DOI: 10.1111/j.1541-0420.2005.00311.x

CONSULTANT'S FORUM

Continuous Toxicity Monitoring in Phase II Trials in Oncology

Anastasia Ivanova,*[*] Bahjat F. Qaqish, and Michael J. Schell

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill,
North Carolina 27599-7420, U.S.A.
[*] *email:* aivanova@bios.unc.edu

Summary. The goal of a phase II trial in oncology is to evaluate the efficacy of a new therapy. The dose investigated in a phase II trial is usually an estimate of a maximum-tolerated dose obtained in a preceding phase I trial. Because this estimate is imprecise, stopping rules for toxicity are used in many phase II trials. We give recommendations on how to construct stopping rules to monitor toxicity continuously. A table is provided from which Pocock stopping boundaries can be easily obtained for a range of toxicity rates and sample sizes. Estimation of the probability of toxicity and response is also discussed.

Key words: Phase II trial; Pocock boundary; Sequential monitoring; Unbiased estimate.

1. Introduction

The primary purpose of a phase I clinical trial is to find a dose with the probability of toxicity equal to the maximum-tolerated level (often 0.2 or 0.25), called the "maximum-tolerated dose" (MTD). The response rate of the dose established in a phase I trial is evaluated in subsequent phase II trials. Because phase I trials use small sample sizes, the estimate of the MTD is imprecise. As a result, the dose chosen for the phase II trial can have a toxicity rate that is much higher than the maximum-tolerated level. To avoid an excessive number of toxicities, a stopping rule for toxicity is often implemented in the phase II trials. One example is a study of mitoxantrone or floxuridine in patients with minimal residual ovarian cancer after the second-look laparotomy conducted by the Southwest Oncology Group (Muggia et al., 1996). The study plan was to assign 37 patients to each arm, although accrual would be stopped in the arm if 13 or more of the first 20 patients on that arm experienced toxicity. Toxicity was defined as not being able to tolerate at least two courses of treatment. The stopping rule for toxicity uses the fact that the probability that 13 or more of 20 will not tolerate a treatment is less than 5% if the true proportion is 0.4 (Liu, 2001). Bryant and Day (1995) proposed a two-stage design, where the trial is terminated after the first stage if the observed toxicity rate is too high or the response rate too low. A similar strategy was adopted by Conaway and Petroni (1995), who considered two- and three-stage designs. However, when toxicity events are severe, continuous monitoring of them is preferred. That was the case for the Lineberger Comprehensive Cancer Center (LCCC) 9818 trial of taxol plus herceptin in patients with metastatic breast cancer, designed by one of us (MJS). Cardiac toxicity was the primary concern. It was decided that the cardiac toxicity should be monitored continuously throughout the trial and the trial stopped as soon as there is evidence that the toxicity rate is much higher than that of historic controls. The total planned sample size for the trial was 60. The stopping boundary designed for the trial yielded a probability of stopping the trial of 0.05 if the true rate of the cardiac toxicity was 0.09.

In this article, we investigate the Pocock and O'Brien–Fleming boundaries as possible stopping boundaries for toxicity. It is known that maximum likelihood estimates may be biased when a sequential procedure is used (Whitehead, 1986). We evaluate the bias of the maximum likelihood estimates of toxicity and response rates and suggest alternative ways of estimation. Section 2 addresses computation of the boundary. Sections 3 and 4 address the estimation of toxicity and response in trials where stopping for toxicity is possible. In Section 5, we draw conclusions.

2. Designing a Stopping Boundary for Toxicity

Let K be the sample size planned for a single-stage phase II study. Let θ denote the true toxicity rate of the dose chosen for the study. Our goal is to construct a stopping boundary based on toxicity such that the probability of early stopping is at most ϕ if the toxicity rate is equal to θ_0. The values of θ_0 and ϕ are elicited from the principal investigator of the trial. One possible choice of θ_0 is the maximum-tolerated toxicity rate, which is the probability of toxicity of the true MTD. Because we do not want the probability of early stopping to be high when the toxicity rate is equal to the maximum-tolerated toxicity rate, it is reasonable to choose $\phi = 0.05$. To monitor toxicity continuously, we need to specify the stopping boundary for each k, $k = 1, \ldots, K$. We will investigate the Pocock (1977)

