A Novel Phase II Design to Minimize Trial Duration and Improve the Success Rate of Follow-up Phase III Trial

Ye Cui[1], Taofeek K. Owonikoko[2], Zhibo Wang[3], Sungjin Kim[3], Dong M Shin[2], Fadlo R. Khuri[2], Jeanne Kowalski [3,4], Zhengjia Chen[3,4]

1. Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303.
2. Department of Hematology and Medical Oncology, Emory University, Atlanta, GA 30322.
3. Biostatistics and Bioinformatics Shared Resource at Winship Cancer Institute GA 30322.
4. Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322.

A Phase II trial is an expeditious and low cost trial with the primary goal of screening potentially effective agents prior to confirmatory Phase III trial. The success rate of Phase III oncology trials remains very low despite the success demonstrated in the preceding Phase II trials. This discordance is mainly due to the different endpoints used in Phase II (disease response) and III (survival) trials. While a robust disease response is expected to translate into survival improvement, this is NOT guaranteed. Moreover, disease response can be determined quickly whereas survival estimation requires a long period of follow up. We propose a novel 2-stage screening design for phase II trials whereby percent of tumor size change endpoint is used as an initial screening to select potentially effective agents within a short time interval followed by a second screening stage where progression free survival is estimated to confirm the efficacy of agents. This design can improve trial efficiency and reduce cost by early stopping the evaluation of an ineffective agent based on low percent of tumor size change. The second survival endpoint screening will substantially increase the success rate of follow-up Phase III trial by using the similar outcomes. We conducted simulation studies to investigate the underlying statistical considerations to optimize the significant levels of the two screening stages in the design.

Key words: Double screening, Phase II design, Success rate of Phase III trials, Percent of Tumor Size Change, Progression free survival, Cost and length of trial

1. Introduction

A Phase II trial in oncology is conducted to evaluate whether a new treatment has sufficient anticancer effect to precede subsequent Phase III study. It usually enrolls fewer than 100 people but may include as many as 300 (NCI website). As a screening trial of subsequent Phase III study, tumor response rate (RR) has been widely adopted as the primary endpoint. The assumption behind it is that higher RRs in Phase II trials associates with longer survival time which is the endpoint of the further Phase III trial. In conventional Phase II trials, tumor shrinkage is measured and assessed via the Response Evaluation Criteria in Solid Tumors (RECIST 1.1) (Eisenhauer EA, 2009), and the proportion of objective responders that two consecutive assessments of complete (CR) or partial response (PR) at least 4 weeks apart is defined as RR.

Nowadays, cancer clinical study is exceedingly important, since cancer is the leading cause of death world widely. However, high failure rate (e.g. 50-60%) of the subsequent Phase III trials obstructs the development of new treatments; and it attracts concerns over the response rate as a primary endpoint (An MW, 2011). Apprehensions about adopting response rate as a primary endpoint are recently well discussed. First, the simplicity achieved by creating only 2 response groups via RECIST 1.1 criteria has a cost: categorize of continuous data may be consequent with a loss of information (Pivot X, 2009). More fundamentally, the ultimate goal of a new drug development is to prolong survival rather than to raise RR. These concerns prompted us to propose a new two-stage Phase II design, which evaluates tumor size changes as a continuous endpoint in Stage I and estimates progression free survival (PFS) in Stage II. Our approach that includes a screening stage of survival time in order to improve Phase III success rate is beneficial to both pharmaceutical companies and patients.

Previously researches on tumor size changes as a continuous variable have been proposed to evaluate antitumor activity (Lavin PT, 1981) (Wang Y, 2009). We adopt the approach by Wang *et al.* which characterizes tumor shrinkage with both treatment effect and linear tumor growth effect. And a Wilcoxon ranked test is applied to compare tumor size data in different treatment groups in Stage I. Moreover, many recent researches confirm PFS as the best estimate of overall survival (OS) (Buyse M, 2000). Although OS remains regulatory gold standard (Halabi S, 2009) and is more reliable in classifying event status; PFS has the advantage of short median survival (Greg Y, 2007) and more informative within the protocol. PFS is a more sensitive indicator of treatment effect and is adopted in Stage II screening in our design.

