Published in final edited form as:
Stat Biopharm Res. 2009 August 1; 1(3): 213–228. doi:10.1198/sbr.2009.0014.

Flexible Phase I Clinical Trials: Allowing for Nonbinary Toxicity Response and Removal of Other Common Limitations

Richard F. Potthoff and **Stephen L. George**
Richard F. Potthoff is Senior Research Scientist, Cancer Statistical Center, and Stephen L. George is Professor of Biostatistics, Department of Biostatistics and Bioinformatics, and Executive Director, Cancer Statistical Center, Duke University Medical Center, Hock Plaza, 2424 Erwin Road, Suite 802, Durham, NC 27705

Abstract

Phase I clinical trials are often subject to severe limitations. The most important one is that they typically allow only for binary response—toxic (1) or nontoxic (0)—rather than a range of responses from 0 to 1. They also may not allow a new patient to be treated until results for all previous patients are available. They may assign patients to doses in groups of two or more, rather than individually. They may require the selected dose to be one of a few prespecified doses. The flexible method proposed here addresses these four limitations. It adopts a quasi-Bayesian approach incorporating a logistic dose–response model with two parameters for the mean response. The response at any dose follows a beta distribution, which entails a third parameter. The choice of dose for a patient is based on a utility function that reflects the latest estimates of toxicity and of the variance of the estimate of the maximum tolerated dose (MTD). Simulations show that the method works well, and that a nonbinary toxicity measure leads to a far more accurate MTD estimate than does a binary one.

Keywords

Bayes; Beta distribution; Continual reassessment method; Dose finding; Maximum tolerated dose; Toxicity scoring

1. Introduction

In this article we present a Phase I design and analysis scheme with several features that are, for the most part, absent from currently available designs. Chief among these features is allowing the response variable to take values over a range from 0 to 1 instead of being restricted to the two values of 1 (toxic) and 0 (nontoxic).

There are two potential advantages to allowing a range of values. First, there will often be at least some patients whose reactions to the drug under study defy a clear-cut binary categorization as either toxic or nontoxic. That is, there may be degrees of toxicity that can be more accurately reflected along a continuum from 0 to 1 rather than as a dichotomy.

Second, allowing a range of values for the toxicity variable, rather than just two values, results in greater statistical efficiency, thus allowing more accurate estimation of the maximum tolerated dose (MTD) with the same number of patients. Our simulation results, reported later, confirm the increase in accuracy. As a rough analogy that also contrasts dichotomous versus

potth001@mc.duke.edu.

continuous measure, note that, for estimating a normal-population mean, the efficiency of the sample median relative to the sample mean is only $2/\pi$.

Although permitting toxicity scores to take values besides 0 and 1 is the leading benefit of our proposed scheme, it is by no means the only one. Three more advantages are also not found in most currently available schemes. First, even when toxicity results are not yet available from all previous patients, our scheme can still assign a dose to a new patient at once, thus avoiding delays that extend trial duration. Second, we assign doses to patients individually, rather than in groups of two or more who all receive the same dose. We can thus expect to gain greater accuracy stemming from greater flexibility in dose assignment. Third, the dose selected at the end of the trial can be determined along a continuum and thus need not be one of the trial doses. This is an advantage if it is feasible to supply the tested agent in a dose other than the trial doses. We are aware of no competing method that enjoys all four of the main benefits of our scheme.

Section 2 reviews some of the literature and discusses issues related to Phase I design. Section 3 suggests possible techniques for scoring toxicity outcomes using a continuum.

Our mathematical model, covered in Section 4, is quasi-Bayesian. It uses a two-parameter logistic dose–response curve to represent the mean toxicity response. The distribution of the response at any dose follows a beta distribution whose mean is specified by this curve and whose variance is governed by a third parameter. We use a few Bayesian tools mainly to circumvent nonpositive estimates of the slope parameter of the curve.

Section 5 presents our proposed design scheme. Whenever a dose is to be assigned to a new patient, the scheme uses all of the currently available data to evaluate a utility function for each eligible dose and then chooses the dose with the highest utility. The utility function reflects both the estimated toxicity for a dose and the estimated accuracy of the MTD estimate if the dose were to be chosen. Basing dose decisions on toxicity and accuracy together (rather than on one or the other alone) seems to be novel; a simple linear utility function can take both into account.

In Section 6 we show results of extensive simulations that evaluate our proposed scheme under varying conditions. Section 7 provides concluding remarks.

2. Issues Involving Phase I Designs

The discussion in this section covering Phase I design issues and related literature is not exhaustive. For more detail, see Potter (2006), Yuan, Chappell, and Bailey (2007), and Potthoff and George (2007), along with their references.

Not all previous Phase I work is based on binary scoring of toxicity. The standardized toxicity criteria for cancer trials specify several severity levels for each type of defined toxicity (Cancer Therapy Evaluation Program 2003), but these are often dichotomized in the analysis. The toxicity index of Rogatko et al. (2004, p. 4646) offers a way to combine ratings (of 0, 1, 2, 3, or 4) across different toxicity types. A design scheme of Gordon and Willson (1992) makes use of individual toxicity grades, but only for purposes of defining the dose-escalation rules. Techniques of Wang, Chen, and Tyan (2000) differentiate between toxicity grades 3 and 4, but not among 0, 1, and 2. A method by Bekele and Thall (2006) uses individual grades of toxicity but is complex mathematically and requires extensive input from oncologists. Yuan et al. (2007) used a quasi-Bernoulli model to extend the continual reassessment method (CRM) of O'Quigley, Pepe, and Fisher (1990) to a setting that allows nonbinary toxicity scores, although the resulting designs lack some of the other features of our approach.

Little earlier effort has been devoted to creating Phase I designs that avoid the delays that stem from waiting for all previous patients' outcomes before assigning a dose to a new patient. One notable exception is the work of Cheung and Chappell (2000), which is aimed especially at settings where toxicity can be fully assessed only after a lengthy wait.

Unlike our proposal, existing designs that assign doses to patients individually (rather than in groups) generally achieve the resulting increase in flexibility only at the expense of greater wait time. In an attempt to reduce (but not eliminate) the attendant delays, Goodman, Zahurak, and Piantadosi (1995) (see also Korn et al. 1994) proposed modifying the CRM by assigning patients in groups of three rather than individually.

Many Phase I designs do not estimate the MTD along a continuum but rather require the dose finally selected to be one of the doses used in the trial. Earlier works that do use a continuum include those of Gatsonis and Greenhouse (1992) and Whitehead and Brunier (1995).

The traditional up-and-down method is not based on a formal model, a characteristic for which it is often criticized (e.g., Storer 1989). Other Phase I designs have used a one-parameter model, a two-parameter logistic-regression model, or isotonic regression.

Although the up-and-down method is non-Bayesian, many other Phase I designs use Bayesian methodology to varying degrees. We use a rough Bayesian approach in selecting patients' doses, but can then produce final MTD estimates that are either frequentist or Bayesian. Whitehead and Brunier (1995) also used a mixture of techniques.

Most Phase I designs use toxicity alone for their criterion for selecting patients' doses, although Whitehead and Brunier (1995) used estimation accuracy alone. We use both, through our utility function. Designs that use toxicity include conservative ones that focus on incurring low toxicity and aggressive ones that seek to avoid doses that are either too low (to be beneficial) or too high. Although we follow the former precept, our utility function is easily modified to fit the latter. Weighing risk against benefit in Phase I trials is not easy to do (Joffe and Miller 2006).

There are still other issues related to Phase I designs. They include starting doses, rules for escalation of doses, variable sample size controlled by a stopping rule versus fixed sample size (we use the latter), and handling of dropouts.

3. Nonbinary Measurement of Toxicity

We consider several ways to refine the measurement or scoring of the variable that represents toxicity so that it is not simply a binary variable but rather is closer to being a continuous variable, one that (for any given dose) may approximately follow a beta distribution as supposed in our model. These ways can be used separately or in any combination. The more of them that are used, however, the greater the amount of information that can be captured and the closer the variable will come to being continuous. Some possibilities are as follows.

First, different grades of toxicity can retain their individualities instead of being dichotomized. For example, if there are six toxicity grades such that ordinarily grades 0, 1, and 2 would be scored as 0 and grades 3, 4, and 5 as 1, then the scores might be set (respectively) to 0, 0.2, 0.4, 0.6, 0.8, and 1 for the six grades.

Second, if toxicity can be scored for more than one relevant organ system or toxicity type, then one can obtain a score for each. After that, the scores can be combined across the systems or types, either through a weighted average with suitable weights, or some other way.

Finally, in some cases it may be feasible and appropriate for two or more raters, working independently, to provide toxicity scores. The scores of the different raters can then be averaged to obtain a final single value to use for the patient's response variable for toxicity.

Because a binary toxicity variable is ingrained in the traditions of Phase I clinical trials, a change of culture would be involved in shifting to a toxicity measurement that is more like a continuous variable. Note, though, that if one combines ratings based on sets of ordered toxicity categories to bring about a single final toxicity measure as just described, the discrete toxicity categories that form the basis of such a measure can be the same ones that have long been in use.

