ENROLLMENT AND STOPPING RULES FOR MANAGING TOXICITY IN PHASE II ONCOLOGY TRIALS WITH DELAYED OUTCOME

Guochen Song

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Heath in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Anastasia Ivanova, PhD

Jason Fine, PhD

Michael Hudgens, PhD

Stergios Moschos, MD

Bahjat F. Qaqish, PhD

ABSTRACT

GUOCHEN SONG: Enrollment and Stopping Rules for Managing Toxicity in Phase II Oncology Trials with Delayed Outcome
(Under the direction of Dr. Anastasia Ivanova)

Stopping rules for toxicity are routinely used in phase II oncology trials. If the follow-up for toxicity is long, it is desirable to have a stopping rule that uses all toxicity information available, not only information from patients with full follow-up. Further, to prevent excessive toxicity in such trials, an enrollment rule is needed. The enrollment rule informs an investigator about the maximum number of patients that can be enrolled depending on the current enrollment and all available information about toxicity. We give recommendations on how to construct Bayesian and frequentist stopping and enrollment rules to monitor toxicity continuously in phase II oncology trials with a long follow-up.

ACKNOWLEDGMENTS

I would like to thank Dr. Anastasia Ivanova for her continuous encouragement and assistance during this research and thank her for introducing the clinical monitoring field to me and sharing her extensive knowledge within this field with me. I'm grateful to Dr. Olga Marchenko, my manager at Quintiles, for her support and encouragement in the past one year and half. I'm thankful to Dr. Moschos, co-author on one of the papers, from whom I learned a lot about how to communicate with professionals that are not statisticians, and to Mr. Samuel Middleton, who helped to improve the readability of the paper. I would like to thank Dr. Bajat Qaqish, who provided useful comments on the Bayesian approach, Dr. Michael Hudgens for making valuable suggestions on how to improve the document, and Dr. Jason Fine, who kindly found time for my final exam despite his grading obligations. My thanks are extended to Dr. Follmann at NCI with whom I had communication that helped in developing the Bayesian Chapter.

The research was supported in part by NIH grant RO1 CA120082-01A1. I thank Drs. Stergios Moschos, Thomas Shea and Matt Foster for the permission to use the LCCC studies as examples.

I'm indebted to my wife, Jennifer Song, who took care of the family while I was working on weekends and nights, and also thank her for correcting many of my grammar mistakes and listening to talks about statistics applied in clinical trials. I would like to

thank my children, Abigail and Zachary, who provided so much joyfulness in the past few years even though they slowed me down in the research process.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction

In a phase I oncology trial the maximum tolerated dose is estimated. The maximum tolerated dose is usually defined as the dose with the probability of dose limiting toxicity equal to the maximum tolerated level (often 0.2 or 0.25). We will refer to dose limiting toxicity as simply toxicity in the remainder of this document. Under most study protocols, a patient that experiences a dose limiting toxicity will be discontinued from the study; therefore, we assume that any one patient can only experience one toxicity (i.e., toxicity is a binary variable where 0 = no toxicity and 1 = toxicity). The efficacy of the estimated maximum tolerated dose is investigated in a phase II trial, usually a single arm study. Since phase I trials use a small sample size, the estimate of the maximum tolerated dose is imprecise and may result in a dose chosen for the phase II trial with a probability of toxicity much higher than the maximum tolerated level. For example, in a phase II study that uses Teniposide (VM26) to treat small cell lung cancer, 32 patients were enrolled and 30 were eligible for evaluation. The study reported 9 early deaths, 5 of which were caused by septicaemia and were attributed to study treatment (Cerny et al., 1988). The high rate of early death seen in this study was unexpected. In another phase II study where vinorelbine plus doxorubicin was used to treat relapsed small cell lung cancer, a total of 34 patients were planned (Johnson et al.,

2004). The study applied a Simon-two stage design (Simon, 1989) where the study would only continue if 1 out of the first 14 patients responded to treatment, and the study would conclude that the treatment does not worth further investigation if fewer than 4 responders at the end of the trial. The trial had 4 responses after 21 patients were registered and 15 patients were evaluable, but accrual was terminated because there were too many toxicities observed in these 15 patients: 11 grade IV neutropenia were reported (73%) and additional 2 patients experienced grade III neutropenia. If a stopping rule was in place, these trials would have stopped early and fewer patients would have been exposed to the toxic treatment.

A stopping rule alone does not prevent a trial from possibly observing too many toxicities. An enrollment rule should be used together with a stopping rule to prevent an excessive number of toxicities in phase II trials. In this research, we propose such strategies to improve the current practice of phase II oncology trials.

The document is organized as follows: in Chapter 1, we present literature review. In Chapter 2, we review continuous monitoring rules currently used in phase II designs and give recommendations on which ones are preferred. In Chapter 3, we develop the enrollment strategies and stopping rules using a frequentist method, and in Chapter 4, we propose a Bayesian enrollment strategy and stopping rule.

1.2 Frequentist stopping boundaries

Let θ be the probability of experiencing toxicity during the follow-up period of the study for a patient at the dose chosen for the study and let θ_0 be the probability of toxicity that considered being acceptable. One way to set up a stopping rule for toxicity is

to specify the shape of the stopping boundary and the probability of stopping the trial for a given value of θ. Ivanova, Qaqish and Schell (2005) examined two types of boundaries widely used in sequential analysis, the Pocock (1977) boundary and the O'Brian-Fleming (1979) boundary, and concluded that the Pocock boundary is the most suitable boundary to monitor toxicity in a phase II oncology trial as it allows stopping early with high probability and, therefore, reduces the expected number of toxicities in the trial.

Let *K* be the sample size planned for a phase II study. The Pocock boundary can be defined through a point-wise probability α_0, such that the trial is stopped if, at each interim analysis, the null hypothesis $\theta = \theta_0$ is rejected at level α_0 in favor of the one-sided alternative $\theta > \theta_0$. The value of α_0 is chosen so that the overall probability of stopping the trial, ϕ, is equal to a specified value, usually $\phi = 0.05$ if the probability of toxicity is equal to the acceptable rate θ_0. We refer to the boundary that allows stopping the study at any point as a continuous boundary, because monitoring for toxicity is done throughout the trial on a continuous basis. Such a boundary can be described through the sequence of integers (b_1, b_2, …, b_K), which can be computed using step-wise significance level α_0. The constants b_k, k = 1,…,K, are equal to the smallest integer such that $P[Y \geq b_k] \leq \alpha_0$, where Y denotes a binomial random variable with parameters k and θ_0. If the number of toxicities in the first k patients is equal to or higher than b_k, the trial is stopped. Another way of implementing this boundary is to compute a one-sided p-value to test the null hypothesis $\theta = \theta_0$ versus a one-sided alternative $\theta > \theta_0$ after each patient is enrolled and their outcome is observed. The trial is stopped if the p-value is less than α_0. It is sufficient to apply the boundary only when a patient has toxicity. The value of

α_0 and the corresponding boundary can be calculated at http://cancer.unc.edu/biostatistics/program/ivanova/.

The Pocock boundary is a boundary that is constant in the standardized test statistic *Z*. The O'Brien-Fleming boundary is constant in the B-value defined as $Z\sqrt{k/K}$, where *k* is the number of patients enrolled at the time of the analysis. If the test statistic is a random variable with normal distribution, the step-wise significance level for O'Brien-Fleming boundary at each stage can be calculated as $\alpha_k = P(Z > C_B(K,\phi)\sqrt{K/k})$, where $C_B(K,\phi)$ is tabulated value that can be found in Jennison and Turnbull (2000). As $C_B(K,\phi)$ is a positive number, clearly α_k increases in *k*. For discrete cases, the O'Brien-Fleming boundary can be computed through trial and error, using $C_B(K,\phi)$ obtained from the normal approximation as the initial value.

The performance of the boundary is characterized by the probability of stopping the trial for a given probability of toxicity, the expected number of toxicities and the expected number of patients enrolled. For a binomial outcome, the exact calculation can be done as in Schultz et al. (1973), where the probability of stopping was calculated using a recursive formula by observing that to stop at stage *k*, the *k*th patient has to experience toxicity and there has to be exactly b_{k-1} toxicities before stage *k*. Given the probability of stopping at each stage, ϕ_k for *k*=1,2,…*K*, the expected enrollment can be calculated as $E(N) = \sum_{k=1}^{K} \phi_k \times k + K(1-\phi)$, and the expected number of toxicities is calculated as $E(Y)=E(N)\times\theta$.

In general, for each given K, ϕ and θ_0, a trial using the O'Brien-Fleming boundary stops with a higher probability, but later in the trial compared to a trial that uses

the Pocock boundary. The Pocock boundary results in fewer expected toxicities on average compared to the O'Brien-Fleming boundary and, therefore, is preferred for monitoring toxicity or adverse events.

The practice of establishing stopping boundaries to stop experiments can be traced back to the 1940s in Wald's work of sequential probability ratio test (SPRT) (Wald, 1945). Assume the likelihood function under the null hypothesis is $P_0(x_k)$ where x_k is data observed at stage k and the likelihood function under the alternative hypothesis is $P_1(x_k)$, the log likelihood ratio test is

$$\log(\Lambda_k) = \sum_{i=1}^{k} \left(\log P_1(x_i) - \log P_0(x_i) \right).$$

The stopping boundary can be expressed as (γ_0, γ_1): if $\log(\Lambda_k) < \gamma_0$, the experiment stops and accepts the null hypothesis as correct, or if $\log(\Lambda_k) > \gamma_1$, the experiment stops and accepts the alternative hypothesis as correct. Otherwise, the experiment keeps going until either γ_0 or γ_1 is crossed. Denote the type I error rate of the SPRT procedure as α, that is, the probability of the procedure accepting P_1 when P_0 is true and β as the type II error, the probability of the procedure accepting P_0 when P_1 is true. The procedure yields power of $1-\beta$. Note that the SPRT contains two one-sided tests at each stage, and the type I and type II errors of the whole procedure are gained through summarizing the whole sampling space. It is an open ended test with no sample size pre-specified. The values (γ_0, γ_1) can be computed for any given type I error rate and power. As the SPRT is an open ended test, it cannot be directly applied to clinical trials. Armitage (1957) proposed methods to restrict the sequential procedures to a pre-specified sample size while retaining the type I error and power. Because the SPRT

boundaries are for both lower and upper boundaries, that is, to stop if the toxicity rate is too low or too high, it cannot be applied to toxicity monitoring as the trial would not stop if the current data deems the toxicity is acceptably low. Goldman (1987) proposed using the upper bound of SPRT to stop for adverse events in trials where an adverse event is a binary outcome. Let θ_A be the probability of toxicity that is not acceptable. Given a current sample size *k*, the expression for the upper boundary for stopping is

$$e = \frac{\ln(1-\beta) - \ln\alpha - k[\ln(1-\theta_A) - \ln(1-\theta_0)]}{(\ln\theta_A - \ln\theta_0) - [\ln(1-\theta_A) - \ln(1-\theta_0)]}.$$

The parameters α and β are nominal parameters for the two one-sided, open ended SPRT. When used with the fixed maximum sample size and only using the upper boundary, α and β might have to be adjusted, through trial and error, to yield desired probabilities. For example, one might want the true type I error rate of the procedure to be $\phi = 0.05$.

As efficacy endpoint is usually the primary endpoint of phase II oncology trials, in multi-stage designs, the trial usually stops when there is evidence that the treatment is inactive. Bryant and Day (1995) proposed to monitor response and safety jointly in two-stage or multiple-stage designs. Let π be the probability of response of the experimental treatment. In this strategy, the null hypothesis is that the treatment is not safe or effective, specifically, the probability of toxicity is θ_0^* or higher **OR** the probability of response is π_0 or lower. The alternative hypothesis is that the probability of toxicity is θ_A^* and the probability of response is π_A, where $\theta_A^* < \theta_0^*$ and $\pi_A > \pi_0$. Note that the null hypothesis will be rejected if the treatment is safe and effective, hence θ_0^* is different from the

tolerated probability of toxicity, θ_0, in toxicity monitoring strategies described above. The study result variables were represented by $\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22})$ as defined below:

	Toxicity	No Toxicity	
Response	X_{11}	X_{12}	X_r
No Response	X_{21}	X_{22}	$X_{\bar{r}}$
	X_t	$X_{\bar{t}}$	

The corresponding probability is denoted as (p_{11}, p_{12}, p_{21}, p_{22}). As toxicity and response are bivariate binomial variables, an association variable is needed to fully specify the distribution. The odds ratio between response and toxicity $\varphi = p_{11}p_{22} / (p_{21}p_{12})$ was used. The stopping boundary is a pair of numbers (c_r, c_t) at stage k, and the trial stops if either more than c_t toxicities or less than c_r responses observed out of the number of patients at stage k. As this rule was usually applied for two stage or three stage designs, it is not a continuous monitoring rule and hence the number of patients at stage k is larger than k. The stopping rule was gained through enumerating all possible pairs of (x_r, x_t) and the pair (x_r, x_t) that satisfies the following situations can be set as the boundary: the probability of recommending an ineffective but safe treatment, $P[X_r \geq x_r, X_t \leq x_t \mid \pi \leq \pi_0, \theta = \theta_A^*, \varphi)$ is smaller than α_R, the probability of recommending an effective but toxic drug, i.e., $P[X_r \geq x_r, X_t \leq x_t \mid \pi = \pi_A, \theta \geq \theta_0^*)$ is smaller than α_T, and the power, $P[X_r \geq x_r, X_t \leq x_t \mid \pi = \pi_A, \theta = \theta_A^*)$ is at least $1-\beta$. The expected maximum sample size when the treatment is either toxic or ineffective can be

calculated for the given α_R, α_T, $1-\beta$ by maximizing over φ, and the boundary that gives the smallest such sample size is chosen as optimal.

Conaway and Petroni (1995) developed a similar method, but instead of controlling α_R and α_T, they proposed to control both the type I error rate over the whole null hypothesis region and the type I error rate at the point null hypothesis. Tournoux et al. (2007) compared the two methods and recommended to use Bryant and Day because it is more flexible. Jin (2007) proposed a method to control the type I error of toxicity and type I error of response separately.

Ray and Rai (2011, 2013) examined to apply the continuous monitoring Pocock boundary (Ivanova et al., 2005) with the Simon's two-stage (Simon, 1989) design. As the correlation between toxicity and response was ignored at the design stage, the authors found from simulation studies that the procedure was unexpectedly conservative, i.e., trials were stopped more often than desired, when the correlation between toxicity and response is high.