2. Methods

We consider a two-stage design with a continuous outcome — percent of tumor size change (PTSC) in the first screening stage and PFS in the second at a predefined time point. Figure 1 gives a schematic of our proposed design, which we will describe in this section.

2.1 Design Framework

Patients who meet the trial "inclusion criteria" with baseline tumor size measurement are given the targeted new treatment. After a predefined period of time, we calculate PTSC according to both baseline and current tumor size. Stage I screening is taken to the experimental group with new treatment, and to the control group or historical data. A decision will be made to early terminate the trial, if there is no significant result found. Otherwise, the experiment continues to second stage, in which we will investigate difference of PFS. Similarly, the trial will be terminated when no significant difference found, or we will proceed to subsequent Phase III trial if the result is significant.

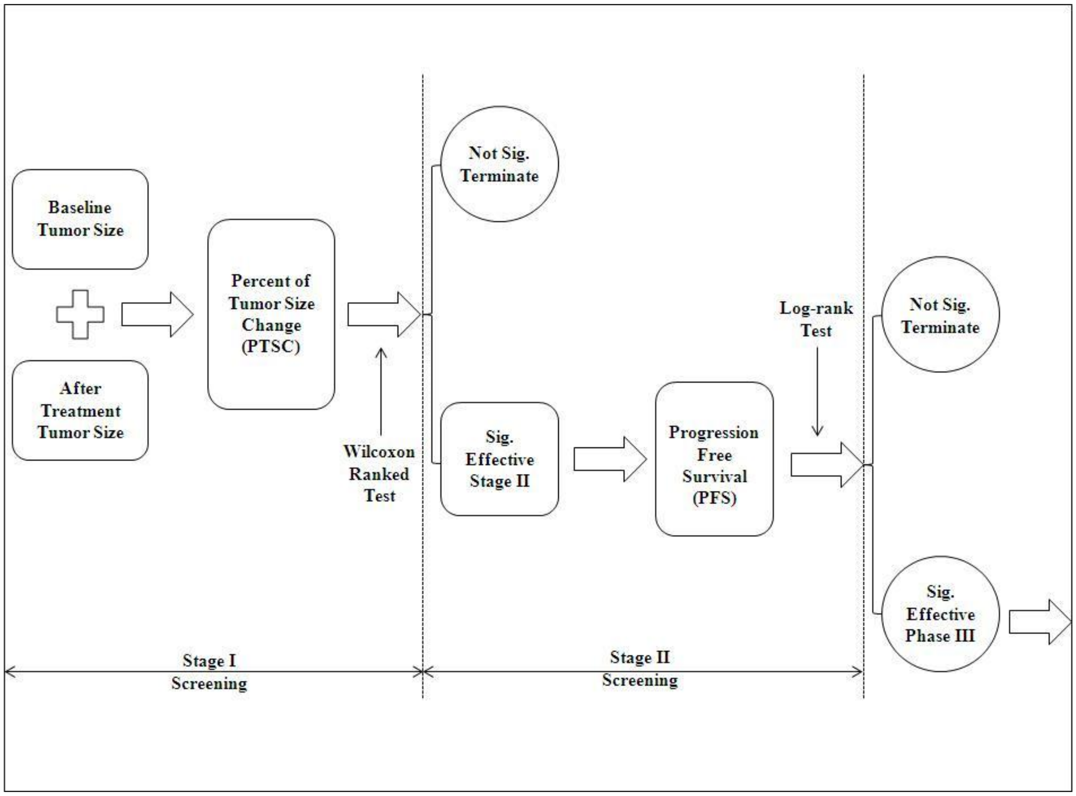

Figure 1: Two-Stage Phase II Design Schematic

2.2 Tumor Size Change Model

In clinical studies, exponential model is widely adopted to formulate tumor growth (Skipper H, 1982):

$$V_i(t) = V_i(0) \cdot \exp(\rho_i \cdot t) \tag{1}$$

Where $V_i(t)$ is the i^{th} patient's tumor size at time point t, $V_i(0)$ is baseline tumor size at the starting point of the patient, and ρ_i is the tumor growth rate. The exponential growth pattern is based on the assumption that no death or treatment intervention, and is considered to be appropriate (Sachs RK, 2001). To simulate data with external agent, Wang *et al.* proposed a mixed exponential-decay and linear-growth model (Wang Y, 2009):

$$V_i(t) = V_i(0) \cdot \exp(-\lambda_i t) + \rho_i \cdot t \tag{2}$$

The model includes treatment effect by an exponential tumor shrinkage rate λ and a linear tumor progression effect with rate ρ. We modified equation (2) to simulate our data with the considertation that the treatment effect applies to tumor's self-progression as well.