The choice of the method of combining toxicity ratings, and of the final measure, will depend on the disease, the drug, and other factors. Appendix A provides an illustrative example that relates to a Phase I trial for pancreatic cancer with the drug bevacizumab.

4. Mathematical Model

We want to obtain a point estimate for the MTD, as well as a confidence interval (either one- or two-sided) for it. The mathematical model is formulated so as to meet these objectives.

The response variable, y, represents toxicity. It can take any value from 0 to 1, with higher y indicating greater toxicity. The dose variable is denoted by x. Although x can take any value, a Phase I study would ordinarily have only a small number, d, of doses. For ease of presentation in what follows, a study with d equally spaced dose levels whose values are $x = 1, 2, \ldots, d$ is assumed. No generality is lost in specifying equally spaced values, because our development can be extended in obvious fashion to a study where the dose levels are not spaced equally (or are equally spaced but on a different scale).

Note also that x need not be the dose itself, but rather could be some monotonically increasing function of the dose. Thus, if $x = \log_2(\text{dose})$ and the x's are consecutive integers, then the doses themselves are successively doubled. Of course, the dose–response curve differs according to how x is defined. A curve where x is the dose itself might not give as good a fit as one where x is taken to be the logarithm of the dose, for example. Hereafter, to simplify the exposition, the variable x will be called the *dose*, irrespective of whether it is the dose itself or some function of the dose.

Because we do not require the toxicity response to be binary, we use the term *targeted mean toxicity*, in place of the less general term *targeted probability of toxicity*, to refer to the toxicity level to which the MTD pertains. This level, which we denote by M, is chosen by the investigator, based generally on the severity and reversibility of the toxicity. Bekele and Thall (2006) gave an example of how a quantity similar to it might be elicited from oncologists.

4.1 The Model for Distribution of Toxicity

Define $m_x = \mathrm{E}(y|x)$. It is the *mean* toxicity (*not* the probability of toxicity) at dose x. For m_x we assume a two-parameter logistic dose–response curve represented by

$$\log \frac{m_x}{1 - m_x} = a + bx, \tag{1}$$

where $b > 0$. The MTD, denoted below by h, is the dose yielding a mean toxicity of M. This is obtained simply from (1) by setting $m_x = M$ and solving for x. The result is

$$h=\frac{\log k - a}{b}, \tag{2}$$

where $k = M/(1 - M)$. We can also write m_x in terms of h and k as a direct result of (1) and (2):

$$m_x=\frac{e^{a+bx}}{1+e^{a+bx}}=\frac{ke^{b(x-h)}}{1+ke^{b(x-h)}}. \tag{3}$$

The distribution of toxicity (y) given dose (x) is assumed to follow a beta distribution with expectation $E(y|x) = m_x$ and variance

$$\mathrm{var}(y|x)=\frac{m_x(1-m_x)}{t+1}, \tag{4}$$

where t is a dispersion parameter ($t > 0$) not dependent on x. Thus, the beta distribution is

$$y|x \sim \frac{1}{\mathrm{B}(\alpha_x,\beta_x)}y^{\alpha_x-1}(1-y)^{\beta_x-1}, \tag{5}$$

where the beta parameters α_x and β_x are given by

$$\alpha_x=tm_x=\frac{tke^{b(x-h)}}{1+ke^{b(x-h)}} \tag{6}$$

and

$$\beta_x=t(1-m_x). \tag{7}$$

Note that lower t corresponds to greater dispersion. In fact, our beta distribution approaches the binomial distribution with mean m_x as $t \to 0$.

The model specified by (1)–(7) is the same as the beta regression model of Ferrari and Cribari-Neto (2004), which they applied to data on the fraction of crude oil converted into gasoline and the fraction of income spent on food. We do not use maximum-likelihood estimation as they did, though, because such estimation may run into difficulties with small sample size or small t.

4.2 Bayesian Elements of the Model

For choosing patients' doses, we use a (quasi-) Bayesian approach, since otherwise there is no good way to handle zero or negative estimates of the slope parameter b that can arise (especially before many patients have accrued) if the estimates of b are based on the data alone. The data can yield a zero estimate of b if, for example, no toxicities have yet been observed.

But, although the basis for the dose selections is Bayesian, the estimation for the MTD from the final data need not be. That is, the final MTD estimation can be either Bayesian or frequentist, as one prefers.

Our parameterization for the Bayesian elements of the model, as well as for some of its other elements, involves choices, based both on rough theoretical considerations and on trial-and-error experimentation. With respect to the two-parameter logistic dose–response curve for mean response, we assume a joint prior distribution for h and b (rather than, e.g., for a and b) such that h and b are independent. Let h and b, respectively, have prior means of E_H and E_B, and prior variances of S_{HH} and S_{BB}. The associated distributions can be normal, but need not be. Upon solving (2) for a and then taking expectations, one obtains

$$E_A = \log k - E_B E_H \tag{8}$$

for the prior mean for a. Further, the prior variance matrix for (a, b) is

$$\mathbf{S} = \begin{bmatrix} S_{AA} & S_{AB} \\ S_{AB} & S_{BB} \end{bmatrix} = \begin{bmatrix} S_{BB}S_{HH} + E_H^2 S_{BB} + E_B^2 S_{HH} & -E_H S_{BB} \\ -E_H S_{BB} & S_{BB} \end{bmatrix}. \tag{9}$$

Although the correlation between h and b is 0, note that the correlation between a and b can be strongly negative.

Besides h and b, our model has one more parameter that is treated in a Bayesian fashion. This third parameter pertains to the variance of the beta distribution. We take it to be

$$v = \frac{1}{t+1}, \tag{10}$$

which must lie between 0 and 1, rather than t itself [see (4)–(7) above], which can vary from 0 to ∞. We denote the prior mean and variance of v by E_V and S_{VV}, respectively, and assume mutual independence of h, b, and v in their joint prior distribution.

The choices of E_H, E_B, and E_V, as well as S_{HH}, S_{BB}, and S_{VV}, must be based on judgment and on discussions with clinical investigators. As an example, if $d = 6$ and if the doses are selected so that the two middle ones bracket the anticipated MTD, then the choice $E_H = 3.5$ would appear reasonable. One would then want to spread the six doses so that they are neither so close together that the resulting prior dose–response curve would be almost flat, nor so far apart that the ordinate of the curve would be close to 0 at dose $x = 3$ or close to 1 at $x = 4$. The choice of $E_B = 0.7$, for instance, might conform to doses selected so that they are anticipated to produce a suitable curve. For doses $x = 1, 2, 3, 4, 5, 6$, respectively, the values of the response function m_x of (3), with $h = 3.5$ and $b = 0.7$, are 0.04, 0.08, 0.15, 0.26, 0.42, 0.59 for $k = 0.25$ (i.e., $M = 0.2$) and 0.08, 0.15, 0.26, 0.42, 0.59, 0.74 for $k = 0.5$ (i.e., $M = 1/3$). In our simulations (Section 6) we used $d = 6$, and set $E_H = 3.5$, $E_B = 0.7$ as the prior means.

E_V should be easier to choose than E_H or E_B, because, at least after a little experience is acquired, one should have a good grasp of how the selected scoring system will cause the toxicity scores to be distributed. If one is able to take advantage of the suggestions of Section

3 and create a scoring system that comes close to producing a continuous toxicity variable, then the appropriate value of E_V will be relatively low. But if not, then the beta distribution will be close to the binomial and a value of E_V close to 1 will be called for.

5. Choosing the Dose for Each Patient

Our design selects a dose for each patient through a utility function that takes into account both estimated toxicity and estimated accuracy of the MTD estimate. One can apply special rules, however, to govern doses at the outset and to restrict escalation and bunching of doses. After the study is finished, one obtains an estimate and a confidence interval for the MTD. The following subsections provide details on each of these matters.

5.1 The Utility Function

The number of patients to be studied, n, is fixed in advance. Except where overridden by the rules governing initial doses and restricting escalation and bunching, the dose-selection routine works as follows. Upon entering the trial, each patient i is assigned a dose (denoted by x_i), even if there are outcome(s) from some previous patient(s) that are not yet known. One uses all available results from the previous patients to calculate the value of a utility function for each of the d doses. Then the new patient i receives whichever dose (x_i) has the highest utility.

Let $U_{i\,x}$ denote the posterior estimate of utility if patient i were to receive dose x ($x = 1, \ldots, d$). Our utility function (calculated just before x_i is chosen) takes the form

$$U_{ix}=Q(1-P_{ix})+\frac{1}{S_{HH+ix}}, \tag{11}$$

where $P_{i\,x}$ is the posterior toxicity estimate at dose x and $S_{H\,H+\,i\,x}$ is the posterior estimated variance for the estimate of h (the MTD) if patient i were to receive dose x. Q is a prespecified parameter ($0 < Q < \infty$) that balances (1 − toxicity) against the precision of the estimated h. One chooses x_i to be the value of x that maximizes $U_{i\,x}$ (subject to the constraints mentioned earlier).

Many ways of estimating $P_{i\,x}$ and $S_{H\,H+\,i\,x}$ for the model of Section 4 could be devised. We tried different estimation methods and present below the one that we judged best.