1.3 Bayesian stopping boundaries

Geller et al. (2005) proposed a Bayesian stopping rule for continuous monitoring of toxicity. Let Y denote the number of toxicity events with $Y \sim binomial(n,\theta)$, where θ is the probability of toxicity and n is the number of evaluable patients. Instead of treating θ as a fixed parameter that needs to be estimated, the Bayesian approach assumes that θ follows a Beta distribution. Before the trial, the prior distribution is defined as $\theta \sim Beta(a,b)$. The parameters a and b reflect prior information about θ. It is roughly

equivalent to data from $a+b$ patients with mean toxicity $a/(a+b)$, that is, if a prior patients had toxicity and b prior patients completed the trial without toxicity. Note that neither a nor b has to be an integer. The posterior distribution of θ follows a Beta distribution $Beta(a+x, b+K-x)$, where x is the number of patients in the current trial with toxicity and $K-x$ is the number of patients who have completed the current trial without toxicity. The posterior mean of the toxicity rate is

$$\mathrm{E}(\theta) = \frac{(a+x)}{(a+b+K)}.$$

From the posterior distribution, the probability of the θ exceeding a critical value θ_0 can be computed as follows:

$$P(\theta > \theta_0 | Data) = \int_{\theta_0}^{1} Beta\{\theta \mid a+x, b+K-x\} d\theta.$$

The stopping rule is constructed such that if this posterior probability is larger than a cut-off, the trial stops. Note that the quantity on the right hand side of the formula increases as x increases. For any given n, one can find the smallest x that satisfies the above equation for a pre-specified cutoff value (e.g, 95%). The sequence of such values of x forms the monitoring boundary. In a peripheral blood stem cell trial where 28 patients were planned (Geller et al., 2005), the rate of transplant related mortality (TRM) was monitored using this rule. They set $\theta_0 = 0.2$ and the cutoff for posterior probability as 0.9, i.e., the trial stops if the posterior probability that the TRM rate is above 0.2 is greater than 90%. The prior was chosen to worth 6 patients with the mean TRM rate as 0.2, i.e., the Beta (1.2, 4.8) prior was used.

The operational characteristics of the stopping rule can also be calculated based on Schultz et al. (1973). For example, the probability of stopping is 15.5% and the expected toxicity is 5.17 when the true toxicity rate is 0.2 using this boundary. This Bayesian approach generates a boundary very similar to the Pocock boundary when a non-informative or weakly informative prior is used and the type I error rate is controlled at the similar level.

Etzioni and Pepe (1994) developed a stopping rule to monitor two specific adverse outcomes at the same time and defined excessive toxicity as either type of adverse outcome exceeds a predefined level. In their example of a marrow transplant study, the dose limiting toxicity was defined as either non-engraftment or relapse. Tolerable probability of non-engraftment, θ_1, was set at a_1 = 30% and the probability of relapse, θ_2, at a_2 = 50%. The adverse outcomes were assumed to be independent and each follows a binomial distribution, given the probability of toxicity θ_1 and θ_2. A piecewise uniform prior were placed on probability of toxicity θ_1 and θ_2: the prior density is $1/(2a_1a_2)$ if $\theta_1 \le a_1$ and $\theta_2 \le a_2$, and $1/(1/2(1-a_1a_2))$ otherwise. The joint posterior distribution is a product of Beta densities and hence traceable. The posterior probability of excessive toxicity can be used to monitor excessive toxicities for any given number of patients, the boundary thus is a continuous monitoring boundary.

Instead of controlling the probability of stopping when the probability of toxicity is tolerable, Yu et al. (2012) proposed to stop the trial when there is evidence that the probability of toxicity is at a certain pre-specified high level θ_1. They argued that methods based on rigorous hypothesis testing setting, such as Pocock or O'Brian-Fleming

boundary, are not suitable. They proposed a group sequential strategy to control the incidence rate of toxicity in the study at a certain level. The frequentist boundary at stage k is the largest s_k that satisfies $P\left(X_k > N\theta_1 - s_k \mid \theta\right) > \xi$, where N is the total number of patients in the trial, ξ is a predetermined constant (suggested to be 0.5 by the authors), $s_k \le n_i$ is a non-negative integer and n_i is the number of patients in stage i, X_k ~Bin($N - \sum_{i=1}^{k} n_i$, θ) is a binomial random variable. In the case previous information is available, they proposed a Bayesian variation of this method by putting a Beta prior on θ. Because existing information is used, the Bayesian version on average enrolls much less patients if the true toxicity rate is high, hence is more effective. The advantage of the proposed boundary are not clear. For example, we compared a boundary from Yu et al. (2012) with θ_1 = 0.3, with the Pocock boundary constructed as described in Ivanova et al. (2005). The Pocock boundary yields better operational characteristics than the boundary in Yu et al. (2012).

For normally distributed outcomes, Freedman and Spiegelhalter (1989) considered trials with up to 5 stages and showed that a Bayesian stopping boundary can have a shape similar to the O'Brien-Fleming or Pocock boundaries depending on the prior, with Pocock boundaries arising from non-informative or slightly informative priors.

1.4 Stopping boundary in trials with delayed outcomes

In many clinical trials in oncology, the outcome is defined as a binary variable that takes a value of 1 if an event is observed during an observation period (0, t^*), and 0

otherwise, where t^* is the follow-up time for toxicity. When t^* is long, it is desirable to make intermediate decisions in the trial based on all the data, including data from patients still under follow-up. Such methods have been developed for phase I oncology trials with long follow-up (Cheung and Chappell, 2000; Ivanova et al., 2007; Bekele et al., 2008).

For example, Cheung and Chappell (2000) developed the time-to-event continual reassessment method (TITE-CRM) in phase I dose finding studies to prevent patients from being assigned to a dose with an unexpected high toxicity. Let θ be the probability of experiencing toxicity by the end of time t^*, and u be the time of follow-up for a patient under observation, $u < t^*$. The contribution of information to likelihood function from this patient can be expressed as a truncated probability, $w(u,T)\theta$, where $w(u,T)$ is a weight function. An obvious choice of such weight function is $w = u / t^*$. Other parameters besides u and t^* can be added to the weight function, for example, the level of dose and dose response curve in a dose finding study. An adaptive weight function may utilize the current number and time of toxicities observed in the study. A similar approach is to use the Kaplan-Meier estimator to define the weight function, which is discussed by Yin (2012).

Bekele et al. (2008) extended the TITE-CRM by calculating the predictive probability of negligible toxicity and excessive toxicity for the current dose. They assumed a latent variable and the probability that the toxicity rate exceeds or is lower than a certain level was calculated through Markov Chain Monte Carlo (MCMC). Using their rule, a trial may stop, escalate to a higher dose level, de-escalate to a lower dose level or simply wait until more information is available.

Following Antonick (1974) and Blum and Sursala (1977), Follmann and Albert (1999) used all information available in interim analysis using a Dirichlet-multinomial model. They divided the interval $(0, t^*)$ into M intervals and tabulated information about adverse events in each of the intervals, with patients completed the trial falls to the interval $M+1$. To make the interim decision, they enumerate all possible outcomes for each interval and compute their probabilities. With a $Dir(\boldsymbol{\alpha})$ prior, the posterior distribution given the observed enrollment profile will be a mixture of Dirichlet distributions for all possible trial results and the probability of the realization for each possible result can be calculated as weight for this distribution. From the mixture of Dirichlet distributions, the posterior distribution of θ can be expressed as a mixture of Beta distributions by summarizing the first M elements of the mixture of the Dirichlet distribution. The posterior probability that θ exceeds θ_0 can be calculated from this posterior distribution. When the number of patients in follow-up gets large, this calculation requires large computing resources and Follmann et al. (1999) proposed a data augmentation method to resolve this problem based on Tanner (1992).

Under the Dirichlet-multinomial framework, Rosner (2005) applied Gibbs sampling for monitoring clinical trials comparing two survival curves.

1.5 Enrollment rules

If many patients are enrolled at once, a stopping rule alone will not prevent from treating too many patients with a regimen that may not be safe. Often, many patients are enrolled at the very beginning of the trial which might lead to excessive toxicities. An enrollment rule informs investigators about how many patients may be enrolled at the

beginning of the trial and guides further accrual based on the information about toxicity in the trial.

Enrollment strategies are often used in phase I dose finding studies to avoid exposing too many patients to potentially toxic compound. For example, in the traditional or the 3 + 3 design, at most 3 patients are enrolled at a time.

In the context of sequential planning of experiments, Schmegner and Baron (2004) defined a risk function that consists of three components: a loss function, a cost from each individual and a fixed cost for each batch. They defined an optimal sequential plan as a plan that minimizes the expected risk. For a trial that enrolls patients in K batches, if the trial enrolls n_k patients at a time, and the cost for enrolling each patient is c_i but there is a fixed cost c_b for each batch, the risk function is:

$$R = E(L(\theta, \{b_k\}) + \sum_{k=1}^{K} (c_i n_k + c_b),$$

where $L(\theta, \{b_k\})$ is the loss function determined by the true toxicity rate and the stopping rule. A pure sequential plan is a plan in which only one patient is allowed at each batch. Clearly, given the true toxicity rate and the stopping rule, the pure sequential plan yields the smallest loss, $L(\theta, \{b_k\})$, comparing to any other plans that allow more than one patients at a time. However, a plan that yields the same loss with less expected risk is possible. Schmegner and Baron (2004) called such plans conservative plans. Take a trial where a total of 6 patients can be enrolled as an example, with a simple stopping rule that the trial will be stopped if 3 or more toxicities are observed. Instead of enrolling patients one by one, this trial can enroll 3 patients at once in the beginning, and if there is no toxicity observed after the first 3 patients, the trial can enroll another 3 patients; if there is

1 toxicity observed after the first 3 patients, the trial can enroll 2 more patients. This enrollment strategy will not result in an increase in the maximum number of toxicities observed in the trial compared to a pure sequential trial and hence is conservative. The goal of the method of Schemegner and Baron (2004) was to reduce sampling costs without significantly sacrificing the accuracy of statistical results. Similar ideas can be used to construct enrollment rules in trials with possible high rate of toxicity to reduce the time of the trial without significantly sacrificing the safety of patients.

CHAPTER 2

SHOULD PHASE II TRIALS ROUTINELY REQUIRE STOPPING RULES FOR TOXICITY?

2.1 Overview

Phase I trials routinely assess safety of new agents. However, given the relatively small number of patients that is usually accrued in phase I trials, the maximum tolerated dose (MTD) of the treatment regimen can be imprecisely defined. This may lead to recommendations for testing new agents in phase II trials at doses that could lead to excessive toxicity. Rigorous stopping rules for toxicity are not currently a standard feature of phase II trials. We propose that phase II trials should systematically assess toxicity in addition to their inherent role to provide preliminary evaluation of efficacy for new agents. We review various toxicity monitoring rules used in phase II oncology trials to stop the trial earlier if the dose limiting toxicity rate is higher than expected from the phase I assessment. We also provide recommendations on which rules we favor to use. We propose that toxicity should be monitored on a continuous basis, that is, after outcome of each new patient is observed, as opposed to *a priori* defined multi-staged rules. Continuous monitoring allows early stopping of trial if toxicity is high, and reduces the expected number of toxicities observed in the trial. In particular, we recommend the

frequentist Pocock stopping rule and the Bayesian rule, both of which allow stopping the trial early if toxicity rate is high.

2.2 Introduction

Over the last 30 years a significant number of novel methodologies have been developed to diversify the classical, single-arm, frequentist, single-stage design of the phase II clinical study, namely to gain preliminary insights in the clinical activity of an agent or treatment combination and to 'weed out' ineffective drugs from further, usually more costly clinical development. Despite the development of novel multi-arm (Rubinstein et al., 2009) or combined phase I-II trial designs (Hoering LeBlanc and Crowley, 2011), most contemporary phase II trials remain single-arm in design and rely on information from phase I studies (El-Maraghi and Eisenhauer, 2008). Novel designs for such single-arm phase II trials include various types of stopping rules for lack of efficacy, ranging from multi-stage enrollments using traditional frequentist approaches (Simon, 1989, Ensign et al., 1994) to Bayesian methodologies that take into account accumulated information from prior experience (prior) as well as data collected (likelihood function) to update and/or adapt the design (Berry, 2012). Such designs have traditionally maintained a single primary outcome, efficacy, whereas further definition of toxicity, which was roughly defined in a relevant previous phase I trial, had been a secondary endpoint. Since efficacy of a particular treatment regimen can be positively associated with toxicity, it may be equally important within the context of a phase II clinical trial not only to adopt stopping rules that address lack of efficacy but also consider stopping rules that penalize excessive toxicity.

To date, 'unacceptable' toxicity occasionally seen in phase II trials is commonly managed by frequent (>50%) dose reduction or refusal for further treatment. This approach raises the question whether sufficient statistical power remains to assess if a given dose of the investigational agent for which the trial was specifically designed is actually effective or whether any lack of efficacy is secondary to frequent dose reductions due to toxicity (Maki et al., 2009, Alberts et al., 2012, D'Adamo et al., 2005, Wyman et al., 2006). Frequentist and Bayesian methods have been developed to evaluate both toxicity and efficacy as bivariate (efficacy, safety) variable. Most of the methods are two-stage and range from equal weighing of response and toxicity to designs with variable trade-offs between these two outcomes (Jin 2007, Thall and Cheng, 1999, Conaway and Petroni, 1995, Bryant and Day, 1995, Tournoux et al., 2007). Detecting excessive toxicity early is paramount; evaluating toxicity formally only once during the trial is not sufficient.

On some occasions serious adverse events (grade ≥4) occur (Dawson et al., 2008, Gibson et al., 2010). An interesting example of grade 5 toxicity (e.g. death) that was observed in a high risk patient population is a phase II study conducted through the Eastern Cooperative Oncology Group (E2205) of preoperative administration of systemic anticancer therapy that consisted of oxaliplatin, infusional 5-fluorouracil, and cetuximab concurrent with external beam irradiation prior to definitive esophagectomy in patients with operable esophageal adenocarcinoma (Gibson et al., 2010). Standard multimodality therapy (the regimen without cetuximab) is associated with a known significant adverse event profile. A Simon two-stage design (Simon, 1989) with near doubling the rate of complete pathologic response was the primary efficacy endpoint. Although toxicity was

an important secondary endpoint and was monitored closely in real-time**,** no particular stopping rules for toxicity were incorporated into the study design. Six treatment-related deaths were observed in the initial cohort of 22 patients and therefore the trial was stopped after stage 1. Even in the face of a known high-risk patient population and treatment regimen, it became clear that the trial exhibited an excessive adverse event rate, even though there also appeared to be an emerging strong efficacy signal and pathologic complete response were seen among multiple patients in the initial cohort, including among patients who later developed grade 5 events. In this trial, a formal stopping rule for adverse event might have assisted the study team in weighing the risk/benefit of the therapy.

A formal rule should be in place to allow stopping the trial at any point during the trial conduct should toxicity rate be unacceptably high. Does observing 3 treatment-related deaths in the first 5 patients warrant stopping the trial? A question like this might be raised by the data and safety monitoring committee for the study if a formal stopping rule is not pre**-**specified within the protocol. If such a rule is continuous, i.e. monitored throughout the study, it will specify the minimum number of toxicities that warrant stopping the study out of the total number of patients who have already been enrolled and have mature toxicity data, also known as the stopping boundary. For example, if the investigators anticipate to 'reasonably' observe approximately 5% of deaths within the study (e.g. allogeneic bone marrow transplantation with high probability of life-threatening graft-versus-host disease), a potential formal stopping rule (Pocock, 1977) would recommend trial suspension if 2 grade 5 events are observed in the first 2-4 patients, 3 events observed in 5-12 patients, 4 events in 13-21 patients, 5 events in 22-31

patients, and if 6 events or more are observed in more than 31 patients. Stopping rules can also be incorporated into more novel phase II designs, such as the phase IB portion of the combined phase I-II trials to monitor extended cohort of patients assigned to the estimated MTD, in which case the stopping rule does not have to be as conservative as the ones applied for a phase II study.