$$V_i(t) = (V_i(0) + \rho_i \cdot t) \cdot \exp(-\lambda_i t) \tag{3}$$

Continuous tumor size modeled with equation (3) would result in the tumor size asymptotically reducing toward zero. λ, the rate for tumor shrinkage and ρ, the progression rate are both restricted to be non-negative. Considering tumor growth kinetics, inidividual patient's tumor size is generated from exponential distribution (Friberg S, 1997). PTSC is calculated as:

$$p_i(t) = \frac{(V_i(t) - V_i(0))}{V_i(0)} \times 100 \tag{4}$$

2.3 Simulation Study and Design Performance

A simulation study is conducted to assess the performance of the two-stage design with respect to sensitivity and specificity. We generated 10,000 trials and compare sensitivity and specificity by two designs — conventional design, and our two-stage design. RR is generated via RECIST 1.1 criteria (Table 1), and PTSC is generated by tumor size model described above respectively as Stage I endpoints. Our two-stage design will adopt PFS as an endpoint in Stage II screening. OS is also simulated as an observation for each patient; log-rank test results to OS are assumed to be true results. We compared outcomes of both designs to the true results.

Table 1: RECIST 1.1 Criteria

Group	*Response Category*	*RECIST 1.1 Criteria*	*Example*
Objective Response	Complete Response (CR)	Disappearance of all target lesions. Any pathological lymph nodes (whether target or non-target) must have reduction in short axis to <10mm.	A 100 percent of decrease in tumor size is defined as CR.
	Partial Response (PR)	At least a 30% decrease in the sum of diameters of target lesions, taking as reference the baseline sum diameters.	A 45 percent of decrease in tumor size is defined as PR.
Non-objective Response	Progressive Disease (PD)	At least a 20% increase in the sum of diameters of target lesions, taking as reference the smallest sum on study (this includes the baseline sum if that is the smallest on study). In addition to the relative increase of 20%, the sum must also demonstrate an absolute increase of at least 5 mm. (Note: the appearance of one or more new lesions is also considered progression).	A 30 percent of increase in tumor size is defined as PD.
	Stable Disease (SD)	Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD, taking as reference the smallest sum diameters while on study.	Both 20 percent decrease in tumor size, and 10 percent increase in tumor size are identified as SD.

A Wilcoxon ranked test is conducted to compare the distribution in terms of PTSC, and chi-square test is used to assess the difference of RR in two treatment groups. The whole design is based on the idea that the time to progression of a patient is related with the tumor shrinkage. Therefore, PFS is generated according to the percent of change, which assumes 1% tumor size shrinkage will result in certain amount benefit of survival time. Log-rank test is performed to assess PFS. OS is simulated as the true observation as well; log-rank test result in OS between two groups is considered to be gold standard. Both conventional design and our two-stage design results are compared with the gold standard.

As it is shown in table 1, sensitivity and specificity of our two-stage design are superior to those of conventional design. 73.55% of treatments that are truly efficient are correctly recognized by our design, while only 36.36% evaluated as active using conventional design. On the other hand, 96.32% of treatments that are not sufficiently active are correctly identified by the two-stage design, whereas a mere of 68.29% are recognized if adopt the traditional design. That is, the ability of our design to correctly identify treatments with sufficient anticancer activities and the ability to correctly identify those with confounding outcomes exceeds conventional design. Moreover, positive predicted value of 74.17% means that 74.17% of treatments identified by two-stage design to be efficient are truly efficient; but only 14.15% of those hit the truth if using conventional design. Similarly, non-significant results with our design have a 96.20% chance of being truly inefficient; while the chance of conventional design is 88.19%. With a high NPV, our new design is considered doing as good as "gold standard".