As noted earlier, one could alter our utility function, if desired, to deter doses that are either too low to be effective or too high. For instance, just replace $Q(1 - P_{i\,x})$ in (11) with $-Q_1(M - P_{i\,x})$ if $P_{i\,x} \leq M$ or with $-Q_2(P_{i\,x} - M)$ if $P_{i\,x} \geq M$ (where usually Q_1 would be $\leq Q_2$).

5.2 The Posterior Toxicity Estimates

The values of $P_{i\,x}$, the posterior toxicity estimates for the d doses x at the time that patient i arrives, are found as follows.

Let n_x denote the number of previous patients at dose x who are *complete* (i.e., whose response scores have been obtained) at the time that patient i becomes available. We use n_x rather than $n_{i\,x}$ to economize on notation. Note that $\Sigma_x\, n_x$ will be $\leq (i - 1)$. For any patients who are not yet complete, we assume there is no interim information related to their toxicity response.

Let Y_x denote the sum of the response scores y for the n_x patients at dose x who are complete when patient i arrives. Before the summing, however, a response score is changed to ε if it is $< \varepsilon$ or to $(1 - \varepsilon)$ if it is $> (1 - \varepsilon)$. (We used $\varepsilon = 0.01$ in our simulations.)

For each new patient, the d triples (Y_x, n_x, x) are fed into PROC LOGISTIC of SAS, which works even though Y_x is not generally an integer. The logistic-regression software is used simply as a computational device. It produces reasonable estimates of the two parameters of (1), based on the usual log-likelihood, even though the usual logistic-regression model is not our actual model. The associated variance matrix, however, is overestimated because values of the response variable will not be simply 0's and 1's. But we correct for the resulting bias by taking into account the variance of the beta distribution through estimating the parameter v of (10).

The change involving ε prevents a calculation failure if no toxicities have yet arisen and all y-values observed to date are 0. It also prevents a result showing separation of points if, so far, all values of y are 0 for doses below some dose x and are 1 for doses $\geq x$. It thereby enables our scheme to work fully in settings where some or even all responses y are either 0 or 1.

Especially when there are not yet many patients, the data may yield a nonpositive estimate of the slope parameter b. If so, we make some adjustments before we use estimates from the data, together with priors, to obtain posterior estimates.

Full mathematical details for calculating the $P_{i\,x}$ values are in Appendix B.

5.3 The Posterior Estimates for the Variances of the Estimated MTD

$S_{H\,H+\,i\,x}$ denotes the posterior estimated variance of the estimate of h if dose x were to be given to patient i. Its reciprocal (the precision of the MTD estimate) is needed, along with $P_{i\,x}$, for the utility function (11). Appendix C gives the details for the calculation of $S_{H\,H+\,i\,x}$.

The calculation exploits the fact that, if there are patients who arrived before patient i but are incomplete, their doses are known even though their responses are not. Although this dose information is not used in finding $P_{i\,x}$, it *is* used in getting $S_{H\,H+\,i\,x}$. Because the choice of a dose for patient i thus takes into account the doses of all $(i-1)$ earlier patients (not just the complete ones), the resulting configuration of patient doses should yield better estimation accuracy.

5.4 Estimates and Confidence Intervals for the MTD

After all n patients are complete, one can obtain a point estimate and a one- or two-sided confidence interval for the MTD. Although the dose-selection routine used a Bayesian approach, the final point and interval estimation can be either Bayesian or frequentist, as one chooses.

The final Bayesian and frequentist point estimates for the MTD are denoted, respectively, by H_+ and h_*. Their respective associated variances are called $S_{H\,H+}$ and s_{hh*}. Appendix D provides the formulas for $S_{H\,H+}$ and s_{hh*}, along with explanation of how H_+ and h_* are obtained using formulas in Appendix B. In case the final data produce a slope estimate ≤ 0, the frequentist point and interval estimation for the MTD will not be available. (But the Bayesian estimation, though available, may itself be of little value in this situation.)

A confidence interval for the MTD can be obtained in the usual manner either with H_+ and $S_{H\,H+}$ (Bayesian interval) or with h_* and s_{hh*} (frequentist). The MTD point estimate, H_+ or h_*, is assumed to be approximately normally distributed, so the half-width of a 95% two-sided interval is taken as 1.96 times the square root of $S_{H\,H+}$ or s_{hh*}. Simulation results reported in Section 6 provide some indication of the accuracy of frequentist confidence intervals.

5.5 Doses at the Start of the Trial

For the first few patients, it may be best, for two reasons, to assign predetermined doses rather than apply the routine of Sections 5.1–5.3 to get doses. First, concerns about the safety of the drug may cause one to use special care with a small number of initial patients, by treating them at the lowest doses, monitoring them quite closely, and perhaps imposing a delay before enrolling new patients so that any unexpected longer-term problems are discovered sooner.

Second, before one can obtain even the crudest sample estimates of all three parameters in our model, one needs response scores from at least three patients. Moreover, estimation of the parameters a and b in (1) requires score data from at least two different doses, and to estimate v (10) from the sample it seems best to have responses from at least two patients at the same dose.

Although our scheme is flexible regarding doses at the outset, for our simulations we assign doses of 1, 1, and 2 to the first three patients and wait for their results (after which we switch to the routine of Sections 5.1–5.3 for the fourth and later patients). Using initial doses 1, 1, 2 in this manner addresses both the safety and the estimation issues.

5.6 Restrictions on Escalation and Bunching of Doses

For safety reasons one may want to prevent doses of successive patients from escalating too rapidly, although a tradeoff can be expected between better MTD estimation resulting from faster escalation and greater safety produced by slower escalation. Various rules to restrict dosage escalation could be devised. Three possibilities are shown below; we used rule (ii) in our simulations. All three define when it is permissible to assign dose $(x + 1)$ once dose x has been assigned. Any given dose is allowed only after all lower doses are tried. The three escalation rules, which are in order from strictest to least strict, are as follows:

- **i.** After a patient at dose x is complete, it is permissible to assign dose $(x + 1)$, but to only *one* patient until that patient is complete.
- **ii.** After a patient at dose x is complete, it is permissible to assign *any number* of patients to dose $(x + 1)$.
- **iii.** After a patient (whether complete or incomplete) has been *assigned* to dose x, it is permissible to assign any number of patients to dose $(x + 1)$.

For better MTD estimation it may be best to prevent too many patients from bunching up on the same dose. For example, if all patients were treated at only two adjacent doses, then the limited spread of the doses could adversely affect the estimation of the dose–response curve (1). Excessive bunching did sometimes occur in our early simulations even though the utility function ought to prevent it. We therefore put some mild restrictions on bunching of doses. We prevented assignment of patient i to dose x if doing so would cause the number of patients at dose x to exceed $(0.4i + 1.5)$. Thus the number of patients receiving any dose x cannot (e.g.) exceed 3 of the first 6 patients, 5 of the first 9, 7 of the first 15, 11 out of 25, or 21 out of 50.

The combination of escalation and bunching restrictions could conceivably cause all d doses to be disallowed. If that happens, dose 1 is the dose chosen.

6. Simulations

6.1 Description of the Simulations

Except for runs reported in Section 6.4 that compare our method with others, all of our simulations are as described here in Section 6.1. There were always $d = 6$ equally spaced doses $(x = 1, 2, 3, 4, 5, 6)$. Service time, the time after arrival until the patient's response is known,

was taken to be constant and equal to one unit of time for each patient. The times between patient arrivals were randomly drawn from the exponential distribution with mean $1/L$, so that arrivals per unit of time followed a Poisson distribution with mean L (a simulation parameter). The toxicity value y for any patient assigned to dose x was drawn from the beta distribution given by (5)–(7), except that the value was changed to 0.01 if $y < 0.01$ or to 0.99 if $y > 0.99$. The respective doses for the first three patients were always 1, 1, 2 (see Section 5.5). We controlled dosage escalation and restricted dose bunching as indicated in Section 5.6.

In all simulations, the prior means were $E_H = 3.5$ for h and $E_B = 0.7$ for b, and the respective prior variances were $S_{HH} = 1.42$ and $S_{BB} = 0.2$. For v, we used $E_V = v$ and $S_{VV} = E_V(1 - E_V)/16$. E_V, the prior mean for v, was taken to be equal to the "true" value of v, on the assumption that the dispersion features of response scores will be largely known in advance.

Seven variables (h, b, v, M, n, Q, and L) were studied in the simulations. One of them (h) was at three levels and the others were at two, thus producing 3×2^6, or 192, total combinations. There were 50 runs for each combination, or 9,600 runs altogether. Of the seven variables, h, b, and v are "unknown" model parameters, though v may be partially controllable; M, n, and Q are selected by the investigator; and L may be largely controllable if patients are not scarce.