Unfortunately, a rigorous stopping rule for toxicity is not a standard feature of most current phase II study designs. In support of this notion, we performed a Medline search to identify articles that focus on phase II oncology trials. We limited our search to articles published in the following five journals: American Journal of Clinical Oncology, Annals of Oncology, British Journal of Cancer, European Journal of Cancer, and Journal of Clinical Oncology. We also focused our search on articles published in 2005 and in 2010 to assess if recently developed dual statistical designs for both efficacy and toxicity were being incorporated in the more recently reported (2010) as opposed to more remotely reported (2005) clinical trials. A total of 255 phase II trials published in 2005 and 204 trials published in 2010 were reviewed. While all publications either summarized toxicities by type (e.g., neurologic, immunogenic-related adverse events) or listed most common toxicities by number and percent of patients, none of the articles described a pre-specified formal stopping rule for toxicity. In this article, we review toxicity stopping rules we found in phase II trial protocols. We group these rules in two categories: rules that are effective in protecting from excessive toxicity by stopping the trial early and rules that fail to detect excessive toxicity early during the trial, and therefore should not be used.

2.3 Bayesian stopping rule

Consider an example of a phase II trial with a total of $K = 20$ patients. This can be a single-stage trial or a two-stage trial where, for example, Simon's two-stage design (Simon, 1989) is used to test the efficacy endpoint. To set up a stopping rule for excessive dose-limiting toxicity (DLT), one needs to specify an acceptable DLT rate, θ_0. Usually, the acceptable DLT rate θ_0 is the rate of toxicities seen at the MTD in the corresponding phase I trial assuming that each patient can develop only a single DLT and that each patient has completed the study, i.e. complete follow-up exists. For example, the rate at the dose chosen by the 3 + 3 design (Storer, 1989) is 0.20 or 0.25 (Reiner, Paoletti and O'Quigley, 1999).

We describe stopping rules based on both the traditional frequentist and the Bayesian statistics. Frequentists use fixed parameters to describe the unknown state of truth. For example, we can use θ to describe the true DLT rate, and assume that there is a true value for θ, e.g., 0.2 or 0.5. Bayesians, on the other hand, describe the unknowns with a certain degree of uncertainty. For example, before data from the phase II trial are available, one may rely on data from a previous phase I trial and assume that the mean DLT rate is near 0.2 and there is a 46% probability that the DLT rate is larger than 0.2. Such assumption is called the 'prior distribution', which in this example can be expressed as Beta(4, 16), where Beta is a statistical distribution defined on (0, 1). Beta(4, 16) can be viewed as reflecting the prior information from 20 patients that were, enrolled in a prior phase I trial, 4 of whom experienced DLTs and 16 patients did not. As data are being collected from the ongoing phase II trial, the 'posterior distribution' is computed, which combines the prior (phase I) and new (ongoing phase II) data. For example, if 5 DLTs are

observed and 25 patients have completed the trial without DLT, the posterior distribution of the DLT rate is Beta(4 + 5, 16 + 25) with corresponding mean DLT rate of 0.18. The probability that the DLT rate is larger than 0.2 is now 33%. On the other hand, if 10 DLTs are observed among these 30 patients, the posterior distribution of the DLT rate is Beta(4 + 10, 16 + 20). Therefore the probability that the DLT rate is larger than 0.2 is 90% and the DLT rate is estimated as 0.28.

Geller et al. (2005) proposed a Bayesian stopping rule for continuous monitoring of toxicity. The trial is stopped if the posterior probability of the DLT rate exceeding θ_0 is equal to or higher than a pre-specified value τ. The value of τ is often chosen based on tradition, e.g., 0.95 or 0.98 is commonly seen.

Lines 1 and 2 of Table 2.1 provide two Bayesian stopping boundaries for a trial of 20 patients and τ = 0.98 for θ_0 = 0.2. A stopping boundary is described by a set of integers $b_1,\ldots, b_K$ such that the trial is stopped if there are b_k or more DLTs observed out of first k patients. The prior distribution, the value of tolerable DLT rate θ_0 and the value of τ uniquely define the set of integers $b_1,\ldots, b_K$ that can be computed before the trial. To use the Bayesian boundary there is no need to compute the probability that the DLT rate is larger than θ_0 = 0.2 given current data, instead one can just check if the number of observed DLTs in the first k patients is equal to or exceeds b_k.

The boundary in line 1 uses prior Beta(0.6, 2.4) reflecting information from the total of 0.6 + 2.4 = 3 patients. The prior might reflect information from a 3-patient dose cohort of phase I trial. Here limited prior information is probably used because the phase I population is different from the phase II population, or the length of follow-up for toxicity in the phase II trial is different from phase I. In this example, the overall

probability of stopping the trial when the DLT rate is equal to the acceptable rate of 0.2 is 0.038. The boundary in line 2 of Table 2.1 uses the prior Beta(4, 16), which reflects information from 20 patients with observed DLT rate of 4/20 = 0.20. Since we have strong prior information that the DLT rate is close to 0.20, stronger evidence is needed in the phase II trial that the DLT rate is high to stop the trial compared to the first boundary. This also is reflected in the overall probability of stopping the trial when the DLT rate is equal to the acceptable rate of 0.2, and this probability is very small and is equal to 0.004.

Another way to set-up a Bayesian boundary is, instead of specifying τ, to specify the overall probability of stopping when the toxicity rate is acceptable, which is the frequentist type I error rate. Line 3 of Table 2.1 shows a boundary with prior Beta(4, 16) where the overall probability of stopping is fixed at 0.05 when the true toxicity rate is 0.2. This probability of stopping is achieved when τ = 0.911 is used. The value of τ is computed by trial and error for a given overall probability of stopping. Defining the boundary based on the overall probability of stopping when the toxicity rate is acceptable, the type I error rate, is not a Bayesian approach but rather a frequentist concept. Nonetheless, this approach is often used because preserving the type I error rate is usually important whether the approach used is Bayesian or frequentist.

Table 2.1. Stopping boundaries for a trial with 20 patients in a trial with acceptable DLT rate of $\theta_0 = 0.2$. The trial is stopped after k patients if the number of observed DLTs in equal to a higher than the corresponding value of the boundary.

Line	Number of patients, k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Bayesian Boundary with prior Beta(0.6,2.4), $\tau = 0.98$	-	-	3	4	4	5	5	5	6	6	6	7	7	7	7	8	8	8	9	9
2	Bayesian Boundary with prior Beta(4,16), $\tau = 0.98$	-	-	-	-	-	6	7	7	7	7	8	8	8	9	9	9	9	10	10	10
3	Bayesian Boundary with prior Beta(4,16), $\tau = 0.91$	-	-	-	4	5	5	5	5	6	6	6	6	7	7	7	7	8	8	8	8
4	Pocock Boundary, type I error rate is 0.05	-	-	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9
5	O'Brien-Fleming Boundary, type I error rate is 0.05	-	-	-	-	-	6	6	6	6	6	6	7	7	7	7	7	7	8	8	8

2.4 The Pocock versus O'Brien Fleming stopping boundaries

Over the years several frequentist sequential boundaries have been developed to use in group-sequential trails to stop early for efficacy. The two most frequently used boundaries are the O'Brien-Fleming (1979) and the Pocock (1977) boundaries (Table 2.1). The O'Brien-Fleming boundary achieves the higher power compared to the Pocock

boundary for a given sample size and type I error rate. That is, when used for sequential monitoring of efficacy, the O'Brien-Fleming boundary yields higher probability of declaring the treatment is efficacious when the treatment is indeed effective compared to the Pocock boundary, which is the reason it is used more often. Similarly, when occasionally used to monitor toxicity, the O'Brien-Fleming boundary yields the higher overall probability of stopping the trial compared to the Pocock boundary when the true DLT rate is higher than tolerable. However, the Pocock boundary allows stopping much earlier than the O'Brien-Fleming boundary and therefore it is usually used to stop the trial for adverse events or toxicity. For example, as shown in Table 1, if the Pocock boundary is used, the trial will be stopped if 3 DLTs are observed in the first 3 patients. In comparison, if the O'Brien and Fleming boundary is used, the earliest stopping point requires the first 6 patients all experience DLT.

The Pocock stopping rule can be alternatively described as repeated testing of toxicity rate after each patient completes toxicity follow-up with the null hypothesis that the DLT rate is equal to $\theta_0 = 0.2$ and type I error rate α'. This is also equivalent to using a confidence interval approach. The trial is stopped after *k* patients if the lower bound of the $1 - 2\alpha'$ level two-sided confidence interval (Clopper and Pearson, 1934) for DLT rate computed when *k* patients completed the trial is above $\theta_0 = 0.2$. Here α' is a point-wise α-level. In the Pocock stopping boundary in Table 2.1, $\alpha' = 0.0196$. For example, if 4 out of 5 DLTs are observed, the $1 - 2\alpha' = 1 - 2\times 0.0196$ exact confidence interval for DLT rate is (0.266, 0.996). The lower bound of the interval, 0.266, is higher than $\theta_0 = 0.2$, or, in other words the confidence interval does not include 0.2, and therefore the null hypothesis that the toxicity rate is equal to $\theta_0 = 0.2$ is rejected and the

trial is stopped. The point-wise α-level, α', can be computed from a given type I error rate α. Ivanova**,** Qaqish and Schell (2005) gave a table of values α' for various sample sizes and tolerable DLT rates θ_0 . Free software to generate the Pocock stopping boundary is available at http://cancer.unc.edu/biostatistics/program/ivanova/. For given K, θ_0 and α, the software computes the stopping boundary and important quantities that describe the boundary's performance. For several values of the true DLT rates the program computes the probability of stopping the trial and declaring that the drug is too toxic, the average number of DLTs and the average number of patients in the trial (Table 2.2). For example, when the DLT rate is 0.4 about half of the trials will be stopped (probability of stopping is 0.55). The software also gives an example write-up that can be used in clinical trial protocols.

Table 2.2 Operating characteristics of the Pocock boundary with 20 patients, tolerable DLT rate of $\theta_0 = 0.2$ and the type I error rate of 0.05

True DLT rate	The probability of early stopping	Expected number of DLTs	Expected the number of enrolled patients
0.2	0.05	3.9	19.5
0.4	0.55	5.8	14.6
0.5	0.83	5.4	10.8
0.6	0.97	4.7	7.8
0.8	1.00	3.6	4.5

2.5 Comparison of the stopping boundaries

As seen from Table 2.1, the Bayesian boundary with Beta(0.6, 2.4) prior and $\tau =$ 0.98 is almost indistinguishable from the Pocock boundary. This is true in general: less informative priors, or lower values of $a + b$, yield Bayesian boundary similar to the Pocock boundary as long as the two boundaries yield similar overall probability of stopping. As the value of $a + b$ increases, that is, the prior becomes more informative, more DLTs are required to occur within the ongoing phase II trial to recommend trial interruption. We need to observe more DLTs to stop the trial because we need to 'override' the prior information that toxicity rate is tolerable. In the example in Table 2.1, under the Bayesian rule with informative prior with the probability of stopping of 0.05 ($\tau =$ 0.911), we stop later than under the Pocock boundary, but stop earlier than under the O'Brien-Fleming boundary. Another method for stopping due to toxicity we saw in phase II protocols is stopping based on the sequential probability ratio test (SPRT) (Wald, 1945, Armitage, 1957, Goldman, 1987). This method leads to a boundary very similar to the Pocock boundary for given sample size and given actual type I error rate.

If minimal prior information is available, either the Pocock boundary or the Bayesian boundary can be used. Even though they are described by a different statistical language, they are almost identical given the total sample size and the probability of stopping the trial when DLT rate is tolerable. If prior toxicity rate information is available, we recommend using the Bayesian boundary as this prior information can be reflected in the prior distribution for toxicity rate.

Since the Bayesian boundary with slightly informative prior is very similar to the Pocock boundary if the two boundaries have the same probability of stopping the trial

when $\theta = \theta_0$, frequentist software at http://cancer.unc.edu/biostatistics/program/ivanova/ can be used to construct the Bayesian boundary. The value τ in the Bayesian boundary can be set to equal to 1 minus step-wise significance level α' corresponding for the common probability of stopping, α, under acceptable DLT rate $\theta = \theta_0$. Software to construct the Bayesian boundary with the type I error restriction for different type of priors is available from the authors upon request.

2.6 Stopping rules we do not recommend using

In this section we review several stopping rules we have seen in phase II clinical trial protocols that we do not recommend using.

2.6.1 Rule when the trial is stopped as soon as *n* DLTs are observed

According to this rule the trial is stopped as soon as *n* DLTs are observed. Despite its simplicity, this rule does not take into account the denominator, that is, the number of patients enrolled in the study at the time of analysis. For example, for a trial with 20 patients and tolerable DLT rate of 0.20, the two rules with the type I error rate closest to 0.05 are 'stop the trial when 8 DLTs are observed' with the type I error rate of 0.032 and 'stop the trial when 7 DLTs are observed' with the type I error rate of 0.087. Obviously, irrespective of DLT rate, at least 7 or 8 DLTs are to be observed before the trial is stopped. In comparison, the maximum expected number of DLTs under the Pocock boundary is 5.8 when the true DLT rate is 0.4 and less for higher DLT rates.

This constant boundary is often used when the acceptable DLT rate is low, e.g., 0.05. According to the Pocock boundary with a maximum of 20 patients, $\theta = \theta_0$ and $\alpha =$

0.05, the trial is stopped if 2 DLTs are observed in the first 2-5 patients, or 3 DLTs are observed in first 6-14 patients, or 4 DLTs are observed in more than 14 patients. This boundary yields less expected DLTs than a constant boundary where the trial is stopped as soon as 3 DLTs are observed. Therefore we recommend using the Pocock or the Bayesian boundary and not the constant boundary.

2.6.2 Stopping rule based on the upper bound of a confidence interval

As mentioned in Section 2.4, stopping according to the Pocock boundary is equivalent to stopping the trial when the lower bound of a confidence interval is above the acceptable DLT rate. In the example in Section 2.4 we compared the lower bound of the confidence interval (0.266, 0.996), 0.266, with θ_0. As the purpose of the stopping rule is to stop the trial if the true DLT rate is high, it might seem reasonable to construct a stopping rule using the upper bound of the confidence interval instead. The trial is stopped if a pre-specified unacceptably high DLT rate, say, 0.5 is included in the interval, that is, if the upper bound of the interval, 0.966, exceeds 0.5. With this approach, the null hypothesis is that the DLT rate is unacceptably high (0.5 in this example) and the trial continues while the null hypothesis is being rejected, i.e., the confidence interval for the DLT rate does not contain 0.5. If the upper bound of the confidence interval exceeds 0.5, the null hypothesis is accepted and the trial is stopped. One problem with this rule is that it is very easy for the upper bound to exceed a given higher DLT rate, e.g. 0.5, in the beginning of the trial because the confidence interval is rather wide when little information is available. For example, if the first patient experiences DLT, the upper limit

of the confidence interval is 1 no matter what significance level is used. With 1-2α' = 0.9, for example, if the true DLT rate is 0.2, the trial has the probability of stopping of 0.2 after the first patient and the overall probability of stopping of 0.82 for a trial with 20 patients.

2.7 Conclusions

We reviewed several stopping rules for toxicity we have infrequently seen in phase II trial protocols. We propose to keep the probability of stopping the trial when the DLT rate is equal to the acceptable DLT rate at 0.05 or lower. In such cases, one needs to have rather strong evidence that the DLT rate is high to stop the trial early for toxicity. The goal is to stop the trial as early as possible if there is 'strong' evidence of high DLT rate. The term 'strong' implies that stopping rules for toxicity should by no means be extremely conservative to the point that they overshadow the main purpose of the phase II study, namely the efficacy assessment. Under this concept we anticipate that stopping rules for toxicity would be activated before efficacy rules infrequently. In addition, we argue that continuous stopping rule for toxicity should be used, that is, the rule that allows stopping the trial at any point. Between the two stopping boundaries most commonly used in clinical trials, the O'Brien-Fleming boundary and the Pocock boundary, we recommend the Pocock boundary as it allows stopping for toxicity as early as possible.