Table 1

Pocock and O'Brien–Fleming boundaries for $K = 20$ that yield probability of stopping of about $\phi = 0.05$ when the true toxicity rate is $\theta_0 = 0.2$

k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Pocock boundary																				
b_k	–	–	3	4	4	4	5	5	5	6	6	6	7	7	7	8	8	8	9	9
O'Brien–Fleming boundary																				
b_k	–	–	–	–	–	6	6	6	6	6	6	7	7	7	7	7	7	8	8	8

and O'Brien–Fleming (1979) boundaries with K stages as possible stopping boundaries for toxicity. Note that because the maximum sample size is fixed, use of open-ended tests, such as the sequential probability ratio test (SPRT), is not appropriate. A boundary is a sequence of integers $(b_1, \ldots, b_K)$. If the number of toxicities in the first k patients is equal to or higher than b_k, the trial is stopped and we refer to this event as early stopping (even though it may occur after all K patients have been treated, if the number of toxicities is b_K). In order to find a boundary, a set of pointwise probabilities $\alpha_1, \ldots, \alpha_K$ are chosen, and b_k is then computed as the smallest integer such that $\Pr\{Y \geq b_k\} \leq \alpha_k$, where Y denotes a binomial variate with parameters k and θ_0. The choice of the α_k's distinguishes the different types of boundaries. A Pocock boundary is obtained by setting $\alpha_1 = \cdots = \alpha_K = \alpha$, where α is such that if $\theta = \theta_0$ the probability of early stopping is as close to ϕ as possible, but not exceeding ϕ. The solution to this problem, then, centers around identification of the α's. For the Pocock boundary, a possible initial estimate of α is $\alpha = \Pr\{Z > C_P(K, \phi)\}$, where $C_P(K,\ \phi)$ are the tabulated values for the normally distributed outcomes (Jennison and Turnbull, 2000, p. 26), and Z denotes a standard normal variate. For the O'Brien–Fleming boundary, an initial estimate is obtained by choosing $\alpha_k = \Pr\{Z > C_B(K, \phi)(K/k)^{1/2}\}$, where $C_B(K, \phi)$ are the tabulated values (Jennison and Turnbull, 2000, p. 29). The discreteness of the binomial distribution requires fine-tuning of these boundaries to bring the overall probability of early stopping as close to ϕ as possible. Jennison and Turnbull (2000, p. 237) suggest that investigators "start with a procedure based on the normal approximation and find the precise error rates . . . by exact calculation." Although other boundaries and general spending-function arguments (Jennison and Turnbull, 2000) can be adapted in a similar fashion, we limit our discussion to these two types of boundaries.

Consider an example with $K = 20$, $\theta_0 = 0.2$, and $\phi = 0.05$. The two stopping boundaries are displayed in Table 1. The Pocock boundary is constructed with $\alpha = 0.01959$ that corresponds to $C_P(K,\ \phi) = 2.054$. The value of $C_P(K,\ \phi)$ for $K = 20$ and the error rate of 0.05 given by Jennison and Turnbull (2000, p. 26) is 2.672. The probability of early stopping is $\phi = 0.0484$. We used $C_B(K,\ \phi) = 1.854$ for the O'Brien–Fleming boundary that gives $\phi = 0.0481$. Note that because of discreteness of the binomial distribution, other choices of $C_P(K,\ \phi)$ and $C_B(K,\ \phi)$ from a small interval around the values given above will produce the same boundaries.

Figure 1 displays the probability of early stopping plotted against toxicity rate for both the Pocock and O'Brien–Fleming boundaries. By construction, the two curves coincide at approximately $\theta = 0.2$ and $\phi = 0.05$. The O'Brien–Fleming boundary yields slightly higher probability of early stopping for higher toxicity rates. However, the average sample size in the trial with high probability of toxicity is larger for the O'Brien–Fleming method compared to the Pocock boundary (Figure 1). This is because the Pocock boundary is at least as likely to stop the trial for every given number of patients except $k = 19$ or 20 (Table 1). The average number of toxicities in the trial is higher for the O'Brien–Fleming boundary (Figure 1) as well. It is also interesting to note that for the Pocock boundary, we expect to see more toxicities if $\theta = 0.4$ than if $\theta = 0.5$. By construction, the α_k values are higher for the Pocock boundary than for the O'Brien–Fleming boundary at the beginning of the trial, allowing earlier stopping if the toxicity rate is high. Because we believe that it is important to stop the trial as early as possible if the toxicity rate is too high, we typically prefer the Pocock boundary.