The "true" values of the first three variables were set to 2, 3.5, and 5 for h; 0.7 and 1.4 for b; and 0.5 and 0.95 for v. Thus h, the MTD, was set at its targeted value (presumed to be 3.5), and also above and below the target to represent situations where the MTD is misjudged in either direction. The slope parameter, b, was simulated at its target value of 0.7, as well as at a higher value that entails a dose–response curve whose mean response is more sensitive to the dose than anticipated. The 0.95 value for the dispersion parameter, v, gives a toxicity distribution that is almost binomial (with most responses close to the extremes of 0 or 1), whereas the 0.5 value leads to fewer toxicity results near 0 or 1.

For M, the targeted mean toxicity, the simulations used values of 0.2 and 1/3, for which the corresponding values of k are 0.25 and 0.5, respectively. Sample sizes (fixed) were simulated at $n = 25$ and $n = 50$ patients. The values for Q, the parameter in the utility function that governs the weight attached to lower toxicity relative to higher precision, were 0.2 and 0.6. The mean number of patient arrivals per unit of time was simulated at $L = 1$ and $L = 3$.

To analyze and interpret the simulation results, we mainly used analyses of variance. Many interaction effects, even high-order ones, were statistically significant, but generally not practically significant. Here we have space to give only a limited report of our findings. Although we obtained extensive simulation results for both the frequentist MTD estimate (h_*) and the Bayesian one (H_+), our account here is confined to the former. Greater detail is available elsewhere (Potthoff and George 2007).

6.2 Results for Toxicity

Analyses of variance of the 9,600 toxicity averages from the simulations indicate that L shows no statistical significance in relation to toxicity, and b and n show statistical but not practical significance. Among the other four factors, all main effects and interactions are significant at $p < 0.00001$. Therefore, in Table 1 we show toxicity averages (displayed as percentages rather than decimal fractions) for all 24 combinations of M, h, Q, and v.

Not unexpectedly, toxicities are much higher for $M = 1/3$ than for $M = 0.2$. With both M-values, though, toxicities are either below M (for $h = 3.5$ and 5) or not far above M (for $h = 2$, where the drug is most toxic). Thus, on this basis the design gives acceptable average toxicity.

Toxicity decreases as h increases. This is no surprise, because with higher h the drug is less toxic relative to anticipations, and so more patients are tested at doses with low toxicities.

As intended by the design, toxicity does turn out to be lower with the higher value of Q (Q = 0.6 versus 0.2). The toxicity difference is generally rather small, though, at least for $v = 0.5$ and lower h. Moreover, even where the difference is larger, one pays a penalty in the form of much greater inaccuracy of the MTD estimate, as will be shown below.

Finally, toxicity is about the same for the two values of v in the left columns of Table 1 but becomes sharply lower for $v = 0.95$ versus $v = 0.5$ as one moves to the right. The difference is especially large at $h = 5$ and $Q = 0.6$. Again, though, any toxicity benefits with $v = 0.95$ generally come at the expense of greater inaccuracy, as will be seen shortly.

6.3 Results for Accuracy of the MTD Estimates

For the simulation results regarding the accuracy of the MTD estimation, our focus here is on the frequentist MTD estimate, h_*. Accuracy has several facets. One might be interested in confidence-interval coverage, bias, absolute or squared error, or the frequency with which $|h_* - h|$ fails to be < 0.5 (half the unit of increment for the doses). Moreover, for bias and error, one might like to look not only at the "horizontal" error, $(h_* - h)$, but also at the "vertical" error, that is, the difference between the toxicity response (3) based on h itself versus the same function (3) predicted with h_* substituted for h. In what follows we broadly convey our results on accuracy.

In 209 runs out of the 9,600, h_* does not exist because the frequentist slope estimate (denoted by b_{*0}) is ≤ 0. We handle these runs by setting h_* to 0 (19 runs) or to 7 (190 runs) if average toxicity is, respectively, > or < M. Likewise, we change h_* to 0 (27 runs) or 7 (210 runs) if h_* exists but has a value < 0 or > 7 (respectively). These changes apply to all analyses below except the ones for confidence intervals, for which any affected runs are simply excluded.

In fact, for those affected runs, Table 2 shows their distribution among the levels of each variable. One gathers that fewer extreme MTD estimates (and thus perhaps better estimation) will occur with lower h, greater b, smaller v, larger n, and smaller Q. There is no effect from L, and little from M.

6.3.1 Confidence-Interval Coverage—We look at the percentage of runs in which a 95% two-sided confidence interval, $h_* \pm 1.96\, s_{h*}$, does not include the "true" value of h. Here, s_{h*} denotes the square root of the quantity s_{hh*} given in Appendix D. Table 2 provides a limited summary. For each level of each experimental variable as well as overall, the last column of the table shows the percentage of coverage failure (based on 9,154 of the 9,600 runs, with exclusions as just indicated). Overall, 8.88% of the intervals fail to include h, versus an ideal 5%. Variation in coverage among levels of the variables is small, but is greatest for variables h and b.

6.3.2 Bias—Two types of bias are considered for the MTD estimates: horizontal and vertical. They refer (respectively) to $(h_* - \mathrm{h})$ and $(m_{h*} - m_h)$. The value of m_x, for $x = h_*$ and h, comes from the rightmost expression in (3) (but note that $m_h = M$).

Various analyses of variance were done for each type of bias, using the values of h_* and h from the 9,600 runs. L has little impact. The other six factors, however, are entangled in a number of interactions that have tiny p-values but are largely of minor practical significance. In view of this complexity, results here are presented in limited form. For both types of bias, Table 3 shows the average bias for each level of the seven experimental variables and overall. Vertical

bias is expressed as a percentage, in similar fashion to Table 1, rather than as a decimal fraction (e.g., the figure of +2.0 for vertical bias for $M = 0.2$ indicates that 0.220 is the average m_{h*}).

Our test for judging main effects shows that differences among levels of the factors (other than L) are significant at $p = 0.013$ or lower in all cases except the vertical bias for M, for which $p = 0.28$. Both horizontal and vertical biases are closer to zero for smaller h, larger b, smaller v, smaller M, larger n, and smaller Q. Biases in Table 3 that are comparatively far from zero include those for $h = 5$, for $v = 0.95$, and for $Q = 0.6$.

In fact, the specific combination $h = 5$, $v = 0.95$, $Q = 0.6$ is of interest, thanks both to its notably low toxicity values in Table 1 and to the greater main-effect biases in Table 3. The average biases for its 800 runs are both sizeable: +0.43 horizontal and +11.1 vertical.

In general, though, our results suggest no major difficulties with respect to bias. The overall average biases are both close to zero (+0.07 horizontal, +2.1 vertical). Although many main effects and interactions are statistically significant, their magnitude is small and for horizontal bias is, in all cases, only a small fraction of the unit of increment (= 1) for the doses.

6.3.3 Absolute Error—For the next facet of accuracy of the MTD estimates, we present the results for absolute error. Results for squared error are similar. Absolute error, however, seems to be less skewed and less sensitive to outliers.

The evaluation for absolute error follows much the same path as for bias. L shows no effect, the six remaining variables are enmeshed in many interactions that have minute p-values but little practical significance, and we again provide a limited summarization, shown in Table 3.

Overall, average absolute error is + 0.42 horizontal and + 7.4 vertical. It is higher for $h = 5$ than for $h = 2$ and 3.5, for $v = 0.95$ versus $v = 0.5$, for $n = 25$ versus $n = 50$, and for $Q = 0.6$ versus $Q = 0.2.v$ In fact, the ratio for $n = 25$ versus $n = 50$ is close to $\sqrt{2}$, as one might expect. Differences among levels of the first six variables in Table 3 are significant at $p < 0.0001$ in all cases except vertical absolute error for b, for which $p = 0.06$.

As before, the combination $h = 5$, $v = 0.95$, $Q = 0.6$, whose toxicities in Table 1 are quite low, merits attention. It has high average absolute error: + 0.92 horizontal and + 18.4 vertical.

Also of interest are combinations of v, Q, and n, the three variables that may provide the best opportunities for an investigator to make choices that will produce a better design. At one extreme, the combination $v = 0.95$, $Q = 0.6$, $n = 25$ has average absolute error of +0.86 horizontal and +15.6 vertical. At the other extreme, the figures for the combination $v = 0.5$, $Q = 0.2$, $n = 50$ are far lower: +0.22 horizontal and +4.0 vertical. Thus, investigators who are in a position to exercise some influence over the values of v, Q, and n can expect to reap a large payoff in terms of smaller absolute error in MTD estimation.

6.3.4 Frequency With Which MTD Estimate Fails to Be Close to Target—Although the foregoing provides results for absolute error, there may be a specific interest in the frequency with which $|h_* - h|$ (horizontal absolute error) is ≥ 0.5. Such a frequency could be particularly relevant if the dose chosen at the end has to be one of the trial doses, because then the frequency would represent the probability of failing to choose the correct dose if the true MTD is $h = 2$ or 5. Put another way, a horizontal absolute error < 0.5 would imply an estimate of the MTD that is (respectively for $h = 2$ or 5) between 1.5 and 2.5 or between 4.5 and 5.5, which would lead to the correct choice of 2 or 5 if rounding is to the nearest integer. Although interpretation is less straightforward for the noninteger value $h = 3.5$, the general concept is comparable.