Table 2: Comparison of Sensitivity and Specificity of Two Designs

		Conventional Design			*Two-Stage Design*		
		Estimated Value	*95% Confidence Interval*		*Estimated Value*	*95% Confidence Interval*	
SE	(%)	36.36	30.36	42.80	73.55	67.44	78.90
SP	(%)	68.29	66.00	70.50	96.32	95.27	97.14
PPV	(%)	14.15	11.56	17.19	74.17	68.06	79.48
NPV	(%)	88.19	86.28	89.87	96.20	95.15	97.04

3. Endpoints of our Two-Stage Design

3.1 Continuous vs. Categorical Endpoint

In clinical studies, categorization is quite commonly adopted by grouping continuous values into ≥ 2 categories, (Naggara O, 2011), not only limited to cancer Phase II trials. The primary reason according to the approach is the need to label patients with an attribute for diagnostic or therapeutic procedures determination (e.g. 'hypertensive', 'obese') (Royston P, 2006).

However, the disadvantage of grouping continuous variables is obvious. A serious loss of power (Lagakos SW, 1988) and higher sample size requirements (Wason JMS, 2011) in effectively detecting possible relationships is the cost of simplicity. At least one third, even higher proportion if the predictor is exponentially distributed of the data is discarded while dichotomizing (Lagakos SW, 1988). Moreover, concerns about clinical benefit make it more controversial nowadays in cancer trials. For example, a patient with 25% tumor shrinks has more clinical benefit than one with 10% tumor increases, but both are labeled by RECIST 1.1 as having stable disease (Karrison TG, 2007). In contrary, a patient is identified as an objective responder with 35% shrinkage may not has much difference with the one with 25% shrinkage, but the latter is not (Karrison TG, 2007). Therefore, it is not surprising that more and more researchers choose to use tumor size changes directly instead of RR.

3.2 Simulation Result

Table 3 lists the simulated *P*-values of the Phase II trials with RR method and PTSC method, respectively. Different combinations of mean percent of tumor size change are assigned to experiment and control groups. Tumor response categories are also listed in the table. According to RECIST 1.1 criteria, CR is defined as a completely disappear of tumor cells, which is only applied to a very small proportion of patients. So, we only compare those mean percent of tumor size change that belong to PR, SD, and PD categories.

Table 3: Comparison of Simulated P-value by Two Designs

Mean Percent of Tumor Size Change with RECIST Category		**Simulated P-values**	
		Conventional Design	**Two-Stage Design**
Placebo Group (μ_0)	**Experiment Group (μ_1)**	**Stage I**	**Stage I**
32.7% (PR)	45.2% (PR)	0.2616	0.0019
-5.0% (SD)	45.2% (PR)	2.42×10^{-11}	9.15×10^{-21}
-22.9% (PD)	45.3% (PR)	9.56×10^{-44}	1.97×10^{-39}
-5.0% (SD)	15.4% (SD)	0.0024	1.93×10^{-7}
-22.9% (PD)	15.4% (SD)	1.88×10^{-26}	1.45×10^{-28}
-27.6% (PD)	-23.0% (PD)	0.0015	0.0231

Table 3 shows that all of the median *P*-values in first screening using PTSC method are statistically significant. Majority of *P*-values with PTSC method are more significant than with RR method, which implies PTSC method is more sensitive at detecting tumor size change. It is interesting to notice that large discrepancies exists between PTSC and RR *P*-values when mean percent of changes are in the same or adjacent categories. Although they are in the same or adjacent categories, the mean percent of changes still have large differences that could be detected by measuring the changes directly. However, adopting dichotomized categories would ignore the natural spread of changes in tumor sizes. It is also noticeable that two of RR *P*-values are more significant than PTSC *P*-values. We notice that tumor shrinkage in one treatment group distributes far away from the category boundary (e.g. 45.3% is far above PR boundary 30% that will be categorized as Objective Response, and -27.6% is opposite will be categorized as Non-objective Response). So the more significant RR *P*-values might overestimate the difference by simply comparing objective response rate. Table 3 also illustrates the simulated Stage II *P*-value results. We found that the PTSC method provides very small improvement while testing the difference in progression-free time.

