Dose-Finding Based on Multiple Toxicities in a Soft Tissue Sarcoma Trial

B. Nebiyou BEKELE and Peter F. THALL

The scientific goal of a phase I oncology trial of a new chemotherapeutic agent is to find a dose with an acceptable level of toxicity. For ethical reasons, dose-finding is done adaptively, with doses chosen for successive cohorts of patients based on the data obtained from previous cohorts. Typically, patients are at risk for several qualitatively different toxicities, each occurring at several possible severity levels. In this article, we describe how we addressed the dose-finding problem in a phase I trial of gemcitabine for treatment of soft tissue sarcoma. The oncologists planning the trial wanted to account for differences in importance among the toxicities that they had identified. They also requested that the dose-finding method utilize the fact that a low-grade toxicity observed at a given dose, although not dose-limiting, provides a warning that a higher grade of that toxicity is likely to occur at a higher dose. Conventional phase I methods reduce each type of toxicity to an indicator of its occurrence at or above a severity level considered dose-limiting, define "toxicity" as the maximum of these indicators, and base dose-finding on that single binary variable. Because conventional methods do not address the aforementioned concerns, we developed a Bayesian method for dose-finding in the sarcoma trial based on a vector of correlated, ordinal-valued toxicities with severity levels varying with dose. We also developed a method for jointly eliciting the prior, a vector of weights quantifying the clinical importance of each level of each type of toxicity, and a target total toxicity burden acceptable to the physicians. Our method assigns each cohort the dose with a current posterior mean total toxicity burden closest to the target. The elicitation process is iterative, with the oncologists repeatedly shown the algorithm's behavior and asked to adjust their weights to ensure that the statistical decisions reflect appropriate clinical behavior. We describe how this methodology has worked in the sarcoma trial, present simulations and sensitivity analyses of the trial under several clinical scenarios, and provide guidelines for general application.

KEY WORDS: Adaptive design; Bayesian inference; Latent variable; Markov chain Monte Carlo.

1. INTRODUCTION

The primary goal of a phase I clinical trial of a new chemotherapeutic agent in oncology is to determine a dose with acceptable toxicity. Because safety is a central concern in such trials, typically patients are treated in successive cohorts, with the dose for each cohort chosen adaptively based on the dose-toxicity data from previous patients in the trial. Thus, a phase I design must provide both an algorithm for sequentially assigning doses to patient cohorts and a rule for selecting a dose, usually called the "maximum tolerated dose" (MTD), at the end of the trial. For convenience, here we refer to these two related statistical problems together as "dose-finding." A difficulty in phase I, especially acute early in the trial, is that doses must be assigned to patients based on very little data. Because any ethically reasonable algorithm must de-escalate when a dose is found to be unacceptably toxic, and only a limited number of toxicities are permitted, little or no data are available on doses with high toxicity probabilities. Thus the dose-finding problem is scientifically difficult due to ethical constraints. Numerous statistical designs for phase I trials have been proposed (Storer 1989; O'Quigley, Pepe, and Fisher 1990; Durham and Fluornoy 1994; Babb, Rogatko, and Zacks 1998; Piantadosi, Fisher, and Grossman 1998; Gasparini and Eisele 2000). Each of these approaches characterizes toxicity as a binary variable, with the underlying statistical model and the algorithm for trial conduct based on the probability of toxicity as a function of dose.

This article is motivated by the problem of designing a phase I trial of presurgical gemcitabine with external beam radiation (EBR) for patients with soft tissue sarcoma. In planning the trial, we worked with a team of three oncologists who have extensive experience in sarcoma treatment and share responsibility for conduct of this trial. The trial was activated in January 2002 and is currently ongoing. Each patient receives a fixed dose of 50 cGy external beam radiation and 1 of 10 doses of gemcitabine, 100, 200, ..., or 1,000 mg/m^2. In virtually all oncology chemotherapy settings, the patient is at risk of several different types of toxicity, each occurring at several possible severity levels. These levels typically are expressed as "grades," taking on integer values from 0 (indicating no toxicity of that type) to 4 (the most severe level). The toxicities used as a basis for dose-finding in the trial are summarized in Table 1, which gives the grades of each type of toxicity that the oncologists consider to be dose-limiting. We explain the "severity weight" listed beside each grade of each toxicity later. These weights play a central role in all that follows. Because it may take up to 6 weeks to evaluate all of the possible toxicities in each patient, to facilitate trial conduct, the cohort size is allowed to vary between three and four, as follows. If the first three patients in a cohort have had all of their toxicities evaluated before a fourth patient is accrued, then that cohort is considered complete, the next gemcitabine dose is chosen, and treatment of the next cohort with that dose is begun. As in many phase I trials, the safety constraint is imposed that no untried dose may be skipped when escalating. The trial will end when at least 36 patients have been accrued and evaluated.

The methodology that we developed for dose-finding in the sarcoma trial is a radical departure from conventional phase I methods. This was motivated by several concerns expressed by the oncologists. They requested that the dose-finding method account for the fact that, clinically, the toxicities that they had identified are not equally important. Additionally, the different toxicities do not occur independently. For example, fatigue and nausea/vomiting are likely to occur together, as are

B. Nebiyou Bekele is Assistant Professor (E-mail: *bbekele@mdanderson.org*); Peter F. Thall is Professor, Department of Biostatistics, Box 447, University of Texas, M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030. The authors are grateful to P. Pisters, M. Ballo, and S. Patel of The M. D. Anderson Cancer Center for numerous discussions that provided essential input for this research. They owe particular thanks to the editor, Thomas Louis, for his encouragment and constructive comments. They also thank an associate editor and two referees for their critical comments, which led to an improvement in the article. Dr. Thall's work was supported in part by NCI grant RO1 CA 83932.