The last column of Table 3 shows the percentage of runs in which the MTD estimate differs from the MTD by 0.5 or more. The column reflects all 9,600 runs, including the ones that Table 2 shows as excluded (because of estimates set to 0 or 7).

Less than 27% of all MTD estimates fail to fall within half a unit of the target. The ordering of the levels of each of the first six experimental variables is the same in the last column of Table 3 as it is for average horizontal absolute error in the third-from-last column.

It is again instructive to look at the two combinations $v = 0.95$, $Q = 0.6$, $n = 25$ and $v = 0.5$, $Q = 0.2$, $n = 50$. The percentages of 1,200 MTD estimates that do not fall within 0.5 of the target are 49.0% for the first combination but only 9.3% for the second.

6.4 Comparison of Our Method With Others

Can one find out how our method compares with competing ones? Several aspects of our method are intrinsic to it and have no counterpart in other approaches (and vice versa). Nonetheless, we did simulate our method to compare it, as well as possible, with the results shown for scenario L2 in Table 2 of Thall and Lee (2003) for their Bayesian logistic regression (BLR) method and for the CRM. In doing so, we were able to control a number of factors. As did they, we made 4,000 runs of $n = 36$ patients each, set the targeted mean toxicity at $M = 0.25$ (so that $k = 1/3$), and used $d = 6$ doses. To use the same "true" logistic dose–response curve as theirs, we set $b = 3.17$ and $h = -0.161$ in (3) above (the value of a is -1.05), thus giving mean toxicities of 0.01, 0.09, 0.26, 0.47, 0.64, 0.76 (as in their Table 2 for L2) for the doses of [log(1, 2, 3, 4, 5, 6) − log(6!)/6]. Because the doses are not 1, 2, 3, 4, 5, 6 as before and are not even equally spaced, we generalized (11) by replacing $1/S_{HH^+ix}$ with $[(x_d - x_1)/(d-1)]^2/S_{HH^+ix}$, where x_d and x_1 are (respectively) the highest and lowest doses; more obvious generalizations were made for some of the formulas in Appendices B and C. For our prior, we used $E_B = 2.398$ and $S_{BB} = 4$, the same as the Thall–Lee values, and we set $E_H = -0.4034$ and $S_{HH} = 0.3435$, which follow from their values of $E_A = -0.1313$ and $S_{AA} = 4$ upon using the E_A formula in (8) and the S_{AA} formula in (9).

In our simulation, we used $v = 0.5$ for the parameter for our beta distribution (as well as $E_V = 0.5$ and $S_{VV} = 1/64$ for the prior); Thall–Lee have nothing comparable, because their toxicity is strictly binary. We used $Q = 0.2$ for our utility function that balances (1 − toxicity) against estimation precision; they have no utility function. We used $L = 1$ for the mean number of patient arrivals per unit of (service) time, but they have no counterpart: They assign patients to a dose in groups of three and then wait until all outcomes are known before assigning a new group, whereas we assign patients individually without any waiting but do not always know all previous outcomes. Our prior has independence between b and h, whereas theirs has independence between b and a. Although their prior for b is normal with truncation to ensure $b > 0$, ours just specifies the mean and variance of b without any distributional assumption and is thus not concerned with the sign of b (though we deal with negative slope estimates in other ways). Their first patients receive the second-lowest dose with no restrictions thereafter, whereas our first two patients are at the lowest dose and the third is at the second lowest. They allow skipping of an untried dose. We do not; per Section 5.6, we restricted escalation according to rule (ii), and also restricted bunching.

Finally, they select one of the six trial doses as the MTD, whereas we estimate the MTD along a continuum. Thus we have to convert each of our estimates to one of the six doses in order to compare our simulation results with theirs. For scenario L2, they take the third of the six doses (with mean toxicity of 0.26) to be the true MTD: BLR and CRM picked the third dose 72% and 79% of the time, respectively. We tallied our estimate as being correct if it was closer to the third dose than to any other, that is, between [log(2 × 3)/2 − log(6!)/6] and [log(3 × 4)/2 − log(6!)/6]. Our method picked correctly 92.8% of the time. Our average toxicity was 0.235,

and thus below $M = 0.25$. Although these results are highly favorable to our method, interpretation may not be clear-cut in view of the points noted in the preceding paragraph.

We did a second, analogous simulation to test our method for the Thall–Lee L3 scenario (using $b = 3.30$ and $h = 0.535$ for the "true" dose–response curve). They reported 49% correct picks for BLR and 47% for CRM, whereas we had 60.0%. Our average toxicity was 0.174.

7. Concluding Remarks

This article has proposed a Phase I design that is highly flexible. Besides permitting a nonbinary measure for the toxicity response, it requires no waiting for results from earlier patients before assigning a dose to a new patient, assigns patients individually rather than in groups, and estimates the maximum tolerated dose along a continuum.

A comparatively simple quasi-Bayesian approach is applied for choosing patient doses. But final MTD estimation can be frequentist rather than Bayesian if preferred.

A nonbinary toxicity measure produces far better accuracy in MTD estimation than a binary one, as shown by our simulation results in Tables 2 and 3 for $v = 0.95$ (for which toxicity scoring is close to binary) versus $v = 0.5$. Toxicity for our design is acceptable (Table 1).

To choose the dose for a patient, we use a utility function that gauges the tradeoff between lower toxicity and better accuracy of the MTD estimate. Although the parameter that controls the penalty placed on higher toxicity works as intended, the toxicity reductions stemming from a higher penalty are mostly only modest. Moreover, where the toxicity reductions are largest, they come at the expense of much worse accuracy in MTD estimation and under conditions where toxicity is already low. Thus, our design, and perhaps others as well, cannot go very far in lowering toxicity without undue sacrifice of accuracy.

The frequency with which new patients arrive has no effect either on toxicity or on accuracy of MTD estimation for the comparison that we did. This result was not fully expected. Its implication is that an investigator who could shorten the trial duration by enrolling patients faster, perhaps by obtaining them from additional sites, would not suffer any adverse consequences (as to toxicity or estimation) from doing so.

There are three variables that seem to be the best to pursue to optimize accuracy (best with respect to both their effect on accuracy and the extent to which the investigator may be able to influence them). They are the degree to which toxicity response scores are close to being continuous, the parameter that penalizes higher toxicity, and the sample size.

Acknowledgments

This work was partially supported by Grant CA33601 from the National Cancer Institute. The authors thank the associate editor and two referees for helpful comments and suggestions.

References

Bekele, BN.; Thall, PF. Dose-Finding Based on Multiple Ordinal Toxicities in Phase I Oncology Trials. In: Chevret, S., editor. Statistical Methods for Dose-Finding Experiments. Chichester, UK: Wiley; 2006. p. 243-258.

Cancer Therapy Evaluation Program. Common Terminology Criteria for Adverse Events. 2003 [(accessed June 25, 2006)]. Version 3.0, DCTD, NCI, NIH, DHHS, March 31, 2003 (http://ctep.cancer.gov/forms/CTCAEv3.pdf), *Publish Date: December 12, 2003*

Cheung YK, Chappell R. Sequential Designs for Phase I Clinical Trials With Late-Onset Toxicities. Biometrics 2000;56:1177–1182. [PubMed: 11129476]

Crane CH, Ellis LM, Abbruzzese JL, Amos C, Xiong HQ, Ho L, Evans DB, Tamm EP, Ng C, Pisters PWT, Charnsangavej C, Delclos ME, O'Reilly M, Lee JE, Wolff RA. Phase I Trial Evaluating the Safety of Bevacizumab With Concurrent Radiotherapy and Capecitabine in Locally Advanced Pancreatic Cancer. Journal of Clinical Oncology 2006;24:1145–1151. [PubMed: 16505434]

DeGroot, MH. Optimal Statistical Decisions. New York: McGraw-Hill; 1970.

Ferrari SLP, Cribari-Neto F. Beta Regression for Modelling Rates and Proportions. Journal of Applied Statistics 2004;31:799–815.

Gatsonis C, Greenhouse JB. Bayesian Methods for Phase I Clinical Trials. Statistics in Medicine 1992;11:1377–1389. [PubMed: 1518998]

Goodman SN, Zahurak ML, Piantadosi S. Some Practical Improvements in the Continual Reassessment Method for Phase I Studies. Statistics in Medicine 1995;14:1149–1161. [PubMed: 7667557]

Gordon NH, Willson JKV. Using Toxicity Grades in the Design and Analysis of Cancer Phase I Clinical Trials. Statistics in Medicine 1992;11:2063–2075. [PubMed: 1293668]

Joffe S, Miller FG. Rethinking Risk-Benefit Assessment for Phase I Cancer Trials. Journal of Clinical Oncology 2006;24:2987–2990. [PubMed: 16809725]

Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A Comparison of Two Phase I Trial Designs. Statistics in Medicine 1994;13:1799–1806. [PubMed: 7997713]

O'Quigley J, Pepe M, Fisher L. Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer. Biometrics 1990;46:33–48. [PubMed: 2350571]

Potter DM. Phase I Studies of Chemotherapeutic Agents in Cancer Patients: A Review of the Designs. Journal of Biopharmaceutical Statistics 2006;16:579–604. [PubMed: 17037260]

Potthoff, RF.; George, SL. Phase I Clinical Trials With Non-Binary Toxicity Response. Duke Biostatistics and Bioinformatics (B&B) Working Paper Series, Working Paper 3. 2007. Available online at http://biostats.bepress.com/dukebiostat/papers/art3

Rogatko A, Babb JS, Wang H, Slifker MJ, Hudes GR. Patient Characteristics Compete With Dose as Predictors of Acute Treatment Toxicity in Early Phase Clinical Trials. Clinical Cancer Research 2004;10:4645–4651. [PubMed: 15269136]

Storer BE. Design and Analysis of Phase I Clinical Trials. Biometrics 1989;45:925–937. [PubMed: 2790129]

Thall PF, Lee S-J. Practical Model-Based Dose-Finding in Phase I Clinical Trials: Methods Based on Toxicity. International Journal of Gynecological Cancer 2003;13:251–261. [PubMed: 12801254]

Wang C, Chen TT, Tyan I. Designs for Phase I Cancer Clinical Trials With Differentiation of Graded Toxicity. Communications in Statistics—Theory and Methods 2000;29:975–987.