Determination of α involves trial-and-error calculations from the values in the Jennison and Turnbull (2000) tables, as

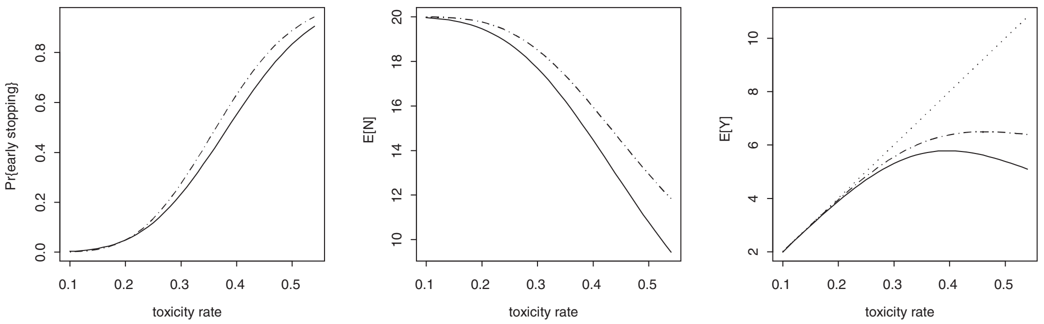

Figure 1. Probability of early stopping, expected sample size, and expected number of toxicities for different toxicity rates. Results for the Pocock boundary are shown by the solid line, O'Brien–Fleming boundary by the dashed line. Expected number of toxicities in a trial with no stopping boundary is shown by the dotted line.

Table 2

Values of α for constructing the Pocock boundaries that yield probability of stopping of about $\phi = 0.05$ when the true toxicity rate is θ_0 and the planned sample size is equal to K

$\theta_0 = 0.1$		$\theta_0 = 0.2$		$\theta_0 = 0.3$	
K	α	K	α	K	α
15–16	0.02685	15, 18–20	0.01959	15–17	0.02530
17–20	0.02566	16–17	0.02666	18–21	0.02162
21–22	0.02389	21–24	0.01941	22–24	0.02097
23–24	0.02238	25–26	0.01806	25–26	0.01823
25–26	0.01853	27	0.01734	27–29	0.01747
27	0.01822	28–30	0.01696	30–31	0.01694
28–31	0.01791	31	0.01629	32–33	0.01525
32	0.01701	32–34	0.01672	34–36	0.01426
33–37	0.01585	35–36	0.01629	37–40	0.01384
38–39, 44	0.01467	37–40	0.01487	41	0.01308
40–43	0.01550	41	0.01442	42–45	0.01215
45	0.01445	42, 44	0.01301	46	0.01166
46, 48–49	0.01403	43	0.01419	47–48	0.01133
47	0.01411	45–47	0.01272	49, 51	0.01130
50–51	0.01315	48	0.01263	50, 52–54	0.01116
52	0.01280	49–51	0.01167	55–56	0.01094
53–57	0.01273	52	0.01166	57–60	0.01073
58–60	0.01258	53–56	0.01161		
		57–59	0.01101		
		60	0.01098		

noted above. Table 2 gives α values that can be used to construct the Pocock boundary for $\phi = 0.05$, $\theta_0 = 0.1$, 0.2, and 0.3, and a wide range of values of K. For example, if $K = 25$, the Pocock boundary with $\alpha = 0.01806$ yields a probability of stopping of almost but not exceeding 0.05 when the true probability of toxicity is 0.2. From an α value in Table 2, boundary b_k can be computed as the smallest integer such that $\Pr\{Y \geq b_k\} \leq \alpha$, where Y is binomial(k, θ_0). The values of α can also be used as starting values when constructing boundaries with different ϕ, θ_0, and K. For example, we used the value of $\alpha = 0.01258$ corresponding to $K = 60$ and $\theta_0 = 0.1$ as the starting value to construct the Pocock boundary for the taxol plus herceptin trial mentioned in Section 1. The Pocock boundary for this trial with $K = 60$ and $\theta_0 = 0.09$ is presented in Table 3. The value $\alpha = 0.01100$ yielded $\phi = 0.0494$ (the original boundary in the LCCC 9818 trial was not a Pocock boundary, although it was similar in spirit).

3. Estimation of the Probability of Toxicity for Trials with Early Stopping for Toxicity

3.1 *Estimation of the Toxicity Rate*

The maximum likelihood estimator (MLE) in a trial where a sequential design is used is known to be biased. Girshick, Mosteller, and Savege (1946) developed unbiased estimators for binomial proportions under various sampling plans. Whitehead (1986) developed bias correction procedures based on the asymptotic normality of MLEs. Jung and Kim (2004) derived the unique uniformly minimum variance unbiased estimator of a proportion in a two-stage phase II design and compared it to the MLE in terms of bias and mean squared error. In this section, we develop an alternative to MLEs that can be used in trials where there is a possibility of early stopping for toxicity.