Journal of the American Statistical Association
March 2004, Vol. 99, No. 465, Applications and Case Studies
DOI 10.1198/016214504000000043

Table 1. Toxicities and Severity Weights in the Sarcoma Trial

	Type of toxicity	Grade	Severity weight
1	Myelosuppression without fever	3	1.0
		4	1.5
	Myelosuppression with fever	3	5.0
		4	6.0
2	Dermatitis	3	2.5
		4	6.0
3	Liver	2	2.0
		3	3.0
		4	6.0
4	Nausea/vomiting	3	1.5
		4	2.0
5	Fatigue	3	.5
		4	1.0

myelosuppression (i.e., low blood cell count, associated with suppression of normal bone marrow function) and fever. They also requested that the dose-finding method utilize the fact that a low-grade toxicity observed at a given dose, although not dose-limiting, provides a warning that a higher grade of that toxicity is more likely to occur at a higher dose level. To explain why these concerns motivated us to develop the methodology described in this article, we first need to explain the limitations of conventional methods.

A typical protocol for a phase I oncology trial lists the possible toxicities that must be monitored. This list often includes transient conditions, such as fatigue, nausea/vomiting, myelosuppression, thrombocytopenia (low platelet count), fever, infection, and dysfunction of specific organs, and irreversible toxicities, such as permanent organ damage or death. In general, the different toxicities do not occur independently. The conventional approach to dose-finding is to reduce each type of toxicity to an indictor of its occurrence at or above a severity level considered dose-limiting, define "toxicity" as the maximum or some other binary-valued function of these indicators, and base the model and dose-finding method on that binary variable. Most phase I protocols define "toxicity" as the occurrence at grade 3 or 4 of several listed toxicities. Although it is logistically convenient to reduce the ordinal scale of given toxicity to the binary variable for which grades 0, 1, or 2 are "no toxicity" and grades 3 or 4 are "toxicity," this common practice discards useful information. For example, if several patients experience a grade 2 toxicity of a given type at a dose level k, then a typical method based on the foregoing binary variable would escalate to level $k + 1$ as if no toxicities had occurred at level k. Clearly, a probability model that distinguishes between grades 0, 1, and 2 rather than combining them as the event "no toxicity" should provide a more reliable basis for predicting the jump from grade 2 to grade 3 or higher as the dose is increased from k to $k + 1$. Furthermore, if each of several different types of toxicity has been defined as a binary variable, then defining "toxicity" as the maximum of these indicators implicitly assumes that the different toxicities are exchangeable and hence equally important. For example, this definition does not distinguish between a patient who has grade 3 fatigue and a patient who has suffered complete kidney failure. Thus dichotomizing all toxicities at the same grade and assuming that they are exchangeable leads to conclusions that simply do not make sense.

To provide a dose-finding method for the sarcoma trial addressing all of these issues, we found it necessary to go beyond the conventional phase I framework. We characterized patient outcome as a vector of correlated, ordinal-valued toxicities by applying the multivariate ordinal probit regression model of Chen and Dey (2000), extended to allow the different toxicities to have different numbers of severity levels. To address the oncologists' concern that qualitatively different toxicities often are not equally important, we used numerical weights to characterize the clinical importance of each severity level of each type of toxicity. During several sessions with the oncologists, we developed a method for jointly eliciting the prior for the model parameters and the toxicity severity importance weights. We defined the total toxicity burden of a patient as the sum of the weights of all toxicities experienced by that patient. This provides the basis for the dose-finding algorithm. We used the elicited weights to identify a target total toxicity burden by constructing a set of hypothetical dose-toxicity scenarios and asking the oncologists in each case whether they would escalate, repeat the same dose, or de-escalate for the next cohort. Our method assigns to each cohort the dose with the current posterior mean total toxicity burden closest to the target. Because this method quantifies the oncologists' experiences in dealing with multiple toxicities in the clinic and incorporates this information into the dose-finding algorithm, it reflects actual clinical practice more closely than do conventional methods.

The methodology evolved iteratively over the course of several sessions with the oncologists. Initially, they specified the five toxicities myelosuppression (M), fever, dermatitis (D), nausea/vomiting (N), and fatigue (F) as binary variables, we defined "toxicity" in the conventional manner as the maximum of these five indicators, and we constructed a continual reassessment method (crm) design (O'Quigley et al. 1990) with a target toxicity rate of 30%. At the second session, during which we presented the crm design, the issues arose that M is positively associated with fever and that M with fever (M^+) is a much more severe event than M without fever (M^-). This led us to combine these two variables into the five-level ordinal toxicity $M_0 < M_3^- < M_4^- < M_3^+ < M_4^+$, where grade is denoted by a subscript and M_0 denotes no M of grade > 2. This in turn motivated the oncologists to refine the other toxicities, so that D, N, and F each became a trinary variable. They also added liver toxicity (L), defined as a four-level variable (see Table 1). We next elicited numerical weights to quantify the clinical importance of each level of each type of toxicity. We subsequently formulated the probability model, developed the idea of using the individual patient's total toxicity burden in terms of the weights as the basis for dose-finding, and elicited a target total toxicity burden for dose-finding and a prior from the oncologists. We repeated this process over the course of several sessions, until the algorithm made decisions that the oncologists considered clinically sensible under all of the scenarios.

In Section 2 we present the probability model and dose-finding algorithm. In Section 3 we describe a method for simultaneously eliciting the prior, toxicity weights, and target total toxicity burden. In Section 4 we return to the sarcoma trial, including specifics of the elicitation process, the current data, and how the algorithm has behaved to date. We present a simulation study in Section 5, and close with a discussion in Section 6.