Whitehead J, Brunier H. Bayesian Decision Procedures for Dose Determining Experiments. Statistics in Medicine 1995;14:885–893. [PubMed: 7569508]

Yuan Z, Chappell R, Bailey H. The Continual Reassessment Method for Multiple Toxicity Grades: A Bayesian Quasi-Likelihood Approach. Biometrics 2007;63:173–179. [PubMed: 17447942]

APPENDICES

A. Possible Toxicity Scoring for Bevacizumab Trial

For a trial of bevacizumab in locally advanced pancreatic cancer (Crane et al. 2006), three main types of toxicity were reported: gastrointestinal (dehydration, gastritis, nausea, vomiting, anorexia, diarrhea, constipation), hematologic (neutropenia, thrombocytopenia, hypomagnesemia), and hand-foot syndrome. The highest grades of drug-related toxicity that were found for these three types were Grades 3, 4, and 2, respectively, using the Common Terminology Criteria for Adverse Events (Cancer Therapy Evaluation Program 2003). Among the 48 patients, gastrointestinal toxicities occurred most often and hematologic ones least often.

The toxicity scoring method to be depicted here, described for a single rater, is one of many that could be devised. The influence that it grants to the toxicity with the highest grade is more

than might exist with some other approaches, but less than with the usual binary measures of toxicity and generally not as much as with the toxicity index of Rogatko et al. (2004).

Let R be the number of toxicities that are to be scored. For a given patient, let G_r ($r = 1, \ldots, R$) denote the patient's grade (defined in a suitable manner) for the toxicity of type r. Let u_r denote the score for that grade, specified in such a way that the score u_r increases monotonically with the grade G_r and runs from 0 (lowest value) to 1 (highest value).

For our example, we will use $R = 3$ so that G_1, G_2, and G_3 refer, respectively, to toxicities that are gastrointestinal, hematologic, and for hand-foot syndrome. We will define G_r to be the *highest* grade of the toxicities that are included in toxicity of type r; for instance, $G_2 = 4$ for a patient who has grade 4 neutropenia but no other hematologic toxicities. The possible grades G_r are 0, 1, 2, 3, 4, and 5, for which the respective scores u_r will be taken as 0, 0.2, 0.4, 0.6, 0.8, and 1. That is, $u_r = G_r/5$ for each r. Note, though, that u_r need not be defined the same for all r, and could be defined in a customized manner if the grade of a toxicity does not properly reflect its seriousness (or lack thereof) in a given trial.

Let ω_r be a weight (> 0) chosen for toxicity of type r, such that $\sum_{r=1}^{R} w_r = 1$. Also define $z = \max_r u_r$, the maximum of the R toxicity scores, and $u = \sum_{r=1}^{R} w_r u_r$, their weighted average. For the example, we choose $\omega_1 = 0.6$, $\omega_2 = 0.3$, and $\omega_3 = 0.1$, so that gastrointestinal toxicity receives the greatest weight and hand-foot syndrome the least.

Now define two functions of z,

$$f_0(z) = (C_0 + 1)z - C_0 z^2$$

and

$$f_1(z) = (C_1 + 1)z - C_1 z^2,$$

and a toxicity response variable y, taken as

$$\begin{aligned} y &= f_0(z) + \frac{u}{z}[f_1(z) - f_0(z)] \\ &= -C_0 z^2 + [(C_0 + 1) - (C_1 - C_0)u]z \\ &\quad + (C_1 - C_0)u, \end{aligned}$$

with the understanding that $y = 0$ if $z = 0$. The constants C_0 and C_1 are to be chosen so that $-1 \le C_0 \le 0$ and $0 \le C_1 \le 1$.

Thus the formula for y provides a value that lies between $f_0(z)$ and $f_1(z)$, which are on opposite sides of the line $y = z$ (f_0 is below and f_1 is above) and which are closer to the line the closer C_0 and C_1 are to 0. From the above formulas, note that $f_0(0) = f_1(0) = 0$ and $f_0(1) = f_1(1) = 1$; that $f_0(z)$ and $f_1(z)$ both have nonnegative derivatives (for all allowable C_0 and C_1) wherever $0 \le z \le 1$; that $f_0(z) \le z$ and $f_1(z) \ge z$ wherever $0 \le z \le 1$; that y is near to $f_0(z)$ if u is near to 0 and $y = f_1(z)$ if $u = z$; and that $f_0(z) = z$ if $C_0 = 0$, $f_1(z) = z$ if $C_1 = 0$, and $y = z$ if both are 0. The formula for y produces a result that is affected heavily by z, the largest of the R toxicity scores, but also somewhat by u, their weighted average.

For our example related to the trial of Crane et al. (2006), we set $C_0 = -0.5$ and $C_1 = 1$, so that $f_0(z)$ is not as far below the line $y = z$ as $f_1(z)$ is above it. The formula for the toxicity response variable then becomes

$$y=0.5[z^2+(1-3u)z+3u].$$

We illustrate with four hypothetical patients (remember that $\omega_1 = 0.6$, $\omega_2 = 0.3$, and $\omega_3 = 0.1$):

- Patient 1 has $(G_1, G_2, G_3) = (2, 0, 0)$, so that $(u_1, u_2, u_3) = (0.4, 0, 0)$. Thus $z = 0.4$, $u = 0.24$, and $y = 0.496$.
- Patient 2 has $(G_1, G_2, G_3) = (0, 0, 2)$, so $(u_1, u_2, u_3) = (0, 0, 0.4)$. Thus $z = 0.4$, $u = 0.04$, and $y = 0.316$.
- Patient 3 has $(G_1, G_2, G_3) = (3, 4, 2)$, so $(u_1, u_2, u_3) = (0.6, 0.8, 0.4)$. Thus $z = 0.8$, $u = 0.64$, and $y = 0.912$.
- Patient 4 has $(G_1, G_2, G_3) = (0, 5, 0)$, so $(u_1, u_2, u_3) = (0, 1, 0)$. Then $z = 1$, $u = 0.3$, and $y = 1$.

B. Details for Calculating $P_{i\,x}$

To start the calculation of the posterior toxicity estimates $P_{i\,x}$, the logistic-regression software referred to in Section 5.2 will provide

$$\begin{bmatrix} a_{*0} \\ b_{*0} \end{bmatrix} = \text{Estimate of} \begin{bmatrix} a \\ b \end{bmatrix}. \tag{B.1}$$

It also produces

$$\mathbf{s}_{*0} = \begin{bmatrix} s_{aa*0} & s_{ab*0} \\ s_{ab*0} & s_{bb*0} \end{bmatrix} = \text{Estimated variance matrix of} \begin{bmatrix} a_{*0} \\ b_{*0} \end{bmatrix}. \tag{B.2}$$

Adjustments always need to be made to $\mathbf{s}_{*0}$ (B.2) to correct for the fact that $y|x$ is beta distributed, rather than binomial as assumed by the logistic-regression software. Additional adjustment, for both (B.1) and (B.2), is needed if $b_{*0} \leq 0$. Let

$$\begin{bmatrix} a_* \\ b_* \end{bmatrix} = \text{Value of} \begin{bmatrix} a_{*0} \\ b_{*0} \end{bmatrix} \text{ after application of any adjustment}$$

and

$$\mathbf{s}_* = \begin{bmatrix} s_{aa*} & s_{ab*} \\ s_{ab*} & s_{bb*} \end{bmatrix} = \text{Value of } \mathbf{s}_{*0} \text{ as adjusted.}$$

The adjustments are as follows. If $b_{*0} > 0$, then

$$\begin{bmatrix} a_* \\ b_* \end{bmatrix} = \begin{bmatrix} a_{*0} \\ b_{*0} \end{bmatrix} \text{(no adjustment),} \tag{B.3}$$

and

$$\mathbf{s}_* = V_+ \mathbf{s}_{*0}, \tag{B.4}$$

where V_+ corrects for nonbinary y. V_+ denotes the posterior mean for v [see (10)], from calculation at patient $(i - 1)$. (V_+ is set equal to E_V, the prior mean for v, to start.) Appendix E provides the method for calculating V_+.