If the investigator does not wish to use prior information about toxicity rate in phase II for various reasons (e.g. the phase I study population is significantly different from the phase II population or, alternatively, very few patients were studied under the

given dose in phase I and therefore the information derived from phase I trial is not very relevant), the Pocock or the Bayesian boundary with non-informative prior can be used, as they are virtually identical in this case. If there is reliable prior information about toxicity rate to use in the stopping rule, we recommend using the Bayesian boundary as it is the only boundary that can formally account for prior information about toxicity. On the other hand, prior information has to be used with caution as various factors including the change in patient population might affect the DLT rate of the investigational treatment. Boundaries mentioned in Section 2.6, are not recommended.

All of the boundaries we investigated use patients with complete-mature follow-up DLT data. These methods, therefore, might not be appropriate for trials that require long-term follow-up to observe DLT or when patients are still under study and/or early in their treatment in which case insufficient information is available. Several methods have been developed for trials with long follow-up for toxicity (Follmann and Albert, 1999, Rosner, 2005) For trials with an unrestricted accrual rate a significant number of patients could be enrolled within a short period of time, which might result in seeing a significant number of DLTs. In that case a stopping rule alone might not prevent from observing excessive toxicity. For clinical trials in oncology where toxicity can be both devastating and take time to develop (e.g. investigational agent plus radiation therapy for patients with brain metastases), we recommend considering both stopping and enrollment rules. An enrollment rule guides the accrual rate of patients not allowing enrolling many patients at once when not much is known about the DLT rate in the trial or when there is evidence that the DLT rate can be high.

CHAPTER 3

FREQUENTIST ENROLLMENT AND STOPPING RULES FOR MANAGING TOXICITY REQUIRING LONG FOLLOW-UP IN PHASE II ONCOLOGY TRIALS

3.1 Overview

Monitoring of toxicity is often done in phase II trials in oncology to avoid an excessive number of toxicities if the wrong dose is chosen for phase II. Existing stopping rules for toxicity use information from patients who have already completed follow-up. We describe a stopping rule that uses all available data to determine whether to stop for toxicity or not when follow-up for toxicity is long. We propose an enrollment rule that prescribes the maximum number of patients that may be enrolled at any given point in the trial.

3.2 Introduction

Many oncology phase II trials implement a stopping rule for toxicity. This is because the toxicity profile of a drug or a drug combination used in a phase II trial might not be well understood by the time the phase II trial commences. A number of stopping rules for toxicity have been proposed for use in a single-arm trial. Ivanova, Qaqish and Schell (2005) argued for the use of the Pocock type (Pocock, 1977) stopping boundary; Geller et al. (2005) proposed a Bayesian stopping rule. Both argued that a continuous

stopping rule, a rule that allows stopping the study at any point in the enrollment process, provides the best protection against observing an excessive number of toxicities and therefore is preferable to two or three-stage stopping rules. The number of stages in a continuous stopping rule is the same as the number of patients in the trial. The continuous Pocock boundary is routinely used in phase II oncology trials conducted by the Lineberger Comprehensive Cancer Center (LCCC). In some oncology trials the follow-up for toxicity is long compared to the accrual rate. This was the case in a phase II study of a novel B-Raf inhibitor administered together with a monoclonal antibody in patients with active melanoma brain metastases (Moschos, 2013). The observation period for dose limiting toxicity (DLT) was 12 weeks. The Pocock boundary was in place to monitor the DLT rate. To use the Pocock boundary one needs to have full follow-up data on all already enrolled patients. If investigators in this study wait for all patients to be fully followed before the next patient is enrolled, the length of the trial with 30 patients will be over 7 years. Is it possible to make the trial shorter without increasing the number of patients potentially exposed to unsafe treatment? The problem can be solved 1) by developing a stopping boundary that uses all data available including data from patients still in follow-up, and 2) by guiding the accrual rate to the trial to avoid the situation where all patients are enrolled at once. A problem of conducting studies with long follow-up arises in phase I oncology trials as well (Cheung and Chappell, 2000, Bekele et al., 2008). As far as utilizing partial data in a phase I dose-finding trial, Cheung and Chappell (2000) proposed to make an assumption about the distribution of time to toxicity in $(0, t^*)$ given toxicity occurs in $(0, t^*)$, where t^* is the length of follow-up for toxicity. The best dose for the next patient is then selected based on all available data.

Bekele et al. (2008) suggested halting enrollment to a phase I trial when there is not enough data to select a safe dose for the next patient. The problem is different in a phase II context as it is a single arm trial and dose reduction in the middle of the trial will make it difficult, if not impossible, to estimate the efficacy and toxicity of the experimental compound at a certain dose level. The main challenge in a phase II trial with long follow-up for toxicity is to maintain the pre-specified probability of stopping the trial under various true toxicity rates, the type I error rate and power, when partial data are used. Methods for phase I trials with long follow-up for toxicity do not apply in phase II context. However, we use the same assumption as in Cheung and Chappell (2000) to utilize partial data. In this paper in Section 3.3 we describe a stopping rule for toxicity based on partial data. Enrollment strategies are discussed in Section 3.4, the simulation study is presented in Section 3.5, example in Section 3.6 and discussion in Section 3.7.

3.3 Stopping for toxicity based on partial data

In a phase II study each patient is followed for toxicity for a fixed period time of t^*. Let n be the number of patients enrolled in the study so far. Let U_i be the random variable denoting the time to toxicity for the ith patient and let $\theta = P(U_i \leq t^*)$ be the probability of toxicity. Denote $Y_{i,n}$ the indicator that the ith patient has experienced toxicity by the time just prior to the entry of the $(n+1)$th patient, $i = 1,\ldots,n$, then $X_n = \sum_{i=1}^{n} Y_{i,n}$ is the random variable denoting the total number of toxicities observed at that time.

In a trial where all patients are fully followed for toxicity when the next patient is assigned, $Y_{i,n}$ does not depend on n and follows a Bernoulli(θ) distribution and X_n ~

binomial(n, θ). Ivanova et al. (2005) considered such trials and argued that the Pocock boundary is the most suitable boundary for monitoring toxicity in a phase II oncology trial as it allows stopping early with high probability and therefore reduces the expected number of toxicities. Let K be the sample size planned for a phase II study and let θ_0 be the acceptable probability of toxicity. The Pocock boundary can be defined through a point-wise probability α, such that the trial is stopped if, at each interim analysis, the null hypothesis $\theta = \theta_0$ is rejected at level α in favor of the one-sided alternative $\theta > \theta_0$. The value of α is chosen so that the overall probability of stopping the trial, ϕ, is equal to a specified value, usually $\phi = 0.05$, if the toxicity rate is θ_0. We refer to the boundary that allows stopping the study at any point as a continuous boundary, because monitoring for toxicity is done throughout the trial on a continuous basis. Let the constants b_k, k = 1,…,K, be the smallest integer such that $P[X_k \geq b_k] \leq \alpha$, then such a boundary can be described through (b_1, b_2, …, b_K). If the number of toxicities in the first k patients is equal to or higher than b_k, the trial is stopped. Another way of implementing this boundary is to compute a one-sided p-value to test the null hypothesis $\theta = \theta_0$ versus one-sided alternative $\theta > \theta_0$ after each patient's outcome is observed. The trial is stopped if the p-value is less than α. In fact, it is sufficient to apply the boundary only when a patient has toxicity. We now explain how to use data from partially followed patients to implement a sequential boundary. We will compute the p-value to test the null hypothesis $\theta = \theta_0$ versus the one-sided alternative $\theta > \theta_0$ based on all information available using assumption on the distribution of time to toxicity in (0, t^*).

Now consider the case when not all n patients are fully followed for toxicity at the time just prior to the entry of the (n+1)th patient. Let $t_{i,n}$ be time elapsed from the start of treatment for the ith patient at the time just prior to the entry of the (n+1)th patient. Following Cheung and Chappell (2000), for $t_{i,n} < t^*$, we have

$$P(U_i \le t_{i,n}) = P(U_i \le t_{i,n} \mid U_i \le t^*)P(U_i \le t^*) = P(U_i < t_{i,n} \mid U_i \le t^*)\theta = \frac{t_{i,n}}{t^*}\theta = w_{i,n}\theta.$$

In other words, a weight $w_{i,n} = P(U_i < t_{i,n} \mid U_i \le t^*)$ is assigned to the ith patient and the probability that the ith patient experiences toxicity when treated for a length of $t_{i,n}$ is $w_{i,n}\theta$. This is equivalent to assuming that the time to toxicity given that toxicity occurs in (0, t^*) follows a uniform distribution on the interval (0, t^*). This and other weighting options were described in Cheung and Chappell (2000) and Yin (2012). The weight $w_{i,n}$ is set to 1 for patients who have already experienced toxicity and patients who have completed follow-up. At the time just prior to the entry of the (n+1)th patient, the number of patients who completed the trial without toxicity is $S_n = \sum_{i=1}^{n} I(w_{i,n} = 1) - X_n$ and the number of patients still under follow-up is $R_n = \sum_{i=1}^{n} I(0 < w_{i,n} < 1)$, $X_n + S_n + R_n = n$. Let x, s and r denote the observed X_n, S_n and R_n respectively. The one-sided p-value for testing the null hypothesis $\theta = \theta_0$ versus $\theta > \theta_0$ is the probability of $X_n \ge x$.

Table 3.1. The Pocock stopping boundary $\{b_k\}$ for K = 30, $\theta_0 = 0.2$ and $\phi = 0.05$ yielding $\alpha =$ 0.0164, and the Pocock enrollment boundary $\{b'_k\}$ with $\alpha' =$ 0.0030

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
b_k	--	--	3	4	4	4	5	5	6	6	6	7	7	7	8
b'_k	--	--	--	4	5	5	6	6	7	7	7	8	8	8	9

k	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
b_k	8	8	8	9	9	9	10	10	10	11	11	11	11	12	12
b'_k	9	9	10	10	10	11	11	11	12	12	12	13	13	13	14

For example, if n = 3, x = 2, $s = 1$ and $r = 0$, then all weights are equal to 1 and the p-value is calculated as $P[X_3 \geq x] = P[X_3 \geq 2] =$ 0.104, where X_3 is a binomial random variable with parameters 3 and θ_0 = 0.2, X_3 ~ binomial(3,0.2) . In another example, the counts right before enrolling the fourth patient are n = 3, $x = 2$, $s = 0$ and $r = 1$ with the first two patients fully followed and time $t_{3,3} = t^* / 2$ and hence $w_{3,3} = 1/2$ for the patient still in follow-up, then $X_3 = Y_{1,3} + Y_{2,3} + Y_{3,3}$, where $Y_{i,3}$ ~ Bernoulli(θ_0) for i =1,2 and $Y_{3,3}$ ~ Bernoulli($\theta_0 / 2$) . Therefore

$$P[X_3 \geq 2] = P[Y_{1,3} + Y_{2,3} \geq 1, Y_{3,3} = 1] + P[Y_{1,3} = Y_{2,3} = 1, Y_{3,3} = 0] = 0.072.$$

Table 3.2. Comparing the new stopping rule that uses all available data when time to toxicity is uniformly distributed in (0, t^*) (Uniform) and exponentially distributed with mean $-t^*/\ln(1-\theta_0)$ (Expon) with the rule that uses fully followed patients only (Full data) for a trial that $K = 30$, $\theta_0 = 0.2$, and $t^* = 12$ weeks with one patient being enrolled every week. For comparison we show results for a trial with instantaneous response ($t^* = 0$). We display the probability of stopping the trial early, the expected number of patients, and the expected number of toxicities

True toxicity		0.2	0.4	0.6	0.9
Probability of stopping	$t^* = 0$	0.05	0.70	1	1
	Uniform	0.03	0.68	1	1
	Expon	0.03	0.69	1	1
	Full data	0.03	0.61	0.99	1
Expected number of patients	$t^* = 0$	29.1	19.0	8.3	4.6
	Uniform	29.6	23.8	15.2	10.0
	Expon	29.6	23.0	13.6	7.1
	Full data	29.8	27.1	21.0	15.8
Expected number of toxicities	$t^* = 0$	5.8	7.6	5.0	3.6
	Uniform	5.9	8.4	6.2	4.6
	Expon	5.9	8.1	5.7	3.9
	Full data	5.9	9.8	9.4	9.3

We illustrate the ability of this rule to stop the trial via simulations. Consider an example of a phase II trial with $K = 30$. To yield the overall probability of stopping of 0.05 when $\theta_0 = 0.2$, we need to use $\alpha = 0.0164$ at each step. The continuous Pocock boundary (Pocock, 1977) for this trial is shown in Table 3.1. In our simulations study t^* = 12 weeks and a new patient is enrolled every week. Table 3.2 contains results for the proposed stopping rule under uniform and exponential distribution of time to toxicity. For comparison, we present data for a trial with instantaneous toxicity outcome, $t^* = 0$, and a trial with $t^* = 12$ weeks where only data from patients who were fully followed (i.e., full follow up time t^* passed from the initiation of treatment for all patients irrespective of their outcome) is used in a stopping rule. As seen from Table 3.2, our rule allows us to stop the trial earlier and to observe less toxicity on average, especially when toxicity rate

is high. We repeated the simulations for various patterns of patient enrollment and the results were very similar to those in Table 3.2.

3.4 Enrollment rule to prevent an excessive number of toxicities

If many patients are enrolled at once, the stopping rule described in the previous section will not prevent treating too many patients on a regimen that may not be safe. Often, many patients are enrolled at the very beginning of the trial which might lead to excessive toxicities. An enrollment rule informs investigators about how many patients may be enrolled at the beginning of the trial and guides further accrual based on the information about toxicity in the trial.

Consider the boundary in Table 3.1. Initially we may enroll 3 patients as it is not possible to stop the trial before 3 patients complete follow-up. If none of these patients experience toxicity in $(0, t^*)$, one may enroll as many as 5 more patients, since there is a possibility to cross the boundary by observing 5 toxicities out of 8 patients, and it is not possible to cross the boundary if less than 5 additional patients are enrolled. More formally, the trial can enroll m new patients such that $r + x + m \leq b_{n+m}, r + x + m - 1 < b_{n+m-1}$ and $n + m \leq K$. Here we assume the worst case scenario that toxicity rate is $\theta = 1$ and therefore every patient in the follow-up will experience toxicity. This approach was referred to as the conservative plan in Schmegner and Baron (2004) who considered it in the context of sequential planning of experiments. This is the most conservative enrollment rule and the number of toxicities we observe will be very similar compared to the trial with instantaneous toxicity response. On the other hand, this rule can lead to a

rather long trial if t^* is long. We considered three ways to relax this rule resulting in three different enrollment strategies.

The first enrollment strategy is described as follows. Let M be the design parameter fixed in advance. One can think of M as the number of extra toxicities we are willing to allow to make the trial shorter. The maximum number of new patients to enroll, m, is determined by $r + x + m \le b_{n+m} + M$, $r + x + m - 1 < b_{n+m-1} + M$ and $n + m \le K$. That is, at any time the maximum number of patients experiencing toxicity cannot exceed the number allowed by the Pocock boundary plus M. As before we assume the worst case scenario that all patients will experience toxicity and allow M extra toxicities beyond what is allowed by the Pocock boundary. When $M = 0$, the rule is equivalent to the conservative enrollment plan. The maximum number of patients to enroll in the trial initially is $b^* + M$, where b^* is the minimum number k such that $k \ge b_k$. In the example in Table 3.1, we can enroll at most $3 + M$ patients initially. If M is as large as $M \ge K - b^*$, all patients can be enrolled in the beginning of the study.