3.3 PFS as Surrogate Endpoint of OS

OS is the traditional and the objectively measured endpoint that is adopted to assess new cancer drugs. However, it requires prolonged follow-up and so may not be optimal for a fast assessment of therapeutic advances (Burzykowski T, 2004). Moreover, many clinical trials now include sequential therapies, and OS as a primary endpoint would not accurately reflect the effect of the investigational drug with multiple lines of treatment (Hotte SJ, 2011). So many researchers proposed surrogate endpoints for OS to evaluate the clinical benefits of new drugs in oncology, where PFS is frequently adopted. For instance, Gill *et al.* used PFS in clinical trials of metastatic colorectal cancer (Gill S, 2011), and Saad *et al.*reviewed PFS as a surrogate endpoint in breast and colorectal cancer treatment (Saad ED, 2010). In this paper, we will treat PFS as the endpoint of Stage II in our design as a predictor of OS.

4. Comparison of Success Rates of Two Designs

We will compare the two-stage Phase II trials with different measurements of tumor shrinkages (continuous measurement of tumor size change vs. grouped tumor response rate). Moreover, we will use the simulation result to discuss the choice of primary test criterion in the first screening stage.

Figure 2 captures the trend of simulated overall success rate in this Phase II trial (significant level of the second screening is set up as 0.05). It shows the success rate of the method that uses tumor size change (PTSC method) directly is superior to the method that uses response rates (RR method) in Stage I screening.

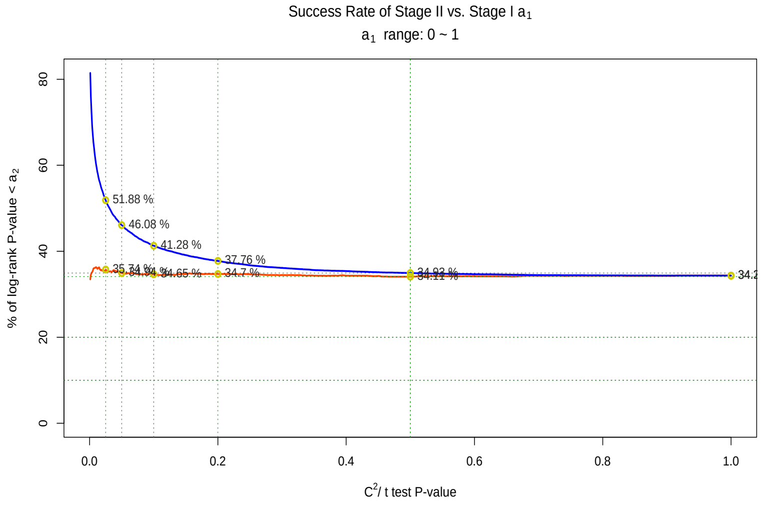

Figure 2: Success Rates of PTSC method and RR method

In this simulation, mean percent of changes in tumor size are assigned as 45.2% and 12.3% to experiment and control groups, respectively. Progression-free survivals are simulated according to tumor shrinkage with regard of PFS increment rate. In this scenario, both distributions are centered within the range of the same category — "SD", but their clinical benefits —

progression-free survival are quite different. Using RR method, however, may underestimate the clinical benefits; because information is lost during the process of categorizing data. In contrary, our design (PTSC method) which uses tumor size change directly is more sensitive in detecting the differences of tumor size related variables.