The outcome of a sequential trial is a pair, (Y, N), the numbers of toxicities and patients, respectively. For the stopping rules discussed here, the likelihood is proportional to a binomial likelihood (Lindsey, 1997) and the MLE is $\hat{\theta}_{\text{MLE}} = Y/N$. If the sample space of (Y, N) contains J points (with positive probability), then an estimator of θ can be expressed as a $J \times 1$ vector containing values of the estimator at each point in the sample space. Let β denote such a vector, and $\hat{\theta}_\beta$ be the associated estimator, $\hat{\theta}_\beta = \beta_j$ if the observed (Y, N) is the jth point in the sample space. The distribution of (Y, N), and hence $\hat{\theta}_\beta$, is completely determined by θ. For a given β, define the bias and the variance by

$$B(\theta, \beta) = \mathrm{E}[\hat{\theta}_\beta - \theta],$$
$$V(\theta, \beta) = \mathrm{var}(\hat{\theta}_\beta),$$

and define a penalty function

$$Q(\theta, \beta, \lambda) = \lambda V(\theta, \beta) + (1 - \lambda) B^2(\theta, \beta),$$

where $\lambda \in (0, 1)$ is an adjustable parameter chosen to achieve a balance between the bias and variability of $\hat{\theta}_\beta$. The function $Q(\theta\ , \beta\ , \lambda)$ depends upon the unknown θ. We propose placing a prior $g(\theta)$ on θ and minimizing the expected penalty

$$R(\beta, \lambda) = \int_a^b Q(\theta, \beta, \lambda) g(\theta)\, d\theta = \mathrm{E}_\theta[Q(\theta, \beta, \lambda)].$$

The interval $[a, b]$ represents a range of interest or a plausible range for θ, where the prior $g(\theta)$ is away from zero, for example, $[a, b] = [0, 1]$. The function $Q(\theta,\ \beta,\ \lambda)$ is the risk function, and $R(\beta,\ \lambda)$ is the Bayes' risk with respect to the prior $g(\theta)$. The estimator is found by minimizing $R(\beta,\ \lambda)$ with respect to β subject to $\beta_j \in [0, 1]$ for all j. Closed-form expressions are not possible, so we adopt a numerical approach. First, we approximate $g(\theta)$ by a discrete probability distribution that puts mass $g_i > 0$ at $\theta = x_i$, $i = 1, \ldots, m$, $\sum_{i=1}^m g_i = 1$. Here, $x_1 < \cdots < x_m$ are points in $[a, b]$. Then, $R(\beta,\ \lambda)$ is approximated by the summation

$$R(\beta, \lambda) \approx \tilde{R}(\beta, \lambda) = \sum_{i=1}^m g_i Q(x_i, \beta, \lambda). \qquad (1)$$

Table 3

Pocock boundary for $K = 60$ that yields probability of stopping $\phi = 0.05$ when the true toxicity rate is $\theta_0 = 0.09$. Only points where stopping is possible are listed.

k	2	4	5	6	8	9	10	12	13	14	15	16	18	19	20	21	22	24	25	26	27	28	29	31	32	33
b_k	2	3	3	3	4	4	4	5	5	5	5	5	6	6	6	6	6	7	7	7	7	7	7	8	8	8
k	34	35	37	38	39	40	41	42	44	45	46	47	48	49	50	52	53	54	55	56	57	59	60			
b_k	8	8	9	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	11	11	11	12	12			

The choice of m is not critical. We report some comparisons in Section 3.2. The mean and variance can be expressed in matrix terms as follows. Let the J-vector f_i contain the probability masses for the J points in the sample space evaluated at $\theta = x_i$. It follows that, at $\theta = x_i$, the mean and variance of $\hat{\theta}_\beta$ are

$$\mu_i = \sum_{j=1}^{J} f_{ij}\beta_j = f_i^{\mathrm{T}}\beta,$$

$$V(x_i, \beta) = \sum_{j=1}^{J} f_{ij}\beta_j^2 - \mu_i^2 = \beta^{\mathrm{T}} D_i \beta$$
$$- \left(f_i^{\mathrm{T}}\beta\right)^2 = \beta^{\mathrm{T}}\left\{D_i - f_i f_i^{\mathrm{T}}\right\}\beta,$$

where D_i is a $J \times J$ diagonal matrix with f_i on the diagonal. It follows that

$$\begin{aligned} Q(x_i, \beta, \lambda) &= \lambda V(x_i, \beta) + (1-\lambda)(\mu_i - x_i)^2 \\ &= \beta^{\mathrm{T}}\left\{\lambda D_i + (1-2\lambda) f_i f_i^{\mathrm{T}}\right\}\beta \\ &\quad - 2(1-\lambda)x_i f_i^{\mathrm{T}}\beta + (1-\lambda)x_i^2. \end{aligned}$$