The second enrollment strategy we consider is to use a separate Pocock boundary for enrollment, the boundary that yields the overall probability of stopping the trial of ϕ', $\phi' \le \phi$, or step-wise probability α' in place of α, $\alpha' \le \alpha$, when $\theta = \theta_0$. Let $\{b'_k\}$ be the set of constants corresponding to the second Pocock boundary. The number of patients we are allowed to enroll, m, is such that $r + x + m \le b'_{n+m}$, $r + x + m - 1 < b'_{n+m-1}$ and $n + m \le K$. An example of an enrollment boundary $\{b'_k\}$ with $\alpha' = 0.003$ is shown in Table 3.1. The trial can start with as many as 4 patients. If, for example, in the middle of the trial there are 2 patients who have completed the trial without toxicity, 3 patients experienced toxicity with 2 patients still in the follow-up, $n = 3 + 2 + 2 = 7$; $x + r = 3 + 2$

= 5, $b_7' = b_8' = 6$, we get $m = 1$ and therefore we may enroll one more patient. If $\phi' = \phi$, the rule is equivalent to the conservative enrollment plan.

In the third enrollment strategy a patient completed t / t^* of the total follow-up contributes $1 - t / t^*$ toxicity to the total toxicity count we will use to calculate allowable enrollment. The total toxicity count, ξ, just prior to the entry of the (n+1)th patient is x plus the sum of $1 - t_{i,n} / t^*$, where the sum is over all the patients still in the follow-up. Similarly to how it was done in Section 3.3, one can compute the p-value to test $\theta = \theta_0$ given $\xi + m$ toxicities out of $n + m$ patients. The maximum number of patients one may enroll is the maximum m such that $m \le K - n$ and

$$P\left[X \ge \xi + m \mid X \sim \text{binomial}(\theta_0, n+m)\right] \le \alpha .$$

In the beginning of the trial, this enrollment rule is the same as the conservative enrollment plan since $\xi = 0$ and $n = 0$. In the example in Table 3.1, say, if there are 2 toxicities observed, 1 patient completed the trial without toxicity and 3 patients are right in the middle of the follow-up, then $\xi = 2 + 3 \times 0.5 = 3.5$ and $n = 2 + 1 + 3 = 6$. Calculating the probability $P\left[X \ge 3.5 + m \mid X \sim Binomial(0.2, 6+m)\right]$ for various m and comparing it with $\alpha = 0.0164$ shows that the maximum number we may enroll is $m = 0$. Since ξ decreases as patients are followed for longer times and given no new toxicity is observed, the number of patients we may enroll, m, increases. Given current data one can compute when additional patients can be enrolled in case no new toxicities are observed. In the description above we took the most conservative approach and used the toxicity rate $\theta = 1$. One can use smaller θ to compute toxicity count ξ.

We examined all three strategies. All three strategies allow flexibility as one can vary parameters M, α' and θ correspondingly. In the trial in Table 3.1, choosing $M = 1$, for example, has very similar results to the second strategy with $\alpha' = 0.0030$, and choosing $M = 3$ has very similar results to the second strategy with $\alpha' = 0.0003$. Choosing $M = 2$ in the first strategy is similar to the third strategy with $\theta = 1$. The third strategy utilizes the partial data in the trial better than other strategies but it is substantially more complex to implement as it requires real time calculations. The first two strategies do not require complex real time calculations and can be easily implemented during the trial according to specifications in a clinical trial protocol. Also the first strategy has a clear interpretation as allowing at most M additional toxicities over the stopping rule. As the performance of the three strategies is similar, we recommend the first enrollment strategy because of its simplicity. We will refer to this strategy as $+M$ enrollment rule in the remainder of the paper.

As mentioned earlier, using just the stopping or just the enrollment rule will not prevent the trial from possibly seeing excessive toxicity. The algorithm below describes how to apply both the stopping rule from Section 3.3 and the $+M$ enrollment rule described in this Section in a clinical trial.

(i) Initial enrollment is $b^* + M$. For example, in Table 3.1, $b^* = 3$.

(ii) When toxicity is observed, calculate the p-value as described in Section 3.3. If p-value is less than α, stop the trial.

(iii) When there is a toxicity or a patient reached the end of follow-up t^* without toxicity, if $x + r \geq b_{x+s+r} + M$, no new patients may be enrolled. If $x + r < b_{x+s+r} + M$ and the enrollment limit has not been reached, find smallest

integer $\tilde{m}$ that satisfies $\tilde{m} = b_{x+s+r+\tilde{m}} + M - (x+r)$. If $\tilde{m} > x+r+s$ then enroll $K-(x+r+s)$ patients, otherwise, enroll $\tilde{m}$ patients.

3.5 Simulation results and discussion of design parameters

In this section we present a simulation study investigating the performance of the $+M$ enrollment rule for various values of M in conjunction with the stopping rule described in Section 3.3. We used the example from Section 3.3 with $K = 30$ and $\theta_0 = 0.2$. Figures 3.1-3.3 show the expected number of toxicities, expected number of patients enrolled and expected length of trial (in units of t^*) for some values of M across the range of true toxicity rate. When $M = 0$ the probability of stopping the trial is almost the same as ϕ. As M increases, assuming that patients are always available to enroll in the trial, the probability of stopping the trial decreases slightly. At the same time, the expected number of toxicities is increasing mostly because many more patients are enrolled before the trial is stopped and not because the probability of stopping gets slightly lower than with $M = 0$. On the other hand, as M increases the trial gets shorter. Note that when $M = 27$ all patients may be enrolled at the beginning of the trial.

Figure 3.1. Expected number of toxicities plotted versus true toxicity rate for different values of M in a trial with $K = 30$, $\theta_0 = 0.2$ and $\phi = 0.05$.

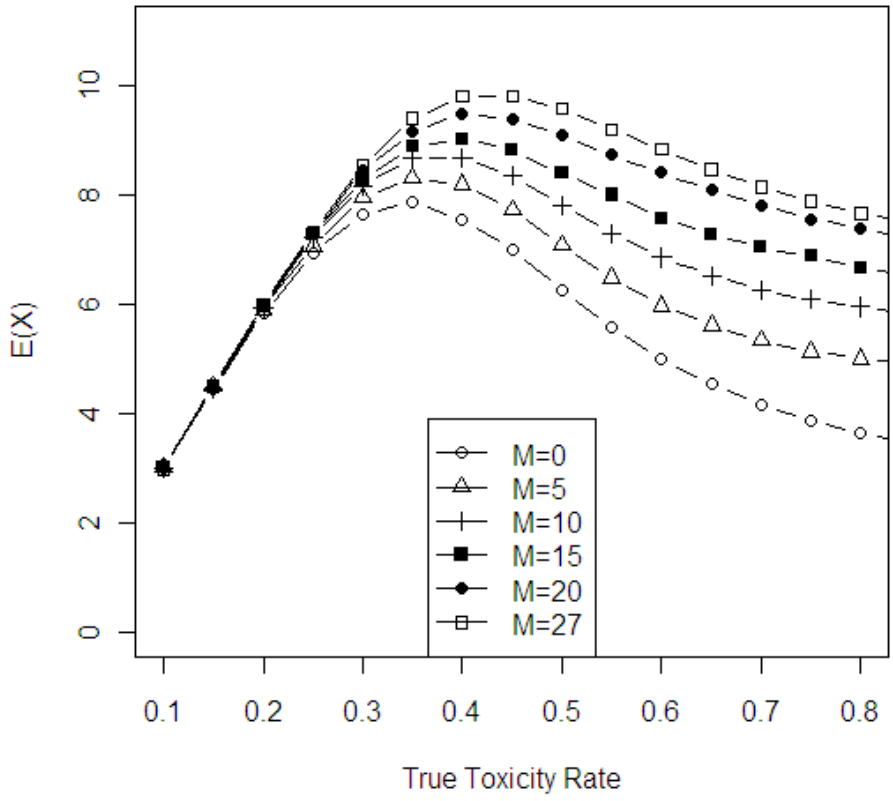


Figure 3.2. Expected number of patients enrolled in the trial versus true toxicity rate for different values of M in a trial with $K = 30$, $\theta_0 = 0.2$ and $\phi = 0.05$.

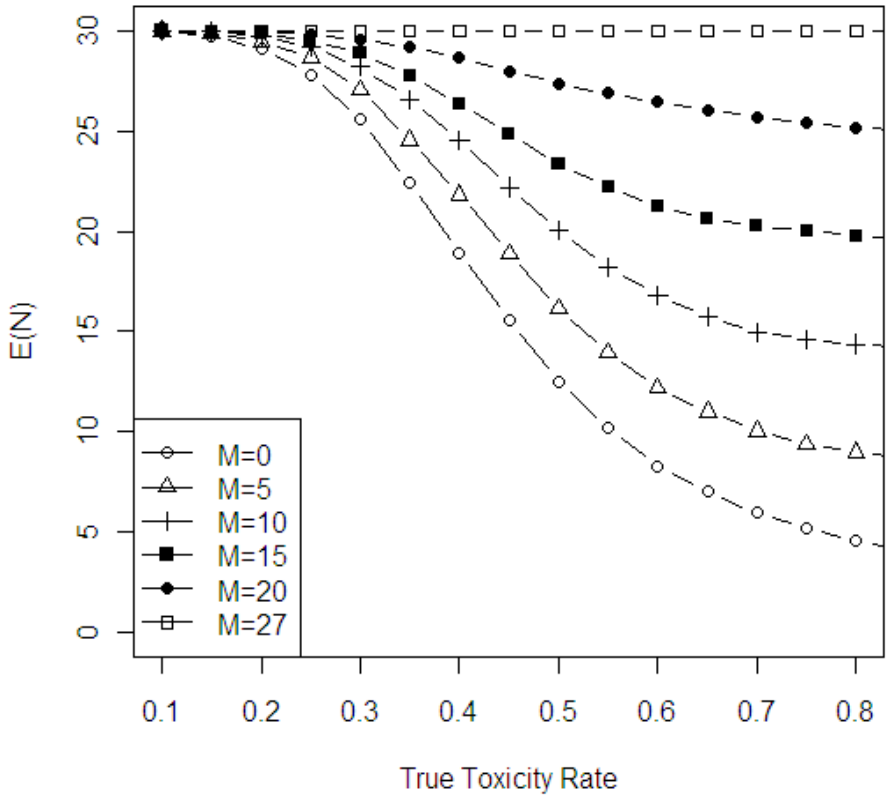
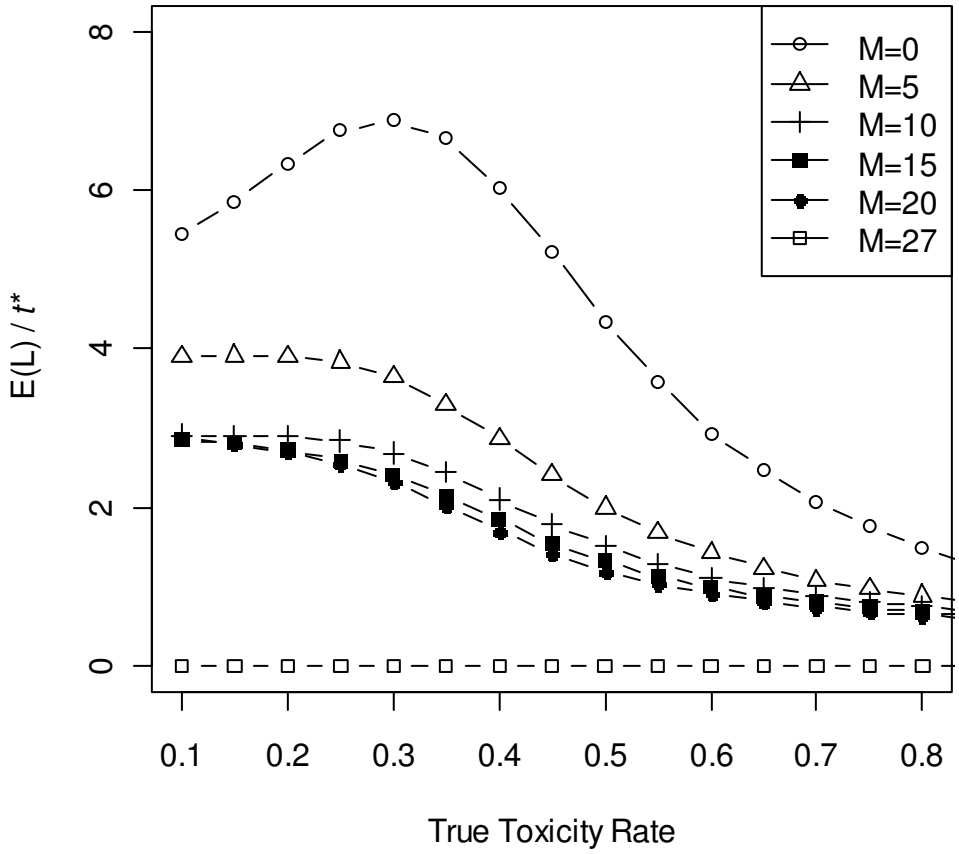


Figure 3.3. Expected length of study versus true toxicity rates for values of M in a trial with $K = 30$, $\theta_0 = 0.2$ and $\phi = 0.05$.



To choose an appropriate *M* for a trial, we notice that for a given *M*, the expected number of toxicities rises as the true toxicity rate increases. If toxicity rate $\theta = 1$, the increase in expected toxicity compared to a sequential trial with instantaneous response is *M*. For $\theta < 1$ the expected increase will never exceed θM. Furthermore, because a stopping rule is in place, the expected increase is smaller than θM . One can estimate the expected increase in toxicity for each stopping and enrollment rule combination and t^* by simulations. Consider the example in Table 3.1 and assume that the true toxicity rate cannot be higher than $\theta = 0.6$. Simulations show that if we chose $M = 5$ in the *+M* enrollment rule, we will see at most one extra expected toxicity when $\theta \leq 0.6$. Therefore $M = 5$ is a good choice of parameter value in the *+M* enrollment rule if we are willing to allow at most one extra toxicity.

3.6 Example

The proposed methodology is used in ongoing LCCC pharmacokinetic study of patients with high risk myelodysplasia and acute leukemia. Dose limiting toxicity outcomes (yes or no) were defined as non-relapse mortality or one of the following toxicities observed during the first 8 weeks from the start of treatment: grade 3 non-hematologic toxicity lasting greater than 7 days, grade 4 non-hematologic toxicity with the exception of drug-related fever, or grade 3/4 hematologic toxicity lasting greater than 42 days. The total number of patients in the study was 22. The probability of stopping the trial was set to 0.05 when the true toxicity rate is equal to the tolerable rate of 0.2. The

investigators preferred the most conservative approach to enrollment and therefore the $+M$ enrollment rule with $M = 0$ was used in this study.

3.7 Conclusions

We propose a frequentist sequential stopping rule for toxicity that utilizes all available data in the trial. To control the number of toxicities in the study, we recommend using the stopping rule with enrollment strategy. The parameter to use in enrollment strategy can be chosen based on the maximum number of extra toxicities or to yield a desired trade-off between the length of the trial and an increase in expected number of toxicities. The continuous sequential boundary (Ivanova et al., 2005) can be generated by using software available at http://cancer.unc.edu/biostatistics/program/ivanova/.