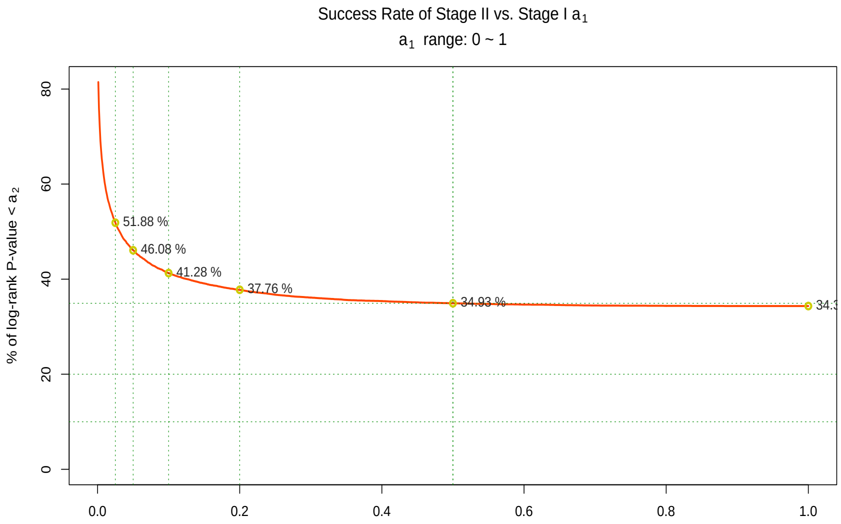

Figure 3: Success Rate of our Two-Stage Design

The main objective of our study for the new Phase II trial is to improve the success rate following Phase III trial. Previous researches (Burzykowski T, 2004; *et al.*) have confirmed PFS as a potential surrogate as OS. So the success rate in the simulation study could well reflect the success rate of the proceeding Phase III trial. A higher percent of success in the two-stage Phase II trial will definitely lead to a better chance of success in further study.

It is demonstrated by Figure 2 that the two-stage Phase II trial success rate changes according to the selection of Stage I significant levels. The half bath-tube shaped curve implies a very sharply decreasing at first and then very smooth trend for the success rate. From 0 to 0.025 the success rate declines more than 30%, while another 15% decrease happened between the range of 0.025 and 0.2. In contrary, the cumulative decrease is a mere 4% starting from 0.2 till 1.With this result, we would like to suggest a rigorous significant level in Stage I screening (e.g. $\alpha_1 \leq 0.05$). Although it is generally expected that a phase II trial with a lax criteria could increase the possibility of new drug discovery and avoid the omission from the phase II trial rejection, this assumption is not supported by our simulation result. A strict criterion in Stage I screening in our Phase II trial is considered to be more significant in practice, and tends to lead more satisfactory results in the further study.

5. Application

Anticancer drug development is an extreme costly and time-consuming process. The cost of launching a new drug in oncology ranges from $800 million to $2 billion, where the mean cost of a Phase III trial is the highest among all three phases (Phase I: $15.2 million, Phase II: $23.5 million, and Phase III: $86.3 million). But the probability of entering Phase III study from previous trials are merely 31.4%, which is relatively low comparing to 71.0% from Phase I to Phase II trials (DiMasi JA, 2003). Thus, it is crucial in clinical studies to enhance the success rate from the mid-stage studies to late-stage Phase III trials. Our proposed new Phase II design, which uses continuous tumor size change and PFS as endpoints in the double screening process, could effectively improves the success rate of the further trial. Based on the assumption that survival time is associated with tumor shrinkage, our design could detect the change in tumor size more sensitively. Early termination is determined with the assessment of Stage I result, which could reduce the length and cost of the trial. We also recommend a strict significant level to evaluate the tumor shrinkage. According to our simulation study, a promising result is more likely to be lead to in Stage II if the Stage I result passes the test with a strict significant level. So, setting up a strict criterion in the first screening stage could effectively reduce the cost by excluding trials unlikely to have significant result in further studies.

According to a new benchmarking report, Goldfarb pointed out the cost per patient of running phase III trial has exceeded $26,000, on average (Goldfarb NM, 2006). Assuming an average phase III trial size of 400 subjects (that is 200 subjects in each treatment arm) are involved in about 500 phase III clinical trials currently, it will cost approximate 5.2 billion. However, high failure rate (50-60%) (An WM, 2011) will cost at least 2.6 billion without returns. Thus, our proposed two-stage Phase II design that could effectively improve the success rate of Phase III trials has a lot practical meanings.