Define F to be the $m \times J$ matrix with ith row f_i^{T}, and let $x = (x_1, \ldots, x_m)^{\mathrm{T}}$. Then, we have

$$\begin{aligned} \tilde{R}(\beta, \lambda) &= \sum_{i=1}^{m} g_i Q(x_i, \beta, \lambda) \\ &= \sum_{i=1}^{m} g_i \beta^{\mathrm{T}}\left\{\lambda D_i + (1-2\lambda) f_i f_i^{\mathrm{T}}\right\}\beta \\ &\quad - 2(1-\lambda)\sum_{i=1}^{m} g_i x_i f_i^{\mathrm{T}}\beta + (1-\lambda)\sum_{i=1}^{m} g_i x_i^2, \\ &= \beta^{\mathrm{T}}\left\{\lambda \sum_{i=1}^{m} g_i D_i + (1-2\lambda)\sum_{i=1}^{m} g_i f_i f_i^{\mathrm{T}}\right\}\beta \\ &\quad - 2(1-\lambda)\sum_{i=1}^{m} g_i x_i f_i^{\mathrm{T}}\beta + (1-\lambda)\sum_{i=1}^{m} g_i x_i^2, \\ &= \beta^{\mathrm{T}}\{\lambda D + (1-2\lambda)F^{\mathrm{T}} G F\}\beta \\ &\quad - 2(1-\lambda)x^{\mathrm{T}} G F \beta + (1-\lambda)x^{\mathrm{T}} G x. \end{aligned}$$

Here, G is a diagonal matrix with diagonal elements $(g_1, \ldots, g_J)$, and matrix D is defined as $D = \sum_{i=1}^{m} g_i D_i$. The estimator β is the minimizer of

$$\beta^{\mathrm{T}}\{\lambda D + (1-2\lambda)F^{\mathrm{T}} G F\}\beta - 2(1-\lambda)x^{\mathrm{T}} G F \beta$$

subject to $0 \leq \beta \leq 1$ component-wise. Note that the term $(1-\lambda)x^{\mathrm{T}} G x$ is dropped because it does not involve β. Because $\lambda\, D + (1-2\lambda)F^{\mathrm{T}} G F$ is strictly positive definite (see the proof in the Appendix), the minimization problem above is a convex quadratic program and the minimizer can be found using standard quadratic programming methods (for example, Fletcher, 1987, Chapter 10). Note that the only approximation in the algorithm above involves replacing the integral in $R(\beta,\ \lambda)$ by the summation in $\tilde{R}(\beta, \lambda)$. The distributions f_i are computed exactly for small K. For larger K, f_i can be approximated by simulating a large number of sequences.

3.2 *Prior Elicitation and Examples*

If there is no information on θ, the prior $g(\theta)$ can be taken as uniform on $[0, 1]$. Often, the estimate of the MTD obtained in the phase I trial is chosen as the dose for a subsequent phase II trial. Hence, data from the phase I trial can be used to construct the prior. Then, a beta prior of the form $g(\theta) \propto \{\theta^{Y^0}(1-\theta)^{K^0 - Y^0}\}^w$ can be used, where Y^0 is the number of toxic observations, $K^0 - Y^0$ is the number of nontoxic observations in the phase I trial, and the weight w, $0 < w \leq 1$, reflects the influence of the prior on the current study (Legedza and Ibrahim, 2001). For example, if we have data from six patients with one toxic and five nontoxic outcomes from a phase I trial where a similar patient population was used, we set $g(\theta) \propto \theta(1-\theta)^5$.

