals 0.032, quite lower than the fixed type II error of 0.10. Thus, with an effect size similar to that defined ($\Delta = 0.20$), a misspecification of −0.05 leads to a major increase of type I error while type II error decreases. A misspecification of +0.05 in both parameters p_0 and p_1 conversely leads to major decrease of the power (type II error equals to 0.2357 instead of 0.10).

3.2. *Variation of the error rates according to the shift in expected response rate*

Fig. 1 represents the variation of α according to the error on p_0 for various optimal designs with fixed $\alpha = 0.10$, $\beta = 0.10$ and an expected effect size $\Delta = 0.10$, each curve corresponding to one of the designs described in Table 1. The X axis represents the true response rate of the null hypothesis. The central value is the initial $p_{0,h}$ specified in the design: moving to the right of the axis corresponds to an underestimation of the p_0 rate, moving to the left corresponds to an overestimation of this rate. The Y axis represents the calculated type I error.

As expected, an under-estimation of p_0 leads to higher type I error, and an over-estimation leads to smaller type I error. Even a small error for p_0, for instance +0.05 misspecification, induces an average increase of approximately 0.20 in type I error: it is approximately 0.30 when the design specification was controlled at 0.10. As illustrated in Fig. 1, the increase of type I error varies only slightly according to the initial null hypothesis: it reaches 0.355 if $p_0 = 0.15$ while we specified $p_{0,h} = 0.10$ in the initial plan and 0.222 if $p_0 = 0.45$ while we specified $p_{0,h} = 0.40$.

Likewise, Fig. 2 represents the variation of β according to the error on p_1. A misspecification of +0.05 in p_1 leads to a better power, when a misspecification of −0.05 induces a major loss of power.

3.3. *Generalisation*

As detailed in Tables 2 and 3, the inflation of the type I error is very similar between optimal and minimax designs for the given parameters. This inflation is slightly lower in the minimax design although the maximal sample size is lower. When the fixed type II error is reduced from 0.10 to 0.05, the misspecification of the initial parameters has an even larger impact on type I error despite the increased sample size.

For an error of +0.05 on p_0, the mean type I error is equal to 0.288 when using an optimal design with a type II error of 0.10 and an effect size of 0.20, compared to 0.272 with the minimax design, all other parameters remaining unchanged. This mean type I error rates increase to 0.319 if the type II error is fixed to 0.05, and to 0.369 if the expected effect size is equal to 0.15.

Comparing the mean actual type I error to that specified initially, the inflation related to the misspecification of p_0 is quite similar whether type I error is initially fixed to 0.10 or 0.05 (mean error equals to 0.354 and 0.281 respectively).

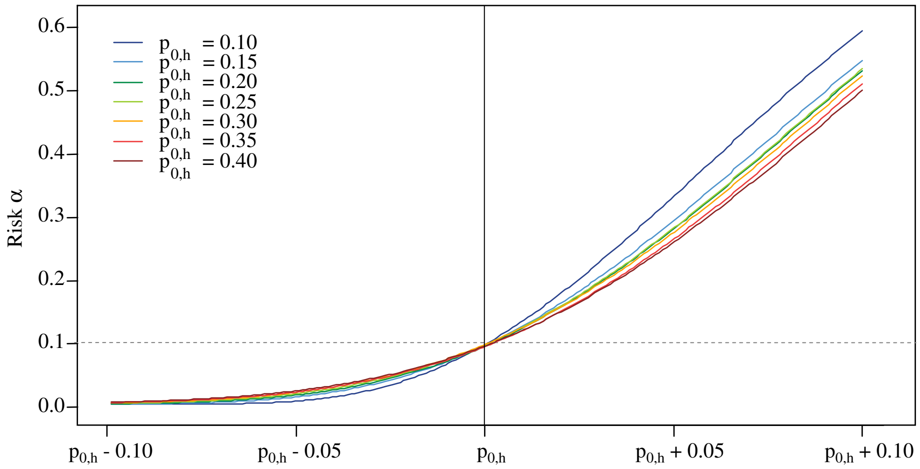

Fig. 1 – Variation of α according to p_0 value, for $p_{1,h} - p_{0,h} = 0.20$, and $\alpha_h = \beta_h = .010$.