2. MODEL AND DOSE–FINDING ALGORITHM

2.1 Modeling Objectives

A statistical model for outcome-adaptive decision making in a clinical trial must provide a practical framework for repeatedly incorporating new data and computing decision criteria. For the sarcoma trial, where the decision is which dose to give the next cohort, we required a model characterizing how the probabilities of the severity levels of each type of toxicity vary with dose, while also accounting for association among the toxicities. In theory, many parametric multivariate ordinal regression models have these properties. A difficulty in phase I trials is that doses must be chosen based on very little data; this problem is especially severe early in the trial. Because the goal is dose-finding rather than model fitting, however, any reasonably tractable model with the aforementioned properties is acceptable, provided that the dose-finding method works well under the assumed model. Because we use computer simulation of the trial design to examine its operating characteristics and calibrate its parameters before actual trial conduct, which requires that the model be fit thousands of times, computational tractability also is an essential requirement.

To obtain a model with all of these properties for the sarcoma trial, we applied the Bayesian multivariate ordinal probit model of Chen and Dey (2000). This is a member of the general class of models developed by Albert and Chib (1993) and Chib and Greenberg (1998), that uses a vector of correlated latent Gaussian variables to induce association among binary, categorical, or ordinal outcomes. The extension of the multivariate ordinal version of this model to accommodate ordinal variables with different numbers of levels is straightforward. Although this model may appear somewhat complicated, we used as simple a version as possible without sacrificing any of the structure described earlier, and in fact found this model quite tractable.

2.2 Probability Model

Let $\mathbf{Y} = (Y_1, \ldots, Y_J)$ denote the vector of ordinal toxicity variables. The jth type of toxicity, Y_j, takes on one of the $C_j + 1$ values $\{y_{j,0}, y_{j,1}, \ldots, y_{j,C_j}\}$, where $y_{j,k}$ is the kth most severe level for $k = 0, \ldots, C_j$. In the sarcoma trial, $J = 5$, and if Y_j refers to, say, dermatitis, then $C_j = 2$, $y_{j,0}$ denotes grade 0, 1, or 2 dermatitis; $y_{j,1}$ denotes grade 3 dermatitis; and $y_{j,2}$ denotes grade 4 dermatitis. Binary Y_j corresponds to $C_j = 1$. To improve numerical stability, we replaced each gemcitabine dose d by $x = \log(d/1{,}000)$, and we refer to x as the "dose." We model association among the Y_j's by introducing the vector of correlated latent variables $\mathbf{Z}^{J\times 1} = (Z_1, \ldots, Z_J)$, which is assumed to be multivariate normal with $\mathrm{E}(Z_j) = \beta_{j,0} + x\beta_{j,1}$ for each j, all variances equal to 1, and correlation matrix $\boldsymbol{\Omega}$. In matrix notation, $\mathrm{E}(\mathbf{Z}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}^{J\times 2J}$ is the block-diagonal matrix with jth block $(1\ x)$ and $\boldsymbol{\beta}^{2J\times 1} = (\beta_{1,0}, \beta_{1,1}, \ldots, \beta_{J,0}, \beta_{J,1})$. The latent variable vector $\mathbf{Z}$ determines the observed outcome vector $\mathbf{Y}$ via the conditions

$$Y_j = y_{j,k} \quad \text{if} \quad \gamma_{j,k} \le Z_j < \gamma_{j,k+1} \quad \text{for } k = 0, 1, \ldots, C_j \text{ and } j = 1, \ldots, J,$$

where the cutoff parameters $\boldsymbol{\gamma}_j^{C_j\times 1} = (\gamma_{j,1}, \ldots, \gamma_{j,C_j})$ must satisfy the constraint $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,C_j} < \gamma_{j,C_j+1} = +\infty$. Denote $A_{j,k} = (\gamma_{j,k}, \gamma_{j,k+1}]$ and $\boldsymbol{\gamma}^{C_+\times 1} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J)$, where $C_+ = C_1 + \cdots + C_J$. We require the variance-covariance matrix $\boldsymbol{\Omega}$ of $\mathbf{Z}$ to be its correlation matrix to ensure identifiability of the posterior distributions, which also requires that $\gamma_{j,1} \equiv 0$. Because $\gamma_{j,0} = -\infty$, $\gamma_{j,C_j+1} = +\infty$, and $\gamma_{j,1} = 0$, if $C_j > 1$, then there are only $C_j - 1$ random cutpoint parameters. Thus, although $\boldsymbol{\gamma}$ has C_+ entries, it actually contains only $\sum_{j=1}^{J} \mathbf{1}(C_j > 1)(C_j - 1)$ parameters, where $\mathbf{1}(A)$ is the indicator of the event A. Denoting the model parameter vector by $\boldsymbol{\theta}$, the marginal distribution of Y_j for a patient treated with dose x is thus

$$\begin{aligned}\pi_{j,k}(x, \boldsymbol{\theta}) &= \Pr(Y_j = y_{j,k} | x, \boldsymbol{\theta}) \\ &= \Phi\{\gamma_{j,k+1} - (\beta_{j,0} + \beta_{j,1}x)\} \\ &\quad - \Phi\{\gamma_{j,k} - (\beta_{j,0} + \beta_{j,1}x)\}, \end{aligned} \tag{1}$$

where Φ is the standard normal cdf. Denote $\boldsymbol{\pi}_j(x, \boldsymbol{\theta}) = (\pi_{j,1}(x, \boldsymbol{\theta}), \ldots, \pi_{j,C_j}(x, \boldsymbol{\theta}))$ for each $j = 1, \ldots, J$ and $\boldsymbol{\pi}(x, \boldsymbol{\theta}) = (\boldsymbol{\pi}_1(x, \boldsymbol{\theta}), \ldots, \boldsymbol{\pi}_J(x, \boldsymbol{\theta}))$. Let $\phi_{\mathbf{W}}(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the pdf of a multivariate normal random vector $\mathbf{W}$ with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and write $\mathbf{W} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a given vector of toxicity severity levels $\mathbf{k} = (k_1, \ldots, k_J)$, the corresponding outcome of $\mathbf{Y}$ is $\mathbf{y}(\mathbf{k}) = (y_{1,k_1}, \ldots, y_{J,k_J})$, and this would arise from the J-dimensional set $\mathbf{A}(\mathbf{k}, \boldsymbol{\gamma}) = A_{1,k_1} \times \cdots \times A_{J,k_J}$ of $\mathbf{Z}$ values. Thus a single patient's contribution to the likelihood is given by