If $b_{*0} \leq 0$, then calculate as though b and its estimate are 0. That is, set

$$\begin{bmatrix} a_* \\ b_* \end{bmatrix} = \begin{bmatrix} \log \frac{\overline{Y}}{1-\overline{Y}} \\ 0 \end{bmatrix}, \text{ where } \overline{Y} = \frac{\sum_{x=1}^{d} Y_x}{\sum_{x=1}^{d} n_x}. \tag{B.5}$$

In addition, set

$$\mathbf{s}_* = \frac{V_+}{\overline{Y}(1-\overline{Y})} \left[\sum_{x=1}^{d} \begin{bmatrix} n_x & n_x x \\ n_x x & n_x x^2 \end{bmatrix} \right]^{-1}. \tag{B.6}$$

The argument to justify (B.5) and (B.6) is as follows. First, one can take the estimate of b to be $b_* = 0$ since, with $b_{*0} \leq 0$, any other value would seem unreasonable. Next, if $b = 0$ so that (1) and (3) are horizontal, then the mean toxicity is constant for all doses and its constant value is estimated by $\overline{Y}$. From the middle expression in (3) with $b = 0$, $\overline{Y}$ estimates $e^a (1 + e^a)$, which indicates that a itself is estimated by a_* as given in (B.5). Under a pure binomial logistic-regression model corresponding to (1), the estimated variance matrix for the estimate of (a, b), as obtained from the matrix of second derivatives of the log-likelihood function, is

$$\left[\sum_{x=1}^{d}\begin{bmatrix} n_x p_x(1-p_x) & n_x x p_x(1-p_x) \\ n_x x p_x(1-p_x) & n_x x^2 p_x(1-p_x) \end{bmatrix}\right]^{-1}, \tag{B.7}$$

where p_x is the same as the middle expression of (3) but with a and b replaced by their estimates. Finally, if p_x uses (B.5) for its estimates of a and b (so that p_x becomes $\bar{Y}$ for all x), and if [as with (B.4)] the multiplicative factor V_+ is applied to correct for the fact that the response is not really binary, then (B.6) follows easily from (B.7).

Now let

$$\begin{bmatrix} A_+ \\ B_+ \end{bmatrix} = \text{Posterior mean vector for} \begin{bmatrix} a \\ b \end{bmatrix}$$

and

$$\mathbf{S}_+ = \begin{bmatrix} S_{AA+} & S_{AB+} \\ S_{AB+} & S_{BB+} \end{bmatrix} = \text{Posterior variance matrix for} \begin{bmatrix} a \\ b \end{bmatrix}.$$

Then (cf. DeGroot 1970, p. 175) it is suitable to set

$$\mathbf{S}_+ = (\mathbf{s}_*^{-1} + \mathbf{S}^{-1})^{-1} \tag{B.8}$$

and

$$\begin{bmatrix} A_+ \\ B_+ \end{bmatrix} = \mathbf{S}_+ \left\{ \mathbf{s}_*^{-1} \begin{bmatrix} a_* \\ b_* \end{bmatrix} + \mathbf{S}^{-1} \begin{bmatrix} E_A \\ E_B \end{bmatrix} \right\}, \tag{B.9}$$

where the expressions for $\mathbf{s}_*$, a_* and b_*, $\mathbf{S}$, and E_A in (B.8) and (B.9) are given, respectively, by (B.4) or (B.6), (B.3) or (B.5), (9), and (8).

Let h_* denote the estimate of h from the sample, and let H_+ denote the posterior estimate of h. Their formulas [refer to (2)] are then

$$h_* = \frac{\log k - a_*}{b_*} \tag{B.10}$$

and

$$H_+ = \frac{\log k - A_+}{B_+}, \tag{B.11}$$

where the statistics on the right sides of (B.10) and (B.11) come from (B.3) and (B.9), respectively. There is no h_* if $b_{*0} \leq 0$, so (B.10) applies only if $b_{*0} > 0$. On the other hand, (B.11) applies even if $b_{*0} \leq 0$ (assuming that $B_+ > 0$).

Note that a formula for H_+ that would use a weighted combination of h_* (B.10) and the prior mean E_H does not work if $b_{*0} \leq 0$.

We can now make use of the rightmost expression of (3) to write the formula for P_{ix}, the posterior toxicity estimate (at each dose x) that is used in the utility function (11). It is

$$P_{ix} = \frac{k e^{B_+(x - H_+)}}{1 + k e^{B_+(x - H_+)}}, \tag{B.12}$$

where B_+ comes from (B.9) and H_+ from (B.11).

C. Details for Calculating $S_{H\,H+i\,x}$

As brought out in Section 5.3, when patient i arrives, the doses for all $(i - 1)$ previous patients (even those who are incomplete) are known. These doses are x_q ($q = 1, \ldots, i - 1$). Define

$$\mathbf{G}_{+1i} = \sum_{q=1}^{i-1} \begin{bmatrix} P_{ix_q}(1 - P_{ix_q}) & x_q P_{ix_q}(1 - P_{ix_q}) \\ x_q P_{ix_q}(1 - P_{ix_q}) & x_q^2 P_{ix_q}(1 - P_{ix_q}) \end{bmatrix}, \tag{C.1}$$

where P_{ix_q} is obtained from (B.12). Further, for $x = 1, \ldots, d$, define

$$\mathbf{G}_{+2ix} = \begin{bmatrix} P_{ix}(1 - P_{ix}) & x P_{ix}(1 - P_{ix}) \\ x P_{ix}(1 - P_{ix}) & x^2 P_{ix}(1 - P_{ix}) \end{bmatrix}. \tag{C.2}$$

If patient i were to receive dose x, then the posterior a variance matrix for $\begin{bmatrix} a \\ b \end{bmatrix}$ would be

$$\mathbf{S}_{+0ix} = (\mathbf{G}_{+1i} + \mathbf{G}_{+2ix})^{-1} \tag{C.3}$$

if (hypothetically) one were using, instead of (1), a pure (binomial) logistic-regression model

$$\log \frac{\Pr\{y = 1 | x\}}{1 - \Pr\{y = 1 | x\}} = a + bx, \tag{C.4}$$

where all responses would be 0 or 1 and P and $P_{i\,x}$ in (C.1)–(C.2) would estimate $\Pr\{y = 1|x\}$. The expressions (C.1)–(C.2) invoke the matrix of second derivatives of the log-likelihood function, similarly to (B.7). But note that (C.1)–(C.3) reflect all previous patients whereas (B.7) can use only the complete patients.

Our model is more general than (C.4), because the distribution of $y|x$ is beta rather than binomial and $P_{i\,x}$ estimates $\mathrm{E}(y|x)$ rather than $\Pr\{y = 1|x\}$. Thus, (C.3) needs to be corrected, in the same way as in (B.4). The corrected posterior variance matrix for $\begin{bmatrix} a \\ b \end{bmatrix}$ is

$$\mathbf{S}_{+ix} = V_{+}\mathbf{S}_{+0ix}. \tag{C.5}$$

Finally, for (11) we need the posterior variance for h if patient i were to receive dose x. We refer to (2) and apply the delta approximation. Plugging in B_{+}, H_{+}, and (C.5) then leads to

$$S_{HH+ix} = \frac{1}{B_{+}^{2}}\begin{bmatrix} 1 & H_{+} \end{bmatrix}\mathbf{S}_{+ix}\begin{bmatrix} 1 \\ H_{+} \end{bmatrix}. \tag{C.6}$$

Note that the posterior variance matrix $\mathbf{S}_{+}$ is not used to get $S_{H\,H+\,i\,x}$, because $\mathbf{S}_{+}$ does not take into account the doses of incomplete patients. $\mathbf{S}_{+}$ does not reflect patient i and does not always reflect all $(i - 1)$ of the previous patients.

D. Details for Final Point and Interval Estimation

Calculations for the final Bayesian and frequentist estimation described in Section 5.4 are as follows. In the Bayesian approach, one uses the completed results from all n patients to obtain (B.3) or (B.5), (B.4) or (B.6), (B.8), and (B.9), after which the final posterior MTD estimate, H_{+}, is found from (B.11). The associated posterior variance is given by

$$S_{HH+} = \frac{1}{B_{+}^{2}}\begin{bmatrix} 1 & H_{+} \end{bmatrix}\mathbf{S}_{+}\begin{bmatrix} 1 \\ H_{+} \end{bmatrix}, \tag{D.1}$$

a formula that is analogous to (C.6).

In the frequentist approach, the final MTD estimate, h_{*}, is calculated from (B.10) after using the completed results from all n patients to obtain (B.3) and (B.4). The associated variance estimate, denoted by s_{hh*}, is

$$s_{hh*} = \frac{1}{b_{*}^{2}}\begin{bmatrix} 1 & h_{*} \end{bmatrix}\mathbf{s}_{*}\begin{bmatrix} 1 \\ h_{*} \end{bmatrix}, \tag{D.2}$$

which is analogous to (C.6) and (D.1). Note that h_{*} (B.10) need not be calculated except at the end of the trial (and not even then if the final estimation is Bayesian), whereas H_{+} (B.11) always has to be obtained anew prior to selecting the dose for each patient (regardless of whether the final estimation is Bayesian or frequentist).