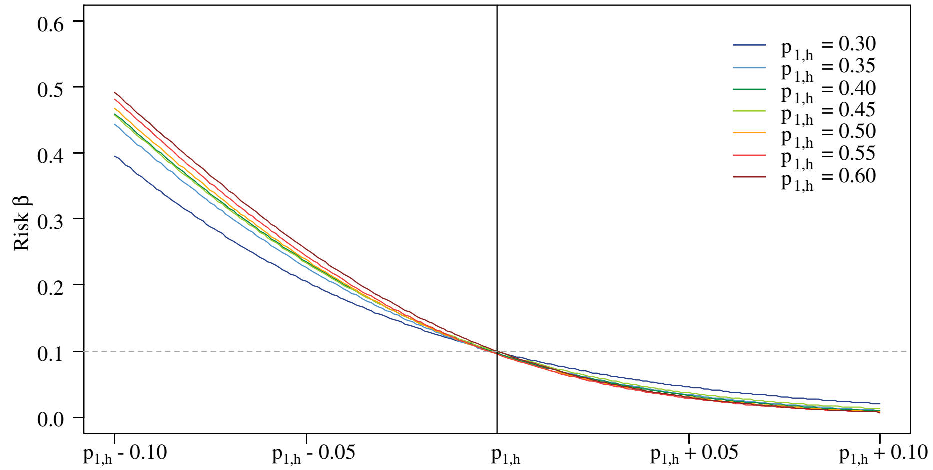

Fig. 2 – Variation of β according to p_1 value, for $p_{1,h} - p_{0,h} = 0.20$, and $\alpha_h = \beta_h = 0.10$.

Table 2 – Mean calculated α over the seven Simon designs obtained for a given effect size and a given β, and for an initially fixed $\alpha = 0.10$. N corresponds to the range of sample size observed in the seven designs.

Error on p_0	$\Delta = 0.20$				$\Delta = 0.15$			
	$\beta = 0.10$		$\beta = 0.05$		$\beta = 0.10$		$\beta = 0.05$	
	Optimal	Minimax	Optimal	Minimax	Optimal	Minimax	Optimal	Minimax
	N=33–47	N=35–42	N=40–62	N=33–55	N=50–88	N=40–73	N=66–107	N=55–96
0.000	0.096	0.092	0.093	0.095	0.099	0.096	0.098	0.096
0.005	0.110	0.105	0.109	0.110	0.118	0.114	0.119	0.117
0.010	0.125	0.119	0.127	0.127	0.139	0.134	0.144	0.139
0.015	0.142	0.135	0.146	0.146	0.161	0.155	0.170	0.165
0.020	0.160	0.151	0.166	0.165	0.186	0.179	0.200	0.193
0.025	0.179	0.169	0.188	0.187	0.213	0.204	0.232	0.223
0.030	0.199	0.188	0.212	0.209	0.242	0.231	0.266	0.255
0.035	0.220	0.208	0.237	0.233	0.272	0.260	0.302	0.290
0.040	0.242	0.228	0.263	0.258	0.303	0.290	0.340	0.326
0.045	0.264	0.250	0.290	0.284	0.336	0.322	0.379	0.363
0.050	0.288	0.272	0.319	0.311	0.369	0.354	0.419	0.402
0.055	0.312	0.296	0.348	0.339	0.403	0.387	0.459	0.441
0.060	0.337	0.320	0.377	0.368	0.438	0.421	0.499	0.480
0.065	0.362	0.344	0.407	0.397	0.473	0.455	0.539	0.519
0.070	0.388	0.369	0.438	0.426	0.507	0.489	0.578	0.558
0.075	0.414	0.394	0.468	0.456	0.541	0.523	0.616	0.596
0.080	0.440	0.420	0.498	0.485	0.575	0.557	0.652	0.632
0.085	0.466	0.446	0.528	0.515	0.607	0.590	0.687	0.668
0.090	0.492	0.471	0.558	0.544	0.639	0.622	0.720	0.701
0.095	0.518	0.497	0.587	0.573	0.669	0.653	0.751	0.733
0.100	0.543	0.523	0.616	0.601	0.698	0.683	0.779	0.763

It is noticeable that a larger sample size does not reduce this inflation. For example, for an error of +0.05 on p_0, the mean α is 0.2984 for the design corresponding to the smallest maximal sample size, and 0.3313 for the design with the largest maximal sample size. Simply stated, using more patients to test the wrong hypothesis does only increase the likelihood of an erroneous conclusion.