$$\begin{aligned}&\mathcal{L}(\mathbf{Y} | \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Omega}, x) \\ &\quad = \prod_{k_1=0}^{C_1} \cdots \prod_{k_J=0}^{C_J} \left\{ \int_{\mathbf{A}(\mathbf{k}, \boldsymbol{\gamma})} \phi_{\mathbf{Z}}(\mathbf{z} | \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Omega})\, d\mathbf{z} \right\}^{\mathbf{1}[\mathbf{Y} = \mathbf{y}(\mathbf{k})]}, \end{aligned} \tag{2}$$

which shows how $\mathbf{Z}$ induces association among the elements of $\mathbf{Y}$ through $\boldsymbol{\Omega}$. Let $x_{(i)}$ denote the ith patient's dose, with $\mathbf{X}_i$ the corresponding matrix. The likelihood for n patients is obtained by substituting $\mathbf{Y} = \mathbf{Y}_i$, $x = x_{(i)}$, and $\mathbf{X} = \mathbf{X}_i$ in (2) and taking the product over $i = 1, \ldots, n$.

Denote the $J(J-1)/2$ unique off-diagonal elements of $\boldsymbol{\Omega}$ by $\boldsymbol{\rho} = (\rho_{1,2}, \rho_{1,3}, \ldots, \rho_{J-1,J})$ and the cutpoint parameter vector by $\boldsymbol{\gamma}$, so that the model parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho})$. A priori, we assume $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, subject to the constraint $\Pr(\beta_{j,1} > 0) = 1$ for all $j = 1, \ldots, J$. That is, we abuse notation in that the prior of $\boldsymbol{\beta}$ is $2J$-variate normal with J entries truncated at 0, but $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ correspond to the untruncated $2J$-variate normal. This constraint ensures that $\Pr(Y_j > y_{j,k} | x) = 1 - \Phi\{\gamma_{j,k} - (\beta_{j,0} + \beta_{j,1}x)\}$ increases with x for each j and $k > 1$, which is necessary for the model to make sense. For each j with $C_j > 1$, we assume that the parameters $\{\gamma_{j,2}, \ldots, \gamma_{j,C_j}\}$ follow independent, uninformative priors on the domain $[0, 10]$, with each $g(\boldsymbol{\gamma}_j) \propto 1$, subject to the constraint $0 < \gamma_{j,2} < \gamma_{j,3} < \cdots < \gamma_{j,C_j}$. The upper limit 10 on the support of the distribution of the $\gamma_{j,k}$'s is chosen for numerical convenience because, relative to a standard normal, the probability mass beyond 10 is vanishingly small. We assume that the elements of $\boldsymbol{\rho}$ are iid $\mathrm{N}(0, 1{,}000)$, truncated to the support $[-1, +1]$, with $\boldsymbol{\Omega}$ positive definite. We describe method for specifying a prior on $\boldsymbol{\beta}$ using elicited information in Section 3.

2.3 Dose-Finding Algorithm

If doses are to be chosen based on multivariate toxicity data, then inevitably some form of dimension reduction must be carried out to obtain a real-valued criterion to use as a basis for deciding whether to escalate, stay at the same dose, or de-escalate. The conventional approach, which reduces each Y_j to $\mathbf{1}(Y_j \geq y_{j,k})$ for a toxicity level $y_{j,k}$ of type j considered by the oncologists to be dose limiting and defines "toxicity" as the maximum of the J indicators, suffers from several pathological properties. Our proposed alternative approach does away with these pathologies, in part by incorporating medical knowledge into the dimension-reduction process. The method requires that positive-valued numerical weights that characterize the importance of each level of each type of toxicity be elicited from the oncologist. For each $j = 1, \ldots, J$, denote the elicited weight of toxicity type j occurring at severity level $y_{j,k}$ by $w_{j,k}$, with $\mathbf{w}_j = (w_{j,1}, \ldots, w_{j,C_j})$ and $\mathbf{w}^{C_+ - J \times 1} = (\mathbf{w}_1, \ldots, \mathbf{w}_J)$. The jth weight vector must satisfy the admissibility requirement $0 = w_{j,0} < w_{j,1} < w_{j,2} < \cdots < w_{j,C_j}$. If the oncologists assign $w_{j,k} = w_{j,k+1}$, then levels k and $k+1$ of Y_j should be combined. The numerical domain of the $w_{j,k}$'s is arbitrary, because the method is invariant to the weights' multiplicative scale. We used the interval 0 to 10 in the sarcoma trial because the oncologists were comfortable with this range.

We define the *severity weight* of Y_j for a patient given dose x to be the random variable W_j taking the value $w_{j,k}$ with probability $\pi_{j,k}(x, \boldsymbol{\theta})$. This replaces the observed toxicity Y_j with the weight-valued variable W_j, by assigning the severity category probabilities of Y_j to the corresponding elicited weights. We define the patient's *total toxicity burden* (TTB) to be $\text{TTB} = \sum_{j=1}^{J} W_j$. The dose-finding algorithm is based on the posterior expected TTB at each dose,

$$\begin{aligned}\psi(\mathbf{w}, x, data) &= \mathrm{E}\{\mathrm{E}(\text{TTB}|x, \boldsymbol{\theta})|data\} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{C_j} w_{j,k}\mathrm{E}\{\pi_{j,k}(x, \boldsymbol{\theta})|data\}. \qquad (3)\end{aligned}$$