CHAPTER 4

BAYESIAN ENROLLMENT AND STOPPING RULES FOR MANAGING TOXICITY REQUIRING LONG FOLLOW-UP IN PHASE II ONCOLOGY TRIALS

4.1 Overview

Stopping rules for toxicity are routinely used in phase II oncology trials. If the follow-up for toxicity is long, it is desirable to have a stopping rule that uses all toxicity information available, not only information from patients with full follow-up. Further, to prevent excessive toxicity in such trials, we propose an enrollment rule that informs an investigator about the maximum number of patients that can be enrolled depending on current enrollment and all available information about toxicity. We give recommendations on how to construct Bayesian stopping and enrollment rules to monitor toxicity continuously in phase II oncology trials with a long follow-up.

4.2 Introduction

The goal of a phase I oncology trial is to estimate the maximum tolerated dose (MTD). The MTD is defined as the dose with the probability of dose limiting toxicity (DLT) equal to the maximum tolerated level (often 0.2 or 0.25). The efficacy of the estimated maximum tolerated dose is investigated in a phase II trial, usually a single arm

study. Since phase I trials use a small sample size, the estimate of the maximum tolerated dose is imprecise and may result in a dose chosen for the phase II trial with a DLT rate that is much higher than the maximum tolerated level. Therefore stopping rules for DLT are routinely used in phase II trials. If the trial is stopped for excessive DLT, the experimental therapy investigated in the phase II trial is concluded to be unsafe. Though not every toxicity is dose limiting, in the remainder of the paper we will use DLT and toxicity interchangeably. Both Bayesian and frequentist boundaries for DLT monitoring in phase II oncology trials have been proposed. Most trials are monitored on a continuous basis, that is, with the possibility of stopping at any point, rather than with a multi-stage rule with only a few stages. Geller et al. (2003) described a Bayesian stopping rule for continuous monitoring of toxicity in phase II oncology trials. Ivanova, Qaqish, and Schell (2005) investigated continuous stopping rules for toxicity and gave a table with critical values for the Pocock boundary (Pocock, 1977) with probability of stopping the trial for excessive toxicity of 0.05 when the toxicity rate is tolerable. Ivanova et al. (2005) also gave estimators of toxicity and response rates that are less biased than the maximum likelihood estimates (MLE).

In some trials the accrual is rapid and the follow-up for toxicity is long. This was the case in a Lineberger Comprehensive Cancer Center (LCCC) trial investigating allogeneic hematopoietic cell transplantation in patients with hematologic disorders. Treatment related mortality (TRM) during the first six months was monitored. A TRM rate of 0.2 was considered acceptable. The total sample size was set to 30. Continuous Pocock stopping boundary yields a probability of stopping of 0.05 when a six-month TRM rate of 0.2 was used. According to this rule the trial should be stopped if 3 TRM

events are observed in the first 3 patients, 4 TRMs in the first 5 etc. Simulations show that if all the patients were fully followed before a new patient was enrolled, the maximum expected number of TRMs, 7.65, occurs when the TRM rate is 0.3. For a higher TRM rate the trial is stopped early and the average number of TRMs is less than 7.65. Accrual was very rapid in the beginning of the trial with 10 patients accrued during the first several weeks. Accrual of several patients at the same time might lead to observing more toxicities than in a trial where all patients are fully followed before a new patient is accrued (we will refer to such a trial as a fully sequential trial). With a TRM rate close to 1, if 10 patients were accrued in the beginning of the trial, as many as 10 TRMs could have been observed if all 10 patients get enough exposure to treatment to develop TRM before the trial is stopped. The local data and safety monitoring committee pointed out that the investigator should not allow rapid accrual to this trial as the TRM rate of the treatment is unknown. At that time the investigators did not know how to guide the enrollment to prevent excessive level of adverse events without slowing the trial considerably. There is a need for a rule that not only stipulates when to stop for excessive toxicity but also prescribes the maximum number of patients that can be enrolled in the trial given current information.

In many clinical trials in oncology the safety outcome is defined as a binary variable that takes a value of 1 if an event is observed during an observation period (0, t^*), and 0 otherwise, where t^* is the follow-up time for toxicity. When t^* is long, it is desirable to make intermediate decisions in the trial based on all the data, including data from patients still under follow-up. One can estimate toxicity rate by assuming that time to toxicity is uniform in (0, t^*) given toxicity occurs before t^* (Cheung and Chappell,

2000). It is a simple approach but might not work well if this assumption is not true. Follmann and Albert (1999) and Rosner (2005) proposed a Bayesian method for monitoring clinical trials with a failure-time endpoint, where interval (0, t^*) is divided into several intervals and toxicity rate in each interval is estimated.

In this paper we show that the probability of stopping the trial based on partial data depends not only on the distribution of time to event and true toxicity rate in (0, t^*), but also on the rate of accrual. Therefore we propose to use an enrollment rule that guides the rate of accrual in conjunction with a stopping rule to ensure correct probability of stopping and to reduce the occurrence of excessive toxicity. In Section 4.3, we describe how to set up a Bayesian stopping boundary that yields a given probability of stopping for trials with immediate response. In Section 4.4, we describe stopping based on all available data when the follow-up is long. In Section 4.5, we introduce an enrollment rule to use in conjunction with the stopping rule. Simulation results are presented in Section 4.6. Section 4.7 contains concluding remarks.

4.3 Bayesian stopping boundary for a trial with immediate response

Let N be the number of patients planned for a study. Let x_t be the observed number of patients who experienced toxicity, and s_t be the observed number of patients who completed the trial without toxicity. Let r_t be the number of patients who are enrolled in the study and still under observation, i.e., those patients who have neither completed the trial nor experienced toxicity. In the case when toxicity outcome is immediately available $r_t = 0.$ For a fully sequential trial, a trial where a patient cannot be

enrolled until toxicity outcomes of all previous patients are observed, r_t is at most 1. Let $n_t = x_t + r_t + s_t$ be the number of patients accrued to the study at time t. For simplicity, we always assume t is time elapsed from the beginning of study and drop it from the subscript in the following discussion. Let p be the probability that a patient experiences toxicity during time (0, t^*). We assume a $Beta(\alpha_1, \alpha_2)$ prior on p, hence the posterior distribution of p is $p \mid x, s \sim Beta(x+\alpha_1, s+\alpha_2)$.

For an M-stage trial, with $M - 1$ interim analyses and one final analysis, the stopping rule can be defined as follows:

$$\text{Stop the trial if } P(p > \theta \mid \text{Data}) > c_j \text{ for } j = 1, 2, \ldots, M.$$

Here θ is the pre-defined tolerable toxicity rate, and c_j are constants selected according to a desirable shape of the stopping boundary and such that the probability of stopping the trial given θ is equal to a pre-specified probability η, usually 0.05. For monitoring toxicity or adverse events, the Pocock boundary is well accepted as it minimizes the expected number of toxicities in the trial (see for example, Ivanova et al., 2005). We will show later that setting $c_1 = \ldots = c_K = c$ gives enough flexibility to generate Pocock-shape boundaries by choosing an appropriate prior distribution. We therefore consider stopping rules of the form

$$\text{Stop the trial if } P(p > \theta \mid Data) > c \text{ for } j = 1, 2, \ldots, M. \quad (1)$$

Because of the discreteness of the binomial distribution, the probability of stopping is not exactly η. We take a conservative approach and choose c to make the probability of stopping just above η. That is, any reduction in the probability of stopping caused by

changing c will lead to the probability of stopping being smaller than η. Note that because of discreteness, the value c is not unique.

Table 4.1. Stopping boundaries that yield probability of stopping of 0.05 when toxicity rate is θ = 0.2 and N = 20. Bayesian stopping boundaries are defined by the prior $Beta(m\theta, m(1-\theta))$

	Number of Patients Enrolled															
Stopping Boundary	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Beta(0.6,2.4)			3	4	4	5	5	5	6	6	6	7	7	7	7	8
Beta(2.4,9.6)				4	5	5	5	6	6	6	6	7	7	7	7	8
Beta(10,40)						6	6	6	6	7	7	7	7	8	8	8
Pocock			3	4	4	4	5	5	6	6	6	7	7	7	8	8
O'Brien-Fleming						6	7	7	7	7	7	7	8	8	8	8

	Number of Patients Enrolled													
Stopping Boundary	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Beta(0.6,2.4)	8	8	9	9	9	10	10	10	10	11	11	11	11	12
Beta(2.4,9.6)	8	8	8	9	9	9	10	10	10	10	11	11	11	11
Beta(10,40)	8	8	9	5	9	9	9	10	10	10	10	11	11	11
Pocock	8	8	9	9	9	10	10	10	10	11	11	11	12	12
O'Brien-Fleming	8	8	9	9	9	9	9	10	10	10	10	10	11	11

For normally distributed outcomes, Freedman and Spiegelhalter (1989) considered trials with up to 5 stages and showed that a Bayesian stopping boundary can have a shape similar to the O'Brien-Fleming (O'Brien and Fleming, 1979) or Pocock boundaries depending on the prior, with Pocock type boundaries arising from non-informative or slightly informative priors. We investigated a continuous boundary, a boundary with as many stages as there are patients, and observed that the same is true for binary outcomes. A non-informative or slightly informative prior creates a boundary of a shape similar to the Pocock boundary, even though a perfect match may not always be found. In Table 4.1 we show boundaries that yield the probability of stopping of 0.05

when toxicity rate is $\theta = 0.2$. For each boundary we give a set of constants b_n, $n = 1,\ldots,N$, such that the trial is stopped if the number of toxicities in n patients with full follow-up is equal to or exceeds b_n. The boundary with the prior $Beta(0.6,2.4)$ is similar to the Pocock boundary. In general, we suggest a slightly informative prior in the form of $Beta(m\theta, m(1-\theta))$, where m is small. This prior reflects the belief that the DLT rate is θ with m being an effective sample size. When $N \leq 60$, $\theta \leq 0.3$, choosing $m = 3$ yields boundaries very similar in shape to the Pocock boundary.

For a given prior we need to find the constant c in (1) to yield the desired probability of stopping for a given toxicity rate. For example, the probability of stopping of 0.05 when toxicity rate is equal to tolerable rate θ. Given θ, prior and c, one can calculate the stopping boundary $b_1,\ldots,b_N$ directly from the posterior, where the condition $X \geq b_n$ is equivalent to $P(p > \theta \mid data) > c$, with X being the random variable representing the number of toxicities. This probability can be computed using a recursive formula proposed by Schultz et al. (1973), by observing that to stop after n patients have been enrolled, the nth patient has to experience toxicity and there has to be exactly b_{n-1} toxicities prior to enrollment of the nth patient:

$$W_n = \binom{n-1}{b_n - 1} - \sum_{j=1}^{n-1} W_j \binom{n-j-1}{b_n - b_j - 1},$$

$$\eta' = \sum_{n=1}^{N} W_n \times \theta^{b_n} \times (1-\theta)^{n-b_n}.$$

Here η' is the unconditional probability of stopping given $\{b_1,\ldots,b_N\}$, and W_n, $n = 2,\ldots, N$, are computed recursively with $W_1 = 1$. Constant c and corresponding b_n can be found by trial and error to yield the required probability of stopping (R code is available from the authors). Table 4.1 lists c for various tolerable toxicity rates and total sample sizes N

for a Beta prior with m = 3, and m = 12 with mean θ. As expected, values 1 - c for the Bayesian boundary with *Beta*(3θ, $3(1-\theta)$) prior are similar to those for the frequentist boundary given in Table 3 of Ivanova et al. (2005). Note that the value of c decreases as the prior gets more informative. This is because such priors put more weight on the belief that $p = \theta$ and hence one needs a lower c to keep the probably of stopping at desired value η.

4.4 Stopping boundary for trials with delayed outcome

For a trial with a long follow-up for toxicity, we would like to utilize all the data available to decide whether to stop the trial or not, not only data from patients with full follow-up. For the fixed time sequence $0 = t_0 < t_1 < t_2 < \ldots < t_K = t^*$, let T_i denote interval (t_{i-1}, t_i), and $\boldsymbol{T} \equiv (T_1, T_2, \ldots, T_K, T_{K+1})$ be the set of the $K + 1$ intervals with $T_{K+1} = (t_K, +\infty)$. Let random variable $Y(T_k)$ for $k \le K$ denote the indicator of toxicity occurring in T_k, with $Y(T_k) = 1$ indicates a DLT occurs in T_k, and 0 otherwise. The random variable $Y(T_{K+1})$ is equal to 1, if a patient does not have a DLT in t^*, and 0 otherwise. Assume that $\boldsymbol{Y}(\boldsymbol{T}) \mid \boldsymbol{p}$, $\boldsymbol{Y}(\boldsymbol{T}) \equiv \left(Y(T_1), \ldots, Y(T_{K+1})\right)$, follows a multinomial distribution with parameter $(1, \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \ldots, p_{K+1})$, $p_k = E[Y(T_k)]$, $p_1 + \ldots + p_{K+1} = 1$. Then $p = p_1 + \ldots + p_K$ is the probability that a patient experiences toxicity during time $(0, t^*)$.

We assume that $\boldsymbol{p}$ has a conjugate Dirichlet prior with parameter $\boldsymbol{\alpha}$, $Dir(\boldsymbol{\alpha})$. The posterior distribution of $\boldsymbol{p}$ after observing all data follows a Dirichlet distribution,

$\boldsymbol{p} \mid \boldsymbol{y}(\boldsymbol{T}) \sim Dir(\boldsymbol{y}+\boldsymbol{\alpha})$, where $\boldsymbol{y}(\boldsymbol{T}) \equiv (y(T_1), y(T_2),..., y(T_{k+1}))$ are observed $\boldsymbol{Y}(\boldsymbol{T})$. To apply a stopping rule we enumerate all possible outcomes for each interval and compute their probabilities. Let I_k denote the number of patients who already experienced toxicity at time interval *k* and let J_k denote the number of patients who will experience toxicity at time interval k among patients currently enrolled. We have $I_k \geq 0$ and $I_1 + ... + I_K = x$, $J_k \geq 0$ and $0 \leq J_1 + ... + J_K \leq r$. Following Antonick (1974), Blum and Sursala (1977) and Follmann et al. (1999), with a $Dir(\boldsymbol{\alpha})$ prior, the posterior distribution given the observed enrollment profile will be a mixture of Dirichlet distributions $Dir(\boldsymbol{\alpha}+\boldsymbol{I}+\boldsymbol{J})$, where $\boldsymbol{\alpha}+\boldsymbol{I}+\boldsymbol{J} = \left(\alpha_1 + I_1 + J_1,...,\alpha_K + I_K + J_K, \alpha_{K+1} + I_{K+1} + J_{K+1}\right)$ for all possible trial results ϑ (i.e., all different possible $\boldsymbol{J}$). The probability of the realization for each $\boldsymbol{J}$ is proportional to $A(\boldsymbol{J})$, where $A(\boldsymbol{J})$ is determined by $\boldsymbol{I}+\boldsymbol{J}$ as follows:

$$A(\boldsymbol{J}) = \prod_{k=1}^{K+1} \frac{\Gamma\left(\alpha_k + I_k + J_k\right)}{\Gamma\left(J_k + 1\right) \times \Gamma\left(\alpha_k + I_k - 1\right)}.$$

Note that formula for $A(\boldsymbol{J})$ given in Follmann et al. (1999) is not correct. The derivation of the above formula is below.