4. Discussion

The conclusions drawn here can be generalised to any of the many proposed single arm designs based on comparisons with historical data. The decision rules upon which these trials are based depend on fixed parameters, especially the response rates p_0 and p_1, which are derived from historical data. Their choice appears often subjective and questionable: when planning a new trial, it is frequent to receive different proposals for p_0 and p_1 according to the interviewed experts.

The estimation of the response rates provided by historical data may not be valid for the current clinical practice, due to patient selection or shift in the outcomes. This shift can be due for example to a change in stage classification, imaging, standard of care or to improvements in surgical or radiation areas. Moreover, the population considered in the actual trial may differ from the population on which the historical rate was based: different ages, performance status, known or

Table 3 – Mean calculated α over the seven Simon designs obtained for a given effect size and a given β, and for an initially fixed $\alpha = 0.05$. N corresponds to the range of sample size observed in the seven designs.

Error on p_0	$\Delta = 0.20$				$\Delta = 0.15$			
	$\beta = 0.10$		$\beta = 0.05$		$\beta = 0.10$		$\beta = 0.05$	
	Optimal	Minimax	Optimal	Minimax	Optimal	Minimax	Optimal	Minimax
	N=35–66	N=33–54	N=49–79	N=41–67	N=66–110	N=55–94	N=80–137	N=70–119
0.000	0.050	0.046	0.048	0.046	0.049	0.047	0.042	0.047
0.005	0.060	0.055	0.058	0.056	0.062	0.059	0.055	0.061
0.010	0.071	0.065	0.071	0.067	0.077	0.073	0.071	0.077
0.015	0.084	0.077	0.085	0.080	0.095	0.090	0.090	0.096
0.020	0.098	0.089	0.102	0.095	0.115	0.108	0.112	0.119
0.025	0.114	0.104	0.120	0.112	0.138	0.129	0.137	0.144
0.030	0.131	0.119	0.140	0.130	0.164	0.153	0.165	0.173
0.035	0.150	0.136	0.162	0.151	0.192	0.179	0.196	0.204
0.040	0.170	0.154	0.185	0.173	0.222	0.207	0.230	0.239
0.045	0.192	0.174	0.211	0.196	0.255	0.238	0.266	0.276
0.050	0.215	0.194	0.238	0.222	0.290	0.270	0.304	0.315
0.055	0.239	0.217	0.266	0.248	0.326	0.304	0.344	0.356
0.060	0.264	0.240	0.296	0.277	0.363	0.340	0.386	0.399
0.065	0.291	0.264	0.327	0.306	0.401	0.377	0.428	0.442
0.070	0.318	0.290	0.359	0.336	0.440	0.415	0.470	0.486
0.075	0.346	0.316	0.392	0.368	0.479	0.453	0.513	0.529
0.080	0.375	0.343	0.425	0.400	0.518	0.492	0.554	0.572
0.085	0.404	0.371	0.459	0.432	0.556	0.530	0.595	0.614
0.090	0.434	0.399	0.492	0.465	0.593	0.568	0.634	0.655
0.095	0.463	0.427	0.525	0.498	0.629	0.605	0.672	0.693
0.100	0.493	0.456	0.558	0.530	0.663	0.641	0.707	0.729

unknown molecular sub-types, etc. Even if the two populations are comparable, there is still variability around the historical response rate due to statistical estimation.[2,12,13] In some cases, one can even face the complete absence of historical data, due to the rarity of the disease, the new definition of the subset of patients to be treated with a targeted drug, or the use of a new primary endpoint. Despite these considerations, in the common single-arm approach, one specific decision rule is defined, based on a single null hypothesis. This rule is then used throughout the trial and the results are described in terms of estimation of the response rate, compared to fixed initial hypotheses, ignoring the high uncertainty surrounding the pre-stated hypotheses.