By straightforward algebra, it follows from (1) that

$$\mathrm{E}(\text{TTB}|x, \boldsymbol{\theta}) = \sum_{j=1}^{J}\left\{w_{C_j} + \sum_{k=1}^{C_j}(w_{j,k-1} - w_{j,k})\Phi(\gamma_{j,k} - \beta_{j,0} - \beta_{j,1}x)\right\},$$

which is increasing in x because $w_{j,k-1} < w_{j,k}$ and $\beta_{j,1} > 0$ for all j and k. Thus $\psi(\mathbf{w}, x, data)$ is increasing in x. We use this as a basis for dose-finding by first eliciting a fixed target TTB value, ψ^*, from the oncologists, and then assigning to each successive cohort the dose with $\psi(\mathbf{w}, x, data)$ closest to ψ^*. To see how this works, for simplicity suppress j and consider one toxicity. If $Y = y_k$ is observed at x, then the posterior of $\{\pi_1(x, \boldsymbol{\theta}), \ldots, \pi_C(x, \boldsymbol{\theta})\}$ must shift probability mass toward $\pi_k(x, \boldsymbol{\theta})$. If y_k is a high level toxicity (i.e., k is high in the range $1, \ldots, C$), then because the w_k's are $\uparrow$ in k, $\sum_{k=1}^{C} w_k\pi_k(x, \boldsymbol{\theta})$ must increase stochastically, and hence also in expectation given the data. By the monotonicity of $\psi(\mathbf{w}, x, data)$ in x, this would tend to decrease the next selected dose. Similarly, observation of low-level toxicities will decrease the chosen dose, on average.

3. ELICITATION PROCESS

Although we developed the foregoing model and method for the sarcoma trial over the course of several sessions with the oncologists, future applications should require much less time and effort, because a computer program is freely available from the first author on request. The method requires eliciting a substantial amount of information, however, including the toxicities, the prior on $\boldsymbol{\theta}$, $\mathbf{w}$, and ψ^*. An effort should be made at the start to include all physicians involved in the trial who are opinion leaders, to avoid having to repeat the process due to subsequent disagreements.

Initially, the physicians must specify the cohort size, c, sample size, N, doses, $\mathbf{d} = (d_1, \ldots, d_K)$, the toxicities to be monitored, and their severity levels. Because there will be N/c cohorts to search among K doses, it may be useful to study several feasible values of N in the computer simulations as a basis for choosing N. Next, the physicians must be asked to specify a numerical severity weight for each level of each toxicity within a given positive-valued numerical range, with 0 corresponding to no toxicity. Because the method is invariant to the particular numerical range of severity weights, the main criterion is that the physicians use a range with which they are comfortable. This establishes $\mathbf{w}$. The process of specifying $\mathbf{w}$ may lead the physicians to modify the toxicities, because quantifying the importance of the possible clinical outcomes in this way requires them to think deeply about the process of treating patients and conducting a dose-finding trial.

Given $\mathbf{d}$, $\mathbf{Y}$, and $\mathbf{w}$, the physicians must specify hypothetical J-variate toxicity outcomes for m hypothetical cohorts, with the severities of the toxicities varying widely between cohorts from "very toxic" to "not toxic at all." As a rough guideline, m should be large enough to obtain a reasonable representation of the range of possible toxicities and severities, but not so large that the elicitation process becomes unduly burdensome to the physicians. Toxicities with larger numbers of levels should be included in more cohorts than toxicities with fewer levels. No single type of toxicity should be the only contributor to all of the "very toxic" cohorts. That is, the toxicity burden should be spread across several different types of toxicity, to the extent that this is clinically reasonable. The statisticians may suggest additional cohorts, as we did, but these must make sense clinically to the physicians. We denote the outcomes of the m hypothetical cohorts by $\mathbf{y}_1^* = (\mathbf{y}_{1,1}^*, \ldots, \mathbf{y}_{1,c}^*), \ldots, \mathbf{y}_m^* = (\mathbf{y}_{m,1}^*, \ldots, \mathbf{y}_{m,c}^*)$.

Next, for each hypothetical cohort, $r = 1, \ldots, m$, the physicians must be asked two questions, (1) whether observation of $\mathbf{y}_r^* = (\mathbf{y}_{r,1}^*, \ldots, \mathbf{y}_{r,c}^*)$ in the first cohort of the trial would cause them to repeat the same dose ($D_r = \text{repeat}$), escalate to a higher dose ($D_r = \text{escalate}$), or de-escalate to a lower dose ($D_r = \text{de-escalate}$) for the next cohort; and (2) what dose, $d_r^* = d(\mathbf{y}_r^*)$, would they consider most likely to produce the outcomes $\mathbf{y}_r^*$ of that hypothetical cohort. Letting $\mathbf{w}_{r,l}^*$ denote the severity weight vector corresponding to $\mathbf{y}_{r,l}^*$, for $r = 1, \ldots, m$ and $l = 1, \ldots, c$, the mean TTB of the rth hypothetical cohort is

$$\overline{\text{TTB}}_r^* = \frac{1}{c}\sum_{l=1}^{c}\sum_{j=1}^{J} w_{r,l,j}^*.$$

The ordered mean hypothetical TTB values are denoted by $\overline{\text{TTB}}^*_{(1)} \le \cdots \le \overline{\text{TTB}}^*_{(m)}$; and the corresponding vector of decisions in this order of increasing TTB values, $D_{(1)}, \ldots, D_{(m)}$. We define an *admissible sequence of decisions* ordered in this way to be one comprising a string of escalations, followed by a string of repeats, followed by a string of deescalations. If the m decisions are not admissible, then, working with the physicians, the hypothetical outcomes, elicited decisions, weights, or possibly other portions of the underlying structure are modified as appropriate. Once an admissible set of decisions is obtained, the target TTB is defined to be the mean of the elicited $\overline{\text{TTB}}^*_r$ values for which the physicians' decision was to repeat the same dose,

$$\psi^* = \frac{\sum_{r=1}^{m} \overline{\text{TTB}}^*_r \mathbf{1}(D_r = \text{repeat})}{\sum_{r=1}^{m} \mathbf{1}(D_r = \text{repeat})}.$$

Because ψ^* is determined by the physicians' subjective input, it is analogous to a fixed target toxicity probability specified by the physicians in the simpler case of one binary toxicity.