Proof: Given $x + s$ were observed and r patients are still in follow-up, the likelihood function can be written as

$$\sum_{J \in \vartheta} \binom{x+s}{I_1 I_2 \ldots I_{K+1}} p_1^{I_1} p_2^{I_2} \ldots p_{K+1}^{I_{K+1}} \binom{r}{J_1 J_2 \ldots J_{K+1}} p_1^{J_1} p_2^{J_2} \ldots p_{K+1}^{J_{K+1}}.$$

With the prior

$$\frac{\Gamma(\alpha)}{\prod_{k=1}^{K+1}\Gamma(\alpha_k)} p_1^{\alpha_1-1} p_2^{\alpha_2-1}\cdots p_{K+1}^{\alpha_{K+1}},$$

the posterior becomes

$$\sum_{J\in\vartheta} \frac{\Gamma(\alpha)}{\prod_{k=1}^{K+1}\Gamma(\alpha_k)} \binom{x+s}{I_1 I_2 \ldots I_{K+1}} \binom{r}{J_1 J_2 \ldots J_{K+1}} p_1^{I_1+J_1+\alpha_1} p_2^{I_2+J_2+\alpha_2}\cdots p_K^{I_K+J_K+\alpha_K} p_{K+1}^{I_{K+1}+J_{K+1}+\alpha_{K+1}}.$$

After some algebra, the posterior becomes

$$\sum_{J\in\vartheta}\prod_{k=1}^{K+1} \frac{\Gamma(\alpha_k+I_k+J_k)}{\Gamma(J_k+1)\times\Gamma(\alpha_k+I_k)} PDir(\alpha_k+I_k+J_k),$$

where *PDir* refers to the density function of Dirichlet. □

We give a heuristic interpretation here as well. When α_k is a natural number, we have

$$\frac{\Gamma(\alpha_k+I_k+J_k)}{\Gamma(J_k+1)\times\Gamma(\alpha_k+I_k)} = \binom{\alpha_k+I_k+J_k-1}{J_k}, \tag{2}$$

or the number of all possible ways of choosing J_k from $\alpha_k+I_k+J_k-1$. For time interval *k*, one can regard α_k as the number of patients from a prior study and I_k as the number of patients who have experienced toxicity in the current study. The expression (2) is equivalent to the number of ways of adding J_k un-ordered objects to $\alpha_k-1+I_k+J_k$ positions, or put the J_k exchangeable patients experiencing toxicity during the remainder of the trial into $\alpha_k-1+I_k+J_k$ pre-determined positions. For a given realization, J_k is fixed for each interval, hence the total number of possible combinations for each realization is the product of (2) over $K+1$ intervals. Note that within each time interval,

only the relative order in which toxicity events occur is used in the calculation. The actual times are not used.

In the following discussion define $\alpha^{(K)} = \sum_{i=1}^{K} \alpha_i$ and similarly $I^{(K)}$ and $J^{(K)}$. Under the Dirichlet-Multinomial setting, $p = p_1 + \ldots + p_K \mid \boldsymbol{\alpha}$ is a random variable following $Beta(\alpha^{(K)}, \alpha_{K+1})$ distribution and the conditional distribution of the number of patients experiencing toxicity is $X \mid n, p \sim binomial\ (n, p)$, where n is the number of enrolled patients. Furthermore, the posterior distribution of p is a mixture of $Beta(\alpha^{(K)} + I^{(K)} + J^{(K)},\ \alpha_{K+1} + n - I^{(K)} + J^{(K)})$ with probability proportional to $A(\boldsymbol{J})$, where the Beta distribution arises from summarizing the first K dimensions of the Dirichlet random variables in the previous paragraph. Hence, the stopping criterion is calculated as

$$P\left(p > \theta \mid Data\right) = \int_{\theta}^{1} \sum \frac{A(\boldsymbol{J})}{\sum A(\boldsymbol{J})} \beta(p \mid \alpha^{(K)} + I^{(K)} + J^{(K)}, \alpha_{K+1} + n - I^{(K)} - J^{(K)}) dp, \quad (3)$$

where $\beta(p)$ is the density function of the Beta distribution, and we can choose $\alpha^{(K)} = m\theta$ and $\alpha_{K+1} = m(1-\theta)$ as in Section 4.4. Let $i(k)$ be the indicator that patient i is currently in the kth interval. There are $K+1-i(k)$ possible results by the end of trial for patient i, i.e., experiences toxicity at k, k + 1, …, K or completes trial at interval $K+1$. For all r patients under observation, there are $Q = \prod_{i=1}^{R} \left(K+1-i\left(k\right)\right)$ possible trial results. Note that there is another error in Follmann et al. (1999) where they mentioned Q summands will be needed in (3) as there are Q possible $\boldsymbol{J}$s. Because patients are exchangeable and the order of patients does not play any role in (3) except for calculating $A(\boldsymbol{J})$, one needs to subtract the number of possible repeated scenarios from

the number of summands (e.g., if there was one patient experienced toxicity in the whole trial, we do not need to distinguish whether it was patient A or patient B). To obtain this number one can obtain all enumeration of possible results for each patient and then combine repeated scenarios. As r gets large, the calculation of this formula becomes intractable, therefore Follmann et al. (1999) proposed to use a data augmentation method which we adopt when Q is larger than 1000. Recall that we used the prior $Beta(m\theta, m(1-\theta))$ in Section 4.3 when $K = 2$. If all K time intervals are of equal length, a reasonable prior for $K > 2$ is a Dirichlet prior $Dir(m\theta / K, m\theta / K, ..., m\theta / K, m(1-\theta))$.

4.5 Enrollment rule for trials with long follow-up

It is reasonable to assume that in trials with a long follow-up, one will check whether to stop the trial or not for excessive toxicity after a toxicity is observed or before a new patient is enrolled. When the decision whether to stop the study is based on partial data, the probability of stopping the trial will depend on the accrual rate as well as on the true toxicity rate. This is because there are fewer opportunities to look at the data in a fast accruing trial. If all patients are fully followed before the next patient is enrolled a trial can be stopped if the first 3 patients have toxicities. This is not the case if all N patients are enrolled in the trial at once. The stopping rule will be applied using b_N and it is not possible to stop the trial if 3 toxicities are observed in three patients while the rest of the patients are still being followed. For example, in a trial with $N = 30$, and a stopping rule with $Beta(0.6,2.4)$ prior, the probability of stopping a trial where all patients are fully followed before new patients are enrolled is 0.05 when $p = 0.2$, and 0.31 when $p = 0.3$.

At the same time, a fast enrolling trial with the same boundary yields the probability of stopping of 0.01 when $p = 0.2$, and 0.17 when $p = 0.3$.

As the rate of accrual is hard to predict, it is hard to control the probability of stopping by modifying the stopping rule, for example, by changing c in (3). We propose to use an enrollment rule to maintain a desirable probability of stopping the trial. An enrollment rule will also prevent the possibility of observing excessive toxicity in trials with fast accrual. The most conservative enrollment rule is not to enroll more patients than those allowed by the stopping rule. For example, in the Bayesian stopping rule with prior *Beta*(0.6, 2.4) (Table 4.1), the earliest the trial can be stopped is when 3 patients are accrued and 3 toxicities are observed. Therefore up to 3 patients can be enrolled at the beginning of the study. Further, the maximum number of enrolled patients is such that the number of patients with toxicities plus the number of patients currently under follow-up, $r + x$, does not exceed b_n. That is, the maximum number of patients to enroll, v, is the largest integer that meets the following criteria:

$$\int_{\theta}^{1} \beta\left(p \mid \alpha^{(K)} + r + x + v, \alpha_{K+1} + s\right) dp \le c \ . \tag{4}$$

The number of toxicities observed in a trial with long follow-up that uses this enrollment rule will be exactly the same as in a trial with immediate response. However, this rule will result in a very long trial. To shorten the trial, we propose to use a different prior in enrollment rule (4) compared to the stopping rule (3). For example, use $\alpha^{(K)} = m\theta$, $\alpha_{K+1} = m(1-\theta)$ with $m = 3$ in (3) and $m = 12$ in (4). This enrollment rule yields cautious enrollment in the beginning of the trial, then, enrollment is guided by the data as the data from the trial overwhelm the prior.

Table 4.2. Values c to generate the stopping boundary with priors *Beta*(3θ, $3(1-\theta)$) and Beta(12θ, $12(1-\theta)$) , θ = 0.1, 0.2, 0.3 for various values of the total sample size *N*

	$\theta = 0.1$		$\theta = 0.2$		$\theta = 0.3$	
Beta prior	(0.3,2.7)	(1.2,10.8)	(0.6,2.4)	(2.4,9.6)	(0.9,2.1)	(3.6,8.4)
N						
15	0.962	0.930	0.965	0.922	0.971	0.924
16	0.962	0.930	0.969	0.927	0.973	0.931
17	0.967	0.934	0.969	0.934	0.975	0.932
18	0.970	0.941	0.973	0.933	0.974	0.938
19	0.970	0.945	0.977	0.936	0.974	0.944
20	0.970	0.946	0.977	0.945	0.976	0.944
21-22	0.971	0.947	0.978	0.945	0.976	0.944
23	0.974	0.947	0.978	0.947	0.980	0.948
24-26	0.973	0.949	0.978	0.947	0.981	0.951
27-29	0.974	0.954	0.978	0.953	0.982	0.955
30-32	0.975	0.958	0.981	0.955	0.984	0.957
33-35	0.977	0.960	0.981	0.961	0.984	0.960
36-39	0.978	0.965	0.982	0.962	0.985	0.965
40-43	0.977	0.965	0.982	0.963	0.985	0.966
44-49	0.980	0.969	0.984	0.967	0.986	0.970
50-51	0.981	0.970	0.986	0.971	0.987	0.972
52-60	0.983	0.972	0.985	0.972	0.987	0.972

4.6 Simulation results and example

Our simulation study illustrates the performance of the enrollment rule and the ability to stop the trial using partial data. We simulate a phase II trial with tolerable toxicity rate $\theta = 0.2$ and sample size N = 30. The constant c = 0.981 yields the probability of stopping of 0.05 when θ= 0.2 in a trial with immediate response (Table 4.2). In the simulation study we assume that an unlimited number of patients are available for enrollment at any time. Time to toxicity, given that toxicity has occurred in (0, t^*),

follows a uniform distribution on (0, t^*). Simulation results are based on averaging over 10,000 runs.

Table 4.3 presents the relationship between toxicity rate, study length and average observed toxicity for selected values of *m*. These relationships are also illustrated graphically in Figures 4.1 and 4.2. As *m* defining $\alpha^{(K)}$ and α_{K+1} in (4) gets larger, the trial length gets shorter, and at the same time the probability of stopping the trial for a given true toxicity rate decreases and the expected number of observed toxicities increases. When *m* = 3, the trial is long, but the maximum expected number of toxicities in the trial is minimized and the probability of stopping the trial when the true toxicity rate is 0.2 is controlled at 0.05. When *m* = 713, all 30 patients may be enrolled at once at the beginning of the study, that is, enrollment is not controlled. The trial time is short and the probability of stopping the trial for excessive toxicity is only 0.01 when the true toxicity rate is 0.2, the same as with no enrollment rule. We propose setting parameter *m* in the enrollment rule to a value that yields an increase in the maximum expected number of toxicities of no more than a certain desired percentage, e.g. 10% or 20%. For example, when *m* = 12 in the enrollment rule there is a 9.7% increase in the maximum expected number of toxicities compared to *m* = 3, the increase is 18.9% when *m* = 15, and 25.0% when *m* = 30. Therefore, if we allow at most a 10% increase in the maximum number of expected toxicities, the enrollment rule with *m* = 12 should be used.

Table 4.3. Probability of stopping (η'), expected number of patients enrolled (E(n)), expected number of toxicities (E(X)), expected duration in days (E(L)) and expected patient days (E(NT)) under the enrollment rule with various m for $N = 30$, $c = 0.981$, $K = 4$, $t^* = 28$ days and the probability of stopping of $\eta = 0.05$ when $\theta = 0.2$

	True Toxicity Rate p	0	0.2	0.3	0.4	0.5	0.6
m=3	η	0	0.050	0.313	0.712	0.951	0.997
	E(n)	30	29.20	25.61	19.17	12.52	8.52
	E(X)	0	5.84	7.67	7.63	6.30	5.14
	E(L)	140	182	199	175	124	86
	E(NT)	840	742	622	448	286	193
m=12	η	0	0.043	0.293	0.706	0.943	0.996
	E(n)	30	29.38	26.22	19.90	13.61	9.64
	E(X)	0	5.87	7.85	7.89	6.71	5.64
	E(L)	112	144	151	126	89	61
	E(NT)	840	747	636	464	308	215
m=15	η	0	0.040	0.297	0.691	0.944	0.995
	E(n)	30	29.46	26.36	20.65	14.60	10.83
	E(X)	0	5.92	7.95	8.10	7.06	6.11
	E(L)	112	136	141	116	80	55
	E(NT)	840	749	638	481	329	238
m=30	η	0	0.035	0.283	0.693	0.941	0.996
	E(n)	30	29.567	26.81	21.29	15.58	11.54
	E(X)	0	5.92	7.95	8.30	7.38	6.42
	E(L)	84	118	116	93	65	44
	E(NT)	840	751	650	494	350	252
m≥713	η	0	0.012	0.175	0.565	0.903	0.988
	E(n)	30	30	30	30	30	30
	E(X)	0	6.01	8.97	11.32	12.83	13.52
	E(L)	28	28	28	26	23	20
	E(NT)	840	762	722	685	645	607

Figure 4.1. Expected number of toxicities E(X) vs. true toxicity rate in a trial with $N = 30$ patients, for different values of m defining a and b in enrollment rule (4). When $m = 3$, the trial is the same as a fully sequential trial and when $m = 713$, enrollment rule allows to enroll all the patients at once.