The above approach was applied using the Pocock boundary from Table 1. Figure 2 shows the bias, standard deviation, and root mean squared error (RMSE) for the MLE, and estimators β corresponding to the uniform prior on $[0, 1]$ and $\lambda = 0.25$, denoted $\beta(0.25)$, and $\lambda = 0.5$, denoted $\beta(0.5)$. The value $\lambda = 0.5$ defines squared error loss, and the corresponding estimator is the posterior mean. The three estimators have comparable bias, but the Bayes' estimates have smaller variance. The RMSE values for both $\beta(0.25)$ and $\beta(0.5)$ are smaller than for the MLE for the toxicity rates above 0.1.

Table 4 displays the MLE of θ (computed as Y/N), the estimates $\beta(0.25)$, and $\beta(0.50)$ for uniform prior on $[0, 1]$,

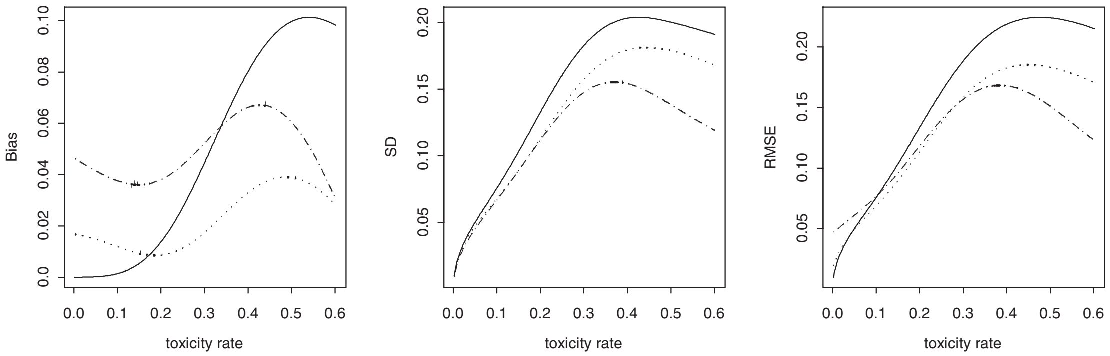

Figure 2. Bias, standard deviation (SD), and RMSE for MLE (solid line), $\beta(0.25)$ (dotted line), and $\beta(0.50)$ (dot–dash line).

Table 4

A phase II trial with 20 patients where the Pocock boundary is used to monitor toxicity. Here, Y is the number of toxicities and N is the number of patients before the trial is terminated. The table presents the probability of obtaining the outcome (Y, N) given θ, p_θ, the MLE of θ calculated as Y/N, the estimates $\beta(0.25)$, and $\beta(0.50)$ for the uniform prior, and estimates $\beta(0.25)$, and $\beta(0.50)$ for the beta prior.

						Uniform prior		Beta prior	
j	Y	N	$p_{0.2}$	$p_{0.5}$	MLE	$\beta(0.25)$	$\beta(0.50)$	$\beta(0.25)$	$\beta(0.50)$
1	3	3	0.01	0.12	1.00	0.80	0.88	0.55	0.45
2	4	5	0.00	0.09	0.80	0.71	0.74	0.55	0.46
3	4	6	0.01	0.14	0.67	0.63	0.61	0.49	0.43
4	5	8	0.00	0.07	0.63	0.60	0.57	0.51	0.44
5	5	9	0.01	0.10	0.56	0.55	0.50	0.46	0.41
6	6	11	0.00	0.05	0.55	0.54	0.49	0.47	0.42
7	6	12	0.01	0.07	0.50	0.50	0.45	0.44	0.40
8	7	14	0.00	0.04	0.50	0.50	0.44	0.45	0.40
9	7	15	0.00	0.05	0.47	0.47	0.41	0.42	0.39
10	8	17	0.00	0.03	0.47	0.47	0.41	0.43	0.40
11	8	18	0.00	0.04	0.44	0.44	0.39	0.41	0.38
12	9	20	0.00	0.02	0.45	0.45	0.39	0.42	0.39
13	0	20	0.01	0.00	0.00	0.05	0.02	0.03	0.07
14	1	20	0.06	0.00	0.05	0.09	0.07	0.07	0.11
15	2	20	0.14	0.00	0.10	0.14	0.11	0.13	0.14
16	3	20	0.21	0.00	0.15	0.18	0.16	0.16	0.18
17	4	20	0.22	0.00	0.20	0.23	0.20	0.20	0.21
18	5	20	0.17	0.01	0.25	0.28	0.24	0.24	0.25
19	6	20	0.10	0.03	0.30	0.32	0.28	0.28	0.29
20	7	20	0.04	0.06	0.35	0.36	0.31	0.33	0.32
21	8	20	0.01	0.05	0.40	0.40	0.35	0.37	0.36