One cannot use the frequentist formulas (B.10) and (D.2) at the end of the trial if $b_{*0} \leq 0$ at that point. But the Bayesian formulas (B.11) and (D.1) still work, of course (unless $B_+ \leq 0$, an outcome that would appear to be rare if not impossible). On the other hand, even the Bayesian results may not be very meaningful if the full trial results are so poor that $b_{*0} \leq 0$ at the end.

E. Formula for V_+

V_+, the posterior estimate of v of (10), is used in (B.4), (B.6), and (C.5). Its calculation is always one patient behind. That is, the new V_+ is calculated *after* the $P_{i\,x}$ values (B.12) have already been found for use with patient i [but before (C.5) and (C.6) are obtained]. It is thus used for getting (B.12) for patient $(i + 1)$ rather than patient i. Although the new V_+ is found using (B. 12), (B.12) in turn depends on the old V_+ through (B.4) or (B.6). For the *first* patient for whom (B.12) is calculated, no V_+ is yet available, so E_V, the prior mean, has to be used instead.

V_+, unlike the estimates of parameters related to the dose–response curve, should change little with added patients if the investigator, through experience, has a good sense of the value of v and is therefore able to give a small value to $S_{V\,V}$, the prior variance. Thus the fact that V_+ is one patient behind will not matter much if it is stable.

As a precaution that may make little difference when V_+ is stable, we perform the final calculations after the end of the study (see Section 5.4) twice in succession. The first time is solely for the purpose of getting a fresh V_+. The second time, though, we obtain the values of (B.11) and (D.1), or of (B.10) and (D.2), that are actually to be reported.

If all responses are certain to be either 0 or 1, then $E_V = 1$, $S_{V\,V} = 0$, and V_+ will be equal to 1 throughout the study. For this degenerate case, the material that follows does not apply.

Otherwise, though, one can obtain V_+ through the formula

$$V_+ = \left(\frac{E_V}{S_{VV}} + \frac{\upsilon_*}{s_{\upsilon\upsilon *}}\right) \Big/ \left(\frac{1}{S_{VV}} + \frac{1}{s_{\upsilon\upsilon *}}\right), \tag{E.1}$$

where υ_* is the sample estimate of v after (B.12) is found and $s_{\upsilon\upsilon *}$ is the estimated variance of υ_*. The following (partially heuristic) development provides our way for getting υ_* and $s_{\upsilon\upsilon *}$.

For $x = 1, \ldots, d$ and $j = 1, \ldots, n_x$ (in the calculations for use with patient i), define

$$c_{xj} = \frac{(y_{xj} - m_x)^2}{m_x(1 - m_x)},$$

where $y_{x\,j}$ is the response score for the jth of the n_x complete patients on dose x and m_x is the (true) mean toxicity at dose x as given by (3). From (4) and (10) it follows at once that

$$\mathrm{E}(c_{xj}) = \upsilon. \tag{E.2}$$

After some lengthy but straightforward algebra involving the expectations of the first four powers of $y_{x\,j}$ along with (5)–(7), one finds that

$$\operatorname{var}(c_{xj}) = \frac{2}{(t+1)^2(t+2)(t+3)} \times \left[(t^2 - 10t - 12) + \frac{3(t+1)}{m_x(1-m_x)} \right].$$

Now define

$$v_{*0} = \frac{\sum_{x=1}^{d} \sum_{j=1}^{n_x} \frac{c_{xj}}{\operatorname{var}(c_{xj})}}{\sum_{x=1}^{d} \sum_{j=1}^{n_x} \frac{1}{\operatorname{var}(c_{xj})}} = \frac{\sum_{x=1}^{d} \frac{\sum_{j=1}^{n_x} (y_{xj} - m_x)^2}{(t^2-10t-12)m_x(1-m_x)+3(t+1)}}{\sum_{x=1}^{d} \frac{n_x m_x(1-m_x)}{(t^2-10t-12)m_x(1-m_x)+3(t+1)}}, \tag{E.3}$$

and note that $E(v_{*0}) = v$ by virtue of (E.2). Let s_{vv*0} denote the variance of v_{*0}. Then

$$s_{vv*0} = \frac{1}{\sum_{x=1}^{d} \sum_{j=1}^{n_x} \frac{1}{\operatorname{var}(c_{xj})}} = \frac{2}{(t+1)^2(t+2)(t+3) \sum_{x=1}^{d} \frac{n_x m_x(1-m_x)}{(t^2-10t-12)m_x(1-m_x)+3(t+1)}}. \tag{E.4}$$

Define $T_+ = (1 / V_{+\,(i-1)}) - 1$, where $V_{+\,(i-1)}$ denotes the latest value of V_+, obtained during the calculations for patient $(i-1)$. Next, define v_* to be the same thing as v_{*0} of (E.3) except that m_x and t are replaced by their respective estimates $P_{i\,x}$ (B.12) and T_+, and except that v_* is to be changed to 1 if (as is possible) it would otherwise turn out to be > 1. In addition, define s_{vv*} to be the same thing as s_{vv*0} of (E.4) except that (again) m_x and t are replaced (respectively) by $P_{i\,x}$ and T_+, and except that the result is multiplied by $\Sigma_x\, n_x / (\Sigma_x\, n_x - 2)$ as a very rough attempt to adjust for the fact that two parameters in (3) are estimated from the data. Then it is reasonable to use v_* and s_{vv*} in (E.1) to obtain the new value of V_+.

Table 1

Average toxicities (percent) for all combinations of *M*, *h*, *Q*, and *v*.

		h = 2		h = 3.5		h = 5	
		Q = 0.2	**Q = 0.6**	**Q = 0.2**	**Q = 0.6**	**Q = 0.2**	**Q = 0.6**
$M = 0.2$	$\upsilon = 0.5$	23.0	22.0	19.8	18.3	16.4	13.6
	$\upsilon = 0.95$	24.5	21.7	18.8	13.9	14.0	5.8
$M = 1/3$	$\upsilon = 0.5$	35.8	34.5	32.3	29.5	28.3	22.1
	$\upsilon = 0.95$	36.9	33.2	31.4	21.1	24.7	7.9

Table 2

Percentage of frequentist 95% two-sided confidence intervals that do not include h (the MTD), with certain runs excluded.

	Number of excluded runs			
Condition	**With $b_{*0} \leq 0$**	**With $h_* < 0$ or > 7**	**Number of included runs (% included)**	**% of intervals without h**
All runs	209	237	9,154 (95%)	8.88
$h = 2$	30	27	3,143 (98%)	7.03
$h = 3.5$	38	40	3,122 (98%)	10.28
$h = 5$	141	170	2,889 (90%)	9.38
$b = 0.7$	147	187	4,466 (93%)	10.48
$b = 1.4$	62	50	4,688 (98%)	7.36
$v = 0.5$	15	72	4,713 (98%)	8.21
$v = 0.95$	194	165	4,441 (93%)	9.59
$M = 0.2$	124	114	4,562 (95%)	8.86
$M = 1/3$	85	123	4,592 (96%)	8.91
$n = 25$	164	136	4,500 (94%)	9.27
$n = 50$	45	101	4,654 (97%)	8.51
$Q = 0.2$	16	57	4,727 (98%)	9.37
$Q = 0.6$	193	180	4,427 (92%)	8.36
$L = 1$	103	116	4,581 (95%)	8.93
$L = 3$	106	121	4,573 (95%)	8.83

Table 3

Horizontal and (expressed as percent) vertical bias and absolute error, and percentage of runs where MTD estimate misses target by ≥ 0.5, for frequentist MTD estimation.

	Average bias		Average absolute error		
Condition	**Horizontal**	**Vertical**	**Horizontal**	**Vertical**	**% of runs where $\lvert h_* - h \rvert \geq 0.5$**
All runs	0.07	2.1	0.42	7.4	26.7
$h = 2$	0.00	0.6	0.33	5.6	20.0
$h = 3.5$	0.07	1.8	0.40	6.9	24.9
$h = 5$	0.14	3.8	0.53	9.8	35.1
$b = 0.7$	0.12	2.6	0.56	7.6	39.2
$b = 1.4$	0.02	1.6	0.28	7.3	14.1
$v = 0.5$	0.03	0.8	0.30	5.2	17.3
$v = 0.95$	0.12	3.3	0.54	9.6	36.1
$M = 0.2$	0.05	2.0	0.45	7.0	30.0
$M = 1/3$	0.09	2.2	0.39	7.9	23.3
$n = 25$	0.09	3.0	0.53	9.3	33.0
$n = 50$	0.06	1.6	0.37	6.5	20.3
$Q = 0.2$	−0.02	0.4	0.33	5.7	21.7
$Q = 0.6$	0.16	3.8	0.51	9.2	31.6
$L = 1$	0.07	2.1	0.42	7.4	26.2
$L = 3$	0.07	2.1	0.42	7.5	27.1