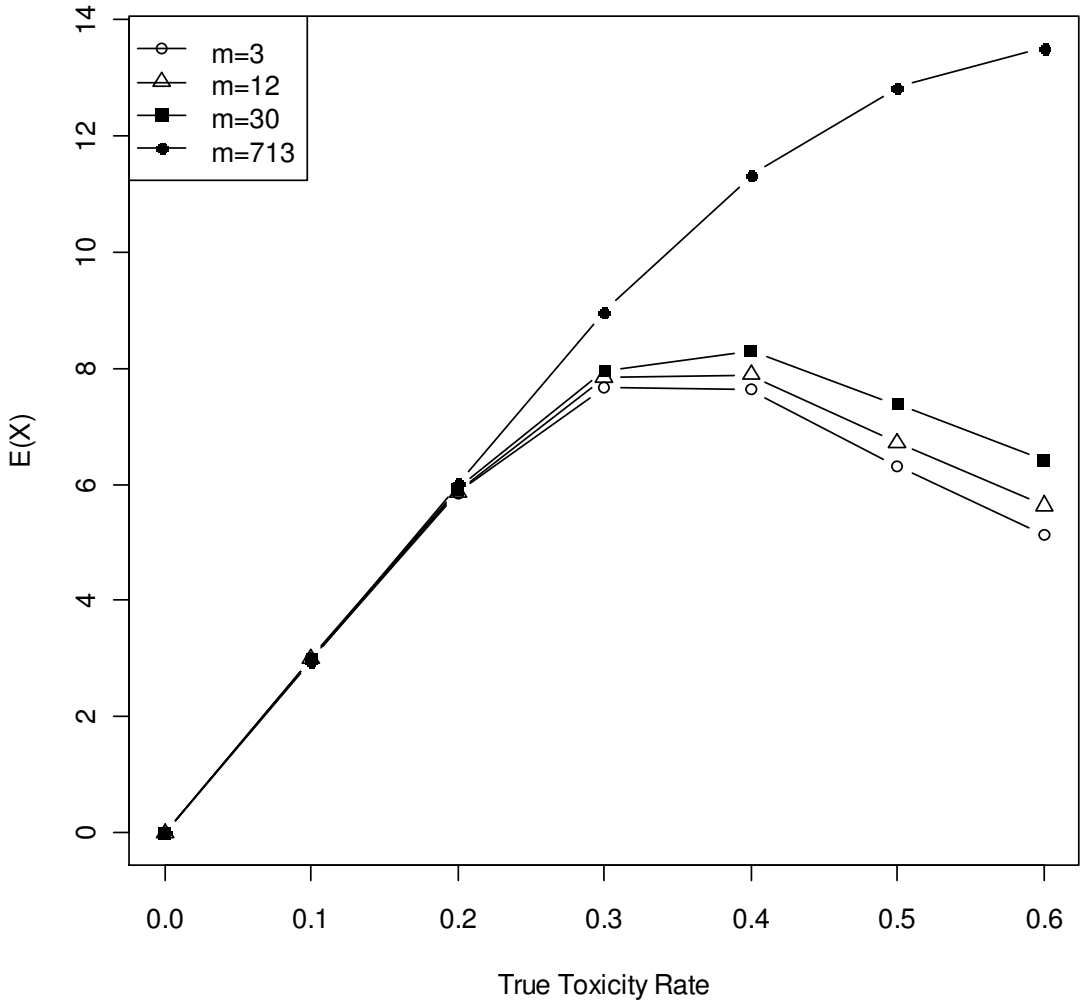
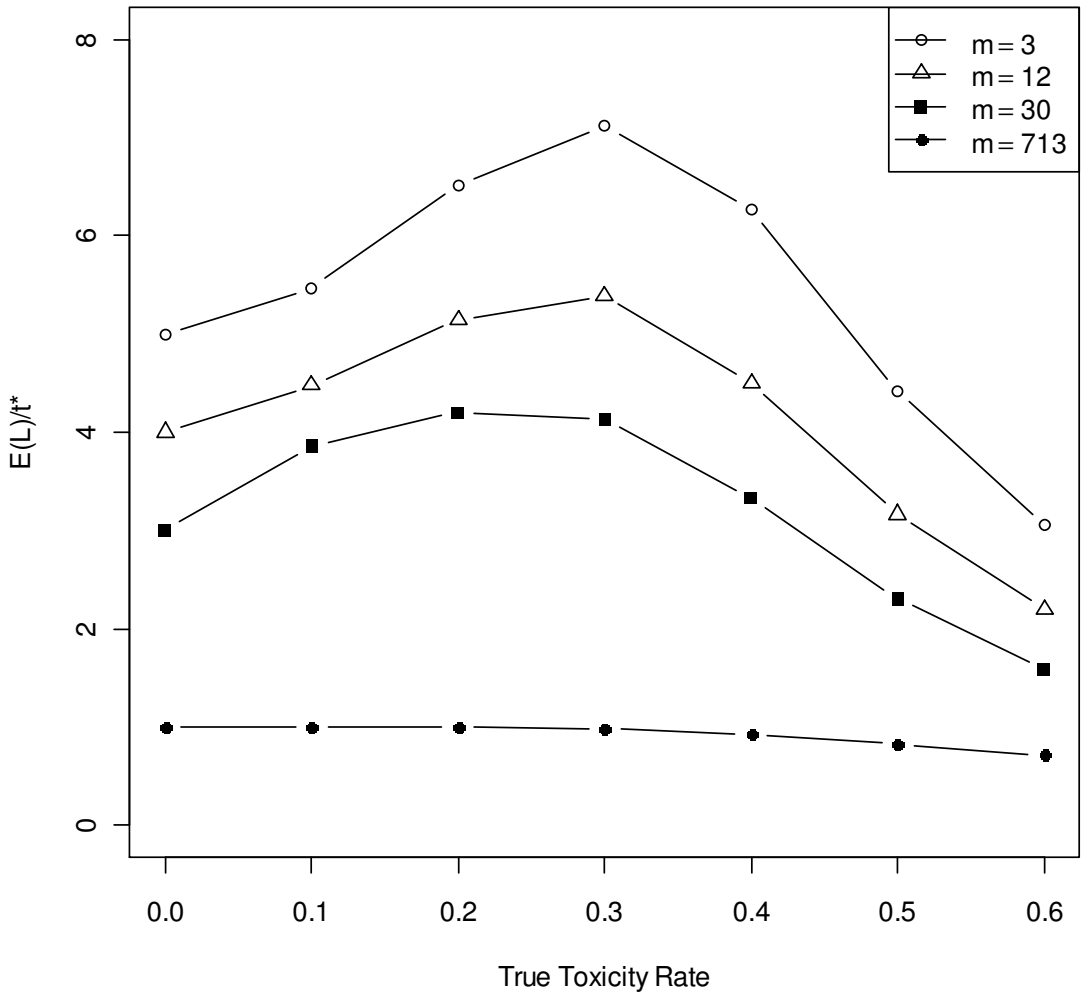


Figure 4.2. Expected length in terms of follow-up time for toxicity $T\,\mathrm{E}(L)/\,t^{*}$ vs. true toxicity rate in a trial with $N = 30$ patients for different values of m in enrollment rule (4). When $m = 3$, the trial is the same as a fully sequential trial and when $m = 713$, enrollment rule allows to enroll all the patients at once.



Example. The proposed methodology was used for enrollment in the LCCC clinical trial addressing the safety of a single dose of a new MEK inhibitor in patients with distinct molecular metastatic melanoma subtypes. In this trial, $N = 60$, $\theta = 0.2$ and $c = 0.987$. The follow-up time for toxicity was 8 weeks and therefore it was convenient to set $K = 8$. The enrollment rule with $m = 15$ was used yielding the maximum toxicity increase compared to a fully sequential trial of about 10%. The maximum occurs when the true toxicity rate is 0.5. Under this enrollment rule the probability of stopping the trial using the Bayesian stopping rule based on partial data is 0.045 when $p = 0.2$. When $p = 0.2$ and assuming that patients are always available to enroll, the expected duration of the trial is 61.3 weeks. When $p = 0.3$, the time increases to 64.4 weeks. When $p = 0.5$, the expected duration is only 24.9 weeks because most trials stop early for excessive toxicity. For comparison, the trial with conservative enrollment with $m = 3$ yields an average length of 86 weeks when $p = 0.3$ and 37.2 weeks when $p = 0.5$, much longer duration than a trial with $m = 15$.

4.7 Discussion

Using a stopping rule for toxicity based on partial data alone may not prevent occurrence of excessive toxicity in a phase II oncology trial with a long follow-up for toxicity. If accrual is fast compared to follow-up length, the probability of stopping might be lower than in a slow accruing trial. An enrollment rule used in conjunction with a stopping rule keeps the probability of stopping close to the required level and guides the rate of enrollment. With the enrollment rule excessive toxicity can be controlled without substantial lengthening of the duration of the trial. A similar enrollment strategy can be

used in a two-arm phase II oncology trial where observing toxicity higher than tolerable level is a concern. We provide design parameters for practical implementation of a Bayesian stopping rule for toxicity. R programs are available upon request. Since 1 - c, where c is the design parameter in the stopping rule (1), is very close to the design parameter in the Pocock boundary (Ivanova et al., 2005), one can obtain a good approximation of 1 – c by using software for the Pocock stopping boundary available at http://cancer.unc.edu/biostatistics/program/ivanova/.

BIBLIOGRAPHY

Alberts, S.R., Fitch, T.R., Kim, G.P., Morlan, B.W., Dakhil, S.R., Gross, H.M. et al. (2012). Cediranib (AZD2171) in patients with advanced hepatocellular carcinoma: a phase II North Central Cancer Treatment Group Clinical Trial. *American Journal of Clinical Oncology* **35**, 329-333.

Antoniak, C.E. (1974). Mixture of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152-1174.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9-26.

Bekele, B.N., Ji, Y, Shen, Y and Thall, P.F. (2008). Monitoring late-onset toxicities in phase I trials using predicted risks. *Biostatistics* **9**, 442-457.

Berry, D.A. (2006). Bayesian Clinical Trials. *Nature Reviews Drug Discovery* **5**, 27-36.

Berry, D.A. (2012). Adaptive clinical trials in oncology. *Nature Reviews Clinical Oncology* **9**, 199-207.

Blum, J. and Susarla, V. (1977). On the posterior distribution of a Dirichlet process given randomly right censored observations. *Stochastic Processes and Their Applications* **5**, 207-211.

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* **51**, 1372-1383.

Cerny, T., Pedrazzini, A., Joss, R.A. and Brunner K.W. (1988). Unexpected high toxicity in a phase II study of teniposide (VM-26) in elderly patients with untreated small cell lung cancer (SCLC). *European Journal of Cancer & Clinical Oncology* **24**, 1791-1794.

Cheung, Y.K. and Chappell, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56**, 1177-1182.

Clopper, C. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26,** 404-413.

Conaway, M.R. and Petroni, G.R. (1995). Bivariate sequential designs for phase II trials. *Biometrics* **51**, 656-664.

D'Adamo, D.R., Anderson, S.E., Albritton, K., Yamada, J., Riedel, E., Scheu, K. et al. (2005). Phase II study of doxorubicin and bevacizumab for patients with metastatic soft-tissue sarcomas. *Journal of Clinical Oncology* **23**, 7135-7142.

Dawson, N.A., Halabi, S., Ou, S.S., Biggs, D.D., Kessinger, A., Vogelzang, N. et al. (2008). A phase II study of estramustine, docetaxel, and exisulind in patients with

hormone- refractory prostate cancer: results of cancer and leukemia group B trial 90004. *Clinical Genitourinary Cancer* **6**, 110-116.

El-Maraghi, R.H. and Eisenhauer, E.A. (2008). Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. . *Journal of Clinical Oncology* **26**,1346-1354.

Ensign, L.G., Gehan, E.A., Kamen, D.S. and Thall, P.F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine* **13**,1727-1736.

Etzioni, R. and Pepe, M.S. (1994). Monitoring of a pilot toxicity study with two adverse outcomes. *Statistics in Medicine* **13,** 2311-2321.

Freedman, L.S. and Spiegelhalter, D.J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials* **10**, 357-367.

Follmann, D. A. and Albert, P. S. (1999). Bayesian monitoring of event rates with censored data. *Biometrics* **55**, 603–607.

Gibson, M.K., Catalano, P.J., Kleinberg, L., Staley, C.A., Montgomery, E., Song, W. et al. (2010). E2205: A phase II study to measure response rate and toxicity of neoadjuvant chemoradiotherapy (CRT) with oxaliplatin (OX) and infusional 5-fluorouracil (5-FU) plus cetuximab (C) followed by postoperative docetaxel (DT) and C in patients with operable adenocarcinoma of the esophagus. *ASCO Annual Meeting* Abstr 4064. Chicago, IL.

Goldman A.I. (1987). Issues in designing sequential stopping rules for monitoring side effects in clinical trials. *Control Clinical Trials* **8**, 327-337.

Geller, N.L., Follmann, D.F., Leifer, E.S. and Carter, S.L. (2005). Design of early trials in peripheral blood stem cell transplantation: a hybrid frequentist-Bayesian approach. In *Advances in Clinical Trial Biostatistics*, N. L. Geller (ed), 40-52. New York and Basel: Marcel Dekker.

Hoering, A., LeBlanc, M. and Crowley, J. (2011). Seamless phase I-II trial design for assessing toxicity and efficacy for targeted agents. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 640-646.

Ivanova, A., Flournoy, N., and Chung, Y**.** (2007). Cumulative cohort design for dose-finding. *Journal of Statistical Planning and Inference* **137**, 2316-2317.

Ivanova, A., Qaqish, B.F. and Schell, M.J. (2005). Continuous toxicity monitoring in phase I trials in oncology. *Biometrics* **61**, 540-545.

Jennison, C. and Turnbull, B.W. (2000). *Group sequential methods with applications to clinical trials.* London, New York: Chapman and Hall/CRC.

Johnson, E., Lake, D., Herndon, J.E. 2nd, Box, J.W., Lynch, T.J. and Green, M.R. (2004). Phase II trial of vinorelbine plus doxorubicin in relapsed small-cell lung cancer: CALGB 9332. *American Journal of Clinical Oncology* **27**, 19-23.

Jin, H. (2007). Alternative designs of phase II trials considering response and toxicity. *Contemporary Clinical Trials* **28**, 525-531.

Maki, R.G., D'Adamo, D.R., Keohan, M.L., Saulle, M., Schuetze, S.M., Undevia, S.D. et al. (2009). Phase II study of sorafenib in patients with metastatic or recurrent sarcomas. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 3133-3140.

Moschos, S. (2013) *PM003: A Phase I Study of Veliparib Combined with a B-RAF inhibitor plus Radiation Therapy [Stereotactic Radiosurgery (SRS) versus Whole Brain Irradiation Therapy (WBRT)] in Patients with Active Melanoma Brain Metastases*. Protocol Letter of Intent, Lineberger Comprehensive Cancer Center.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-556.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191-199.

Ray, H.E., Rai S.N. (2011). An evaluation of a Simon 2-Stage phase II clinical trial design incorporating toxicity monitoring. *Contemporary Clinical Trials* **32**, 428-436.

Ray, H.E., Rai S.N. (2013). Flexible bivariate phase II clinical trial design incorporating toxicity and response on different schedules. *Statistics in Medicine* **32,** 470-485.

Reiner, E., Paoletti, X. and O'Quigley, J. (1999). Operating characteristics of the standard phase I clinical trial design. *Computational Statistics and Data Analysis* **30**, 303-315.

Rosner, G.L. (2005). Bayesian monitoring of clinical trials with failure-time endpoints. *Biometrics* **61**, 239-245.

Rubinstein L., Crowley J., Ivy, P., Leblanc M. and Sargent D. (2009). Randomized phase II designs. *Clinical Cancer Research* **15**, 1883-1890.

Schmegner, C. and Baron, M. (2004). Principles of optimal sequential planning. *Sequential Analysis* **23**, 11-32.

Schultz J.R., Nichol, F.R, Elfring, G.L. and Weed S.D. (1973). Multiple-stage procedures for drug screening. *Biometrics* **29**, 293-300.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1-10.

Storer, B.E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45**, 925-37.

Tanner, YI. A. (1992). *Tools for statistical inference*. Berlin:Springer-Verlag.

Thall, P.F. and Cheng, S.C. (1999). Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics* **55**,746-753.

Tournoux , C., De Rycke, Y., Medioni, J. and Asselain, B. (2007). Methods of joint evaluation of efficacy and toxicity in phase II clinical trials. *Contemporary Clinical Trials* **28**, 514-524.

Yin, G. (2012) *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods.* New Jersey: John Wiley & Sons.

Yu, J., Hutson, A.D., Siddiqui, A.H. and Kedron, M.A. (2012). Group sequential control of overall toxicity incidents in clinical trials—non-Bayesian and Bayesian approaches. *Statistical Methods in Medical Research*, doi: 10.1177/0962280212440535.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16**, 117-186.

Wyman, K., Atkins, M.B., Prieto, V., Eton, O., McDermott, D.F., Hubbard, F. et al. (2006). Multicenter Phase II trial of high-dose imatinib mesylate in metastatic melanoma: significant toxicity with no clinical efficacy. *Cancer* **106**, 2005-2011.