Reference

1. National Cancer Institute (http://www.cancer.gov)
2. Eisenhauer EA,Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumors: Revised RECIST guideline (version 1.1). *Eur. J. Cancer*2009; **45**: 228-247.
3. An MW, Mandrekar SJ, Branda ME, Hillman SL, Adjei AA, Pitot HC, Goldberg RM, Sargent DJ. Comparison of continuous versus categorical tumor measurement-based metrics to predict overall survival in cancer treatment trials. *Clin. Cancer Res.*2011; **17**: 6592-6599.
4. Pivot X, Thierry-Vuillemin A, Villanueva C, Bazan F. Response rate: a valuable signal or promising activity? *Cancer J.*2009; **15**: 361-365.
5. Lavin PT. An alternative model for the evaluation of antitumor activity. *Cancer Clin Trials*. 1981; **4**: 451-457.
6. Wang Y, Sung C, Dartois C, Ramchandani R, Booth BP, Rock E, Gobburu J. Elucidation of relationship between tumor size and survival in non-small-cell lung cancer patients can aid early decision making in clinical drug development. *Clin.Pharmacol*2009; **86**: 167-174.
7. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: application to a study of sorafenib and erlotinib in non-small-cell lung cancer. *J. Natl. Cancer Inst.*2007; **99**: 1455-1461.
8. Buyse M, Thirion P, Carlson RW, Burzykowski T, Molenberghs G, Piedbois P. Relation between tumor response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. Meta-Analysis Group in Cancer.2000
9. Halabi S, Vogelzang NJ, Ou SS, Owzar K, Archer L, Small EJ. Progression-free survival as a predictor of overall survival in men with castrate-resistant prostate cancer. *J. Clin. Oncol.* 2009; **27**:2766-2771.
10. Greg Y. Toward progression-free survival as a primary end point in advanced colorectal cancer. *J. Clin. Oncol.* 2007; **25**: 5153-5154.
11. Skipper H, Schabel F Jr. Quantitative and cytokinetic studies in experimental tumor system. *Cancer Med.* 1982; **2**: 636-648.
12. Sachs RK, Hlatky, Hahnfeldt P. Simple ODS models of tumor growth and anti-angiogenic or radiation treatment. *Math Comput Model*. 2001; **33**; 1297-1305.

13. Friberg S, Matton S. On the growth rates of human malignant tumors: implications for medical decision making. *J Surg Oncol.* 1997; **65**: 284-297.
14. Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Anaylsis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptred aneurysms. *Am. J. Neuroradiol* 2011; **32**: 437-440.
15. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statist. Med.* 2006; **25**: 127-141.
16. Lagakos SW. Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statist. Med.* 1988; **7**: 257-274.
17. Wason JMS, Mander AP. The choice of test in phase II cancer trials assessing continuous tumor shrinkage when complete responses are expected. *Stat Methods Med Res*. 2011; **0**:1-11.
18. Burzykowski T, Molenberghs G, Buyse M. The validation of surrogate end points by using data from randomized clinical trials: a case-study in advanced colorectal cancer. *J. R. Statist. Series A: Statistics in Society.* 2004; **167**: 103-124.
19. Hotte SJ, Bjarnason GA, Heng DYC, Jewett MAS, Kapoor A, Kollmannsberger C, Maroun J, Mayhew LA, North S, Reaume MN, Ruether JD, Soulieres D, Venner PM, Winquist EW, Wood L, Yong JHE. Progression-free survival as a clinical trial endpoint in advanced renal cell carcinoma. *Curr. Oncol.* 2011; **18**: S11-S19.
20. Gill S, Berry S, Biagi J, Butts C, Buyse M, Chen E, Jonker D, Marginean C, Samson B, Stewart J, Thirlwell M, Wong R, Maroun JA. Progression-free survival as a primary endpoint in clinical trials of metastatic colorectal cancer. *Curr. Oncol.*2011; **18**: S5-S10.
21. Saad ED, Katz A, Hoff PM, Buyse M. Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Ann. Oncol.*2010; **21**: 7-12.
22. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J. Health Econ.* 2003; **22**: 151-185.
23. Goldfarb NM. Clinical operations: accelerating trials, allocating resources and measuring performance. *J. Clin. Res. Best Practice.* 2006; **2** (12).
24. Weber WA. Assessing Tumor Response to Therapy*J. Nucl. Med* 2009; **50**: 1S-10S.