The doses $d_1^*, \ldots, d_m^*$ obtained as answers to the second question may be used to construct a prior on $\boldsymbol{\beta}$, as follows. Recall that because we assume vague priors on $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$, only the hyperparameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the prior on $\boldsymbol{\beta}$ must be specified. Beginning with a vague $\text{N}(\mu^o, \Sigma^o)$ prior on $\boldsymbol{\beta}$, with $\mu^o_{j,0} = 0$, $\mu^o_{j,1} = 1$, $\text{var}^o(\beta_{j,0}) = \text{var}^o(\beta_{j,1}) = 10{,}000$ for each j, we computed the posterior of $\boldsymbol{\beta}$ given the hypothetical data $\{d_1^*, \mathbf{y}_1^*, \ldots, d_m^*, \mathbf{y}_m^*\}$. We modified this distribution by multiplying each variance by m, the number of hypothetical cohorts, and setting each off-diagonal element of $\boldsymbol{\Sigma}$ equal to 0. This gave the prior on $\boldsymbol{\beta}$ used at the start of the trial. As a check for internal consistency, it may be assumed in turn that each hypothetical cohort is the first cohort in the trial. If the algorithm makes the same decision as the physicians, then construction may proceed; otherwise, the prior or possibly some other aspect of the model or design must be modified, as appropriate, so that the method behaves in accordance with clinical practice.

4. THE SARCOMA TRIAL REVISITED

Although the oncologists initially decided on a toxicity severity weight domain ranging from 0 (no clinical importance) to 10 (the most severe toxicity of that type seen), as shown in Table 1, the highest weight that they assigned was 6. Table 2 summarizes the 16 hypothetical cohorts used in the elicitation process for the sarcoma trial, with no toxicities of any type denoted by NT. Each toxicity is subscripted by its grade. For example, the first patient in hypothetical cohort 3, with outcomes denoted by $M_3^- + D_3 + N_3$, experienced grade 3 myelosuppression without fever, grade 3 dermatitis, and grade 3 nausea/vomiting. The description of each cohort is followed by the corresponding empirical mean TTB, the elicited decision for the next cohort, and the dose that the oncologists considered most likely to have caused the cohort's outcomes. Using the elicitation method described in Section 3, the three cohorts for which the decision was to repeat the same dose were 1, 9 and 16. This gives a target per patient TTB of $\psi^* = (3.00 + 3.12 + 3.00)/3 = 3.04$, which is the value used to conduct the sarcoma trial.

At this writing, 14 patients have been treated and evaluated. Table 3 summarizes the outcomes and TTBs for these patients. To more fully illustrate how the method works in practice, these data are followed in Table 3 by hypothetical data for the remaining 22 patients in the trial. The third cohort included only three patients, because the toxicities of all of these three patients were evaluated before a fourth patient was available to be accrued to this cohort. The first cohort was treated at 400 mg/m^2. Although $\psi(\mathbf{w}, 700, data_4) = 3.24$ is closest to the target $\psi^* = 3.04$, because of the safety constraint that no untried dose may be skipped when escalating, the second cohort was treated at 500 mg/m^2. Incorporating the second cohort's data, because $\psi(\mathbf{w}, 600, data_8) = 2.85$ is closest to 3.04, the third cohort was treated at 600 mg/m^2. The next value, $\psi(\mathbf{w}, 700, data_{11}) = 3.24$, determined that the fourth cohort should receive 700 mg/m^2. The tabled data from the fourth cohort consist of actual outcomes plus one hypothetical patient, number 15. Because $\psi(\mathbf{w}, 600, data_{15}) = 3.15$, the trial de-escalates to 600 mg/m^2. The remaining decisions reported in the table rely on hypothetical outcomes for patients 16–36. Based on these data, the trial would stay at 600 mg/m^2 for cohort 6 because $\psi(\mathbf{w}, 600, data_{19}) = 3.01$, and then return to 700 mg/m^2 as the dose for the final 17 patients, with $\psi(\mathbf{w}, 700, data_{23}) = 3.31$, $\psi(\mathbf{w}, 700, data_{27}) =$

Table 2. Hypothetical Cohorts Used in the Elicitation Process for the Sarcoma Trial

Cohort	Outcomes	$\overline{TTB}^*$	Decision	d^*
1	$M_4^+, \text{D}_4, \text{NT}, \text{NT}$	3.00	Repeat	400
2	$M_4^-, \text{L}_3, \text{F}_4, \text{N}_4$	1.88	Escalate	200
3	$M_3^- + D_3 + N_3, \text{M}_3^-, \text{M}_3^-, \text{M}_3^-$	2.00	Escalate	300
4	$D_4, \text{D}_4, \text{L}_2, \text{L}_2$	4.00	De-escalate	600
5	$M_3^- + F_3, \text{M}_3^- + F_3, \text{L}_2 + F_3, \text{N}_3$	2.25	Escalate	300
6	$M_4^+, \text{L}_4, \text{D}_4, \text{NT}$	4.50	De-escalate	700
7	$D_3, \text{D}_3, \text{NT}, \text{NT}$	1.25	Escalate	100
8	$M_3^-, \text{D}_3, \text{F}_3, \text{F}_3$	1.25	Escalate	200
9	$D_3, \text{D}_4, \text{L}_2, \text{L}_2$	3.12	Repeat	400
10	$M_3^+ + N_3, \text{M}_3^+ + D_3 + N_3, \text{D}_3 + F_3 + N_3, \text{F}_3 + N_3$	5.50	De-escalate	800
11	$M_3^- + D_3 + F_3, \text{F}_3, \text{F}_2, \text{N}_4$	2.12	Escalate	300
12	$L_3, \text{L}_3, \text{NT}, \text{NT}$	1.50	Escalate	200
13	$M_3^+ + D_3 + F_4, \text{M}_3^+ + D_3 + F_4, \text{M}_3^- + F_4, \text{D}_3 + F_4$	5.62	De-escalate	900
14	$M_3^- + N_3, \text{F}_4, \text{L}_2, \text{D}_3 + N_3$	2.38	Escalate	300
15	$D_4 + F_4, \text{L}_3 + F_4, \text{L}_2 + N_4, \text{N}_4$	4.25	De-escalate	600
16	$M_3^- + F_4, \text{L}_2 + F_3, \text{M}_3^- + D_3 + N_4, \text{L}_2$	3.00	Repeat	500