and estimates $\beta(0.25)$, and $\beta(0.50)$ for the beta prior $g(\theta) \propto \theta(1-\theta)^5$. The Bayesian estimates are less extreme than the MLE. The estimates obtained using a beta prior are closer to its prior mean of 0.25 compared to corresponding estimates from a uniform prior. The conclusions regarding the bias of the MLE of toxicity are similar for the boundary with $K = 60$ and $\theta_0 = 0.09$ (Table 3) compared to the boundary with $K = 20$ and $\theta_0 = 0.2$ (Table 1). The estimates reported in Table 4 were obtained using $m = 500$ in (1). For comparison, the maximum change in β was 0.004 with $m = 50$; 0.002 with $m = 100$; and 0.0005 with $m = 250$.

4. Estimation of the Probability of Response

In this section, we address the concern that when using a toxicity-based sequential stopping rule in phase II trials, the observed proportion of responses (the standard estimator of the probability of a therapeutic response to a drug) may be biased. To address this issue, we specify the joint distribution of toxicity and response. For the ith patient, let the random variables Y_i and X_i, respectively, indicate toxicity and response, both coded as 1 = yes and 0 = no. Further, let $\pi = \mathrm{E}[X_i] = P\{X_i = 1\}$, $\rho = \mathrm{corr}(X_i, Y_i)$, and $X = \sum_1^N X_i$ be the number of responses after N patients. It then follows that

$$\mathrm{E}[X_i \mid Y_i] = \pi + \rho\left\{\frac{\pi(1-\pi)}{\theta(1-\theta)}\right\}^{1/2}(Y_i - \theta),$$

$$\mathrm{E}\left[\frac{X}{N} - \pi \mid (Y, N)\right] = \rho\left\{\frac{\pi(1-\pi)}{\theta(1-\theta)}\right\}^{1/2}\left(\frac{Y}{N} - \theta\right),$$

and, by unconditioning, the bias of X/N is

$$\mathrm{E}\left[\frac{X}{N} - \pi\right] = \rho\left\{\frac{\pi(1-\pi)}{\theta(1-\theta)}\right\}^{1/2}\mathrm{E}\left[\frac{Y}{N} - \theta\right]. \qquad (2)$$

A similar formula was derived by Whitehead (1986) based upon the asymptotic normality of MLEs. However, in the current context, formula (2) is exact for any sample size. To maximize (2) with respect to ρ we note that for fixed θ and π, the upper bound on ρ is min $(r, 1/r)$, where $r = \{\theta/(1-\theta)\}^{1/2}/\{\pi/(1-\pi)\}^{1/2}$ (Qaqish, 2003). Under restriction $\pi > \theta$ and $\rho > 0$, we obtain the bound

$$\frac{\mathrm{E}[X/N - \pi]}{\mathrm{E}[Y/N - \theta]} \leq \frac{1-\pi}{1-\theta}.$$

This has the intuitive interpretation that the bias in X/N relative to that in Y/N is smaller for treatments that are more effective and less toxic. The factor $\rho\{\pi(1-\pi)\}^{1/2}/\{\theta(1-\theta)\}^{1/2}$ in (2) has a maximum of 1, attained at $\pi = \theta$ and $\rho = 1$; and a minimum of –1, attained at $\pi = 1 - \theta$ and $\rho = -1$. These are not realistic scenarios. Nevertheless, they indicate that $\mathrm{E}[X/N - \pi] \leq \mathrm{E}[Y/N - \theta]$. In essence, the bias in X/N cannot exceed the bias in Y/N.

Perhaps more important is the bias in X/N if a study has reached the maximum planned sample size, $N = K$, as has occurred in the taxol plus herceptin trial mentioned in Section 1. For the boundary shown in Table 3, and assuming $\theta = 0.1$, we obtain $\mathrm{E}[Y/N - \theta] = 0.01940$, and conditionally $\mathrm{E}[Y/N - \theta \mid N = K] = -0.00487$. Assuming $\pi = 0.3$ and $\rho = 0.2$ and applying the bias formulas given above yields $\mathrm{E}[X/N - \pi] = 0.00593$ and conditionally $\mathrm{E}[X/N - \pi \mid N = K] = -0.00149$. Thus, if the trial runs to the end, the bias introduced by the toxicity-based stopping rule is practically negligible. The bias of the MLE increases with θ (Figure 2), but for larger values of θ the drug would be considered too toxic and the estimation of π would not be a major concern. To give investigators an idea about the potential magnitude of bias, we recommend doing bias calculations as described above for a range of parameter values deemed reasonable or appropriate for a particular trial.