As we have demonstrated, a misspecification of these design parameters could lead to uncontrolled type I and II errors, whatever the specified hypotheses tested. An under-estimation of p_0 (historical rate) induces an increased type I error, when an over-estimation of p_1 implies a loss of power. Even a small error in the choice of the initial parameters has a major impact on the results, whatever the choice of the design (optimal or minimax), the fixed parameters (α, β, Δ) or the sample size.

In this context, randomised phase II trials, in which a control group is treated with a standard treatment appear attractive, as they reduce the potential bias resulting from inter-trial variability when historical controls are used.[2] In a recent article, Tang et al. compared the errors rates between single-arm and randomised phase II trials.[12] They showed that two factors highly influence type I and II errors in single-arm trials: a misspecification of the p_0 rate and a change in the patient population. Thus, in single-arm studies, type I error can reach 48% instead of the pre-specified 20% when considering both a change in patients selection and a shift in the true response rate from the historical data, whereas it remains closed to 20% in randomised designs.

The two main objections to the use of randomised phase II trials are the increased number of patients it implies,[12] and the fact that it would resemble a phase III trial.[3] This last criticism is not justified, since randomised phase II trials still differ from phase III trials on several points: the type I and II errors in phase II trials are usually higher (around 10% or 20%), the patients included in the study are not necessarily representative of the target population, and the end-points generally differ between phase II and phase III. Concerning the higher number of patients to be included compared to a single-arm study, this is an unavoidable consequence of the fact that the hypothesis tested in the single-arm trial can never be verified in practice to be valid. Randomised phase II trials will be very cost-effective if they prevent ineffective treatments being taken forward to large, and very costly, phase III trials that will be negative. Moreover, one could imagine that a positive randomised phase II study could evolve towards a phase III trial.[14]

The impact of the choice of the tested hypothesis must be carefully considered when planning a single arm phase II trial. As the result of such trials can be biased, the use of randomised phase II design should be more often considered.

Conflict of interest statement

The authors have no conflict of interest to disclose.

Role of the funding source

The study was supported by the Ligue Nationale Contre le Cancer and Cancéropole Ile-de-France. The funding sources had no involvement in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Acknowledgements

We thank Pr. Daniel Sargent for his thoughtful and constructive comments on an earlier draft of the manuscript, and the Ligue Nationale Contre le Cancer and Cancéropole Ile-de-France for funding. The funding sources had no involvement in the design and conduct of the study; collection, management, analysis and interpretation of the data; and preparation, review, or approval of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ejca.2011.03.013.

REFERENCES

1. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov* 2004;**3**:710–5.
2. Ratain MJ, Sargent DJ. Optimizing the design of phase II oncology trials: the importance of randomization. *Eur J Cancer* 2009;**45**:275–80.
3. Cannistra SA. Phase II trials in *Journal of Clinical Oncology. J Clin Oncol* 2009;**27**:3073–6.
4. Gehan EA. The determination of the number of patients required in follow-up trial of a new chemotherapeutic agent. *J Chronic Dis* 1961;**13**:346–53.
5. Fleming TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982;**38**:143–51.
6. Koyama T, Chen H. Proper inference from Simon's two-stage designs. *Stat Med* 2008;**27**:3145–54.
7. Chen TT, Ng TH. Optimal flexible designs in phase II clinical trials. *Stat Med* 1998;**17**:2301–12.
8. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989;**10**:1–10.
9. Tan SB, Machin D. Bayesian two-stage designs for phase II clinical trials. *Stat Med* 2002;**21**:1991–2012.
10. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clin Trials* 2008;**5**:93–106.
11. Schlesselman JJ, Reis IM. Phase II clinical trials in oncology: strengths and limitations of two-stage designs. *Cancer Invest* 2006;**24**:404–12.
12. Tang H, Foster NR, Grothey A, et al. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. *J Clin Oncol* 2010;**28**:1936–41.
13. Rubinstein LV, 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**:7199–206.
14. Fazzari M, Heller G, Scher HI. The phase II/III transition: toward the proof of efficacy in cancer clinical trials. *Control Clin Trials*, 2000;**21**:360–8.