NOTE: Myelosuppression, dermatitis, liver toxicity, fatigue, and nausea/vomiting are denoted by *M*, *D*, *L*, *F*, and *N*, with each subscripted by grade. With myelosuppression, presence and absence of fever are denoted by superscripts "+" and "−". NT denotes no toxicities of any type, and d^* is the dose considered by the oncologists most likely to have caused the cohort's outcomes.

Table 3. Patient-by-Patient Illustration of the Method Used in the Sarcoma Trial

Patient	*Dose (mg/m^2)*	*Myelosuppression*	*Dermatitis*	*Liver*	*Fatigue*	*Nausea*	*TTB*
1	400	Grade 3 without fever	Grade 3	None	None	Grade 3	5.0
2	400	Grade 3 without fever	None	None	None	None	1.0
3	400	Grade 3 without fever	None	None	None	None	1.0
4	400	None	None	None	None	None	0
5	500	None	Grade 3	None	None	None	2.5
6	500	None	None	None	None	None	0
7	500	Grade 3 without fever	Grade 3	None	None	None	3.5
8	500	Grade 4 without fever	None	Grade 2	None	None	3.5
9	600	None	Grade 3	None	None	None	2.5
10	600	None	None	Grade 3	None	None	2.0
11	600	None	Grade 3	None	None	None	2.5
12	700	Grade 3 without fever	None	None	None	None	1.0
13	700	None	None	Grade 3	None	None	3.0
14	700	None	None	None	Grade 3	None	.5
15	700	Grade 3 with fever	Grade 4	None	Grade 3	Grade 3	13.0
16	600	Grade 3 without fever	None	None	None	Grade 3	2.5
17	600	Grade 3 without fever	None	None	None	None	1.0
18	600	None	Grade 3	Grade 2	None	None	4.5
19	600	None	Grade 3	None	None	None	2.5
20	600	None	None	None	None	None	0
21	600	None	Grade 3	None	None	None	2.5
22	600	Grade 4 without fever	None	None	None	None	1.5
23	600	Grade 4 without fever	None	None	None	None	1.5
24	700	Grade 3 without fever	None	Grade 2	None	None	3.0
25	700	Grade 3 without fever	None	None	None	None	1.0
26	700	Grade 3 without fever	None	None	None	Grade 4	3.0
27	700	Grade 3 without fever	None	None	None	None	1.0
28	700	Grade 3 with fever	None	None	None	None	5.0
29	700	Grade 3 without fever	Grade 3	None	None	Grade 3	5.0
30	700	None	None	None	Grade 3	None	.5
31	700	None	None	None	Grade 3	None	.5
32	700	Grade 4 without fever	None	Grade 2	None	Grade 3	5.0
33	700	Grade 3 without fever	Grade 3	None	None	None	3.5
34	700	None	None	None	None	None	0
35	700	None	None	Grade 3	None	None	3.0
36	700	None	None	Grade 3	None	None	3.0

NOTE: The first 11 patients are real; the remaining 25 are hypothetical.

3.09, $\psi(\mathbf{w}, 700, data_{31}) = 3.03$, $\psi(\mathbf{w}, 700, data_{35}) = 3.02$, and, finally, $\psi(\mathbf{w}, 700, data_{36}) = 3.03$ determining 700 mg/m^2 to be the MTD.

For the sarcoma trial, conventional methods typically would define one binary "toxicity" as the maximum of the indicators $\mathbf{1}$(myelosuppression grade $\geq$ 3), $\mathbf{1}$(dermatitis grade $\geq$ 3), $\mathbf{1}$(liver toxicity grade $\geq$ 3), $\mathbf{1}$(nausea/vomiting grade $\geq$ 3), and $\mathbf{1}$(fatigue grade $\geq$ 3). For example, a conventional method would consider a patient with grade 2 liver toxicity (TTB = 2.5) to have "no toxicity" and a patient with grade 4 fatigue (TTB = 1) to have "toxicity," and furthermore would not distinguish the latter patient from a patient with grade 4 myelosuppression with fever, grade 4 dermatitis, and grade 4 liver toxicity (TTB = 18). Consequently, the proposed algorithm based on the TTB with target 3.04 for $\psi(\mathbf{w}, x, data)$ makes more sensible decisions. For example, if three of the four pa-

tients in the first cohort treated at 400 mg/m^2 had either grade 4 fatigue or grade 3 myelosuppression without fever and one also had grade 3 nausea, for TTB values {0, 1, 1, 2.5} and empirical mean TTB = 1.125, then for these data $\psi(\mathbf{w}, 500, data_4) = 2.10$, and the algorithm would escalate to 500 mg/m^2, whereas a conventional method would score three toxicities in these four patients and de-escalate to a lower dose.