5. Conclusions

This article has addressed the formal incorporation of stopping rules for toxicity in phase II trials. It was shown that the estimation of toxicity after using a sequential boundary is feasible. The proposed estimator is obtained using a combination of exact calculations and quadratic programming methods. The effect of incorporating the boundary on the estimation of the probability of response is minimal unless toxicity and response are highly correlated. When the toxicity rate is low, the trial runs to the end with high probability, and the bias introduced in the estimation of the response probability is generally low. On the other hand, if the toxicity rate is high, the trial stops early with high probability and in those cases unbiased estimation of the response probability should not be a major concern. In this article, we only considered a single-stage phase II trial. The methods for constructing estimators described in this article can be extended to trials where there is a need to continuously monitor toxicity and a multistage design is used to monitor response. Software is available at `www.bios.unc.edu/~qaqish/software`.

Acknowledgements

The authors thank Dr. Lisa Carey for the permission to use protocol LCCC 9818 and the referees and an associate editor for their constructive comments.

References

Bryant, J. and Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* **51,** 656–664.

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

Fletcher, R. (1987). *Practical Methods of Optimization*, 2nd edition. New York: Wiley.

Girshick, M. A., Mosteller, F., and Savege, L. J. (1946). Unbiased estimates for certain binomial sampling problems with applications. *Annals of Mathematical Statistics* **17,** 13–23.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* London, New York: Chapman and Hall/CRC.

Jung, S. H. and Kim, K. M. (2004). On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine* **23,** 881–896.

Legedza, A. T. R. and Ibrahim, J. G. (2001). Heterogeneity in phase I clinical trials: Prior elicitation and computation using the continual reassessment method. *Statistics in Medicine* **20,** 867–882.

Lindsey, J. K. (1997). Stopping rules and the likelihood function. *Journal of Statistical Planning and Inference* **59,** 167–177.

Liu, P. Y. (2001). Phase II selection designs. In *Handbook of Statistics in Clinical Oncology,* J. Crowley (ed), 119–127. New York: Dekker.

Muggia, F. M., Liu, P. Y., Alberts, D. S., Wallace, D. L., O'Toole, R. V., Terada, K. Y., Franklin, E. W., Herrer, G. W., Goldberg, D. A., and Hannigan, E. V. (1996). Intraperitoneal mitoxantrone or floxuridine: Effects on time-to-failure and survival in patients with minimal residual ovarian cancer after second-look laparotomy—A randomized phase II study by the Southwest Oncology Group. *Gynecological Oncology* **61,** 395–402.

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.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* **90,** 455–463.

Whitehead, J. (1986). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics* **42,** 461–471.

Received August 2003. *Revised August* 2004.
Accepted August 2004.

Appendix

Proof. The fact that matrix $A = \lambda D + (1 - 2\lambda)F^{\mathrm{T}}F$ is positive definite for all $\lambda \in (0, 1)$.

We show that $t^{\mathrm{T}}At > 0$ for any n × 1 vector $t \neq 0$. Write

$$\begin{aligned} t^{\mathrm{T}}At &= \sum_{i=1}^{m} \rho_i t^{\mathrm{T}}\left[\lambda D_i + (1-2\lambda) f_i f_i^{\mathrm{T}}\right] t \\ &= \lambda \sum_{i=1}^{m} \rho_i t^{\mathrm{T}} \left[D_i - f_i f_i^{\mathrm{T}} + \frac{1-\lambda}{\lambda} f_i f_i^{\mathrm{T}} \right] t \\ &= \lambda \sum_{i=1}^{m} \rho_i \left[t^{\mathrm{T}}\left\{D_i - f_i f_i^{\mathrm{T}}\right\} t + \frac{1-\lambda}{\lambda}\left(t^{\mathrm{T}} f_i\right)^2 \right]. \end{aligned}$$

Now, $D_i - f_i f_i^{\mathrm{T}}$ is an $n \times n$ positive semidefinite multinomial covariance matrix of rank $n - 1$ and $t^{\mathrm{T}}(D_i - f_i f_i^{\mathrm{T}})t = 0$ only if t is a multiple of a vector of all ones. But for such t, the term $(t^{\mathrm{T}} f_i)^2$ is strictly positive. Thus, $t^{\mathrm{T}}At > 0$ for any $t \neq 0$.