5. SIMULATION STUDY AND SENSITIVITY ANALYSES

5.1 Simulation Study Design

To assess the method's average behavior, we performed a simulation study of the sarcoma trial. Due to the inherent complexity of patient outcome, specifying a reasonably representative set of possible dose-toxicity probabilities to study is not straightforward. To obtain a manageable set of dose-toxicity scenarios for the simulation study, we considered only cases where the target TTB occurred at 200, 500, or 800 mg/m^2, and we categorized the main source of toxicity as being either those having high-severity (HS) weights ($w \geq 5$) or greater, or low-severity (LS) weights ($w \leq 2$). We considered the remaining toxicities, grade 3 dermatitis ($w = 2.5$) and grade 3 liver toxicity ($w = 3$), intermediate and included them in either group. Each of the six scenarios was characterized by $10C_+ = 130$ fixed probabilities $p_{j,1,d}, \ldots, p_{j,C_j,d}$, for $j = 1, \ldots, 5$ and $d = 100, \ldots, 1000$, where $p_{j,k,d} = \Pr(Y_j = y_{j,k}|d)$ and $p_{j,0,d} = 1 - \sum_{k=1}^{C_j} p_{j,k,d}$. We chose these probabilities nonparametrically to satisfy $\sum_j \sum_k w_{j,k} p_{j,k,d} = 3.04$ for $d = 200$ under scenarios 1 and 2, $d = 500$ under scenarios 3 and 4, and $d = 800$ under scenarios 5 and 6. Figure 1 summarizes the six scenarios graphically in terms of the TTB as a function of dose. We induced association among the elements of each simulated $(Y_1, \ldots, Y_5)$ by generating a sample of standard normal random variables $\mathbf{Z}^{5\times 1}$ with specified correlation matrix, then defining the correlated uniform(0, 1) random variates $\mathbf{U}^{5\times 1} = (\Phi(Z_1), \ldots, \Phi(Z_5))$, and finally denoting $P_{j,k,d} = \sum_{r=0}^{k} p_{j,r,d}$ for $k = 0, \ldots, C_j$ and $P_{j,-1,d} = 0$, defining $Y_j = y_{j,k}$ if $P_{j,k-1,d} \leq U_j < P_{j,k,d}$. We elicited the correlations from the oncologists in terms of the latent variables, $\mathbf{Z}$, underlying the toxicities, as follows. The only toxicities that the oncologists considered correlated a priori were F and N. We asked the oncologists the following question: "For two randomly chosen patients, if you observe that the first patient is more fatigued than the second, what is the probability of observing more severe nausea/vomiting in the first patient than the second?" The oncologists assigned a probability between .55 and .60 to this event. Denoting the latent variables corresponding to F and N for the two patients by $(Z_{i,F}, Z_{i,N})$ for $i = 1, 2$ and $q = \Pr(Z_{1,F} > Z_{2,F}|Z_{1,N} > Z_{2,N})$, if we assume that q is symmetric in that $q = \Pr(Z_{1,F} < Z_{2,F}|Z_{1,N} < Z_{2,N})$, then q is related to Kendall's τ via $\tau = 2q - 1$. Because the Pearson's correlation ρ between Z_F and Z_N between τ and satisfies the relationship $\rho = \sin(\tau\pi/2)$ (Kruskal 1958), $.55 < q < .60$ implies that $.15 < \rho < .31$. We used the average $\rho = .23$. We simulated the trial 1,000 times under each scenario; each reported value is the average over these replications.

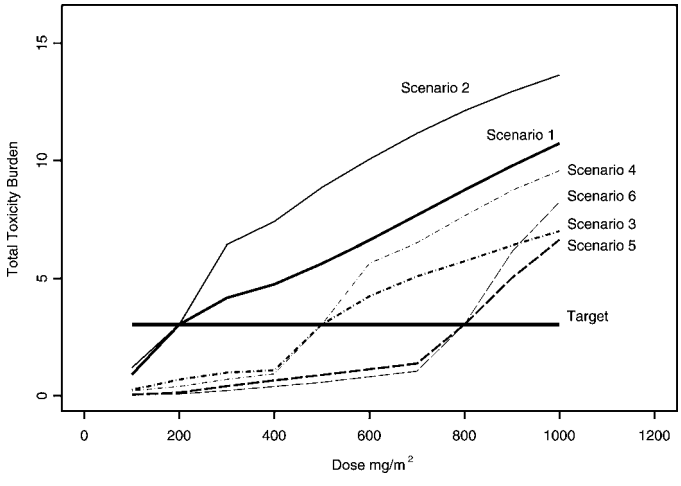

Figure 1. TTB as a Function of Dose Under Each of the Six Dose-Toxicity Scenarios Considered in the Simulation Study.

5.2 Numerical Methods

We followed the computational framework developed by Albert and Chib (1993) for one polytomous outcome, extended by Chib and Greenberg (1998) to accommodate correlated binary outcomes and by Chen and Dey (2000) to the correlated ordinal case. Denote the outcome indices of the ith patient by $\mathbf{k}_i = (k_{i,1}, \ldots, k_{i,J})$ for $i = 1, \ldots, n$, and write $\mathbf{Z}^{(n)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ and $\mathbf{Y}^{(n)} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Because $\Pr\{\mathbf{Y}_i = \mathbf{y}(\mathbf{k}_i)|\mathbf{Z}_i, \boldsymbol{\theta}\} = \mathbf{1}\{\mathbf{Z}_i \in \mathbf{A}(\mathbf{k}_i, \boldsymbol{\gamma})\}$, by Bayes's theorem the joint posterior of the latent variables and parameters is given by

$$f\left(\mathbf{Z}^{(n)}, \boldsymbol{\theta}\,\middle|\,\mathbf{Y}^{(n)}\right) = \prod_{i=1}^{n} f\left(\mathbf{Z}_i, \boldsymbol{\theta}|\mathbf{Y}_i = \mathbf{y}(\mathbf{k}_i),\, x_i\right)$$

$$\propto f(\boldsymbol{\theta}) \prod_{i=1}^{n} \mathbf{1}\{\mathbf{Z}_i \in \mathbf{A}(\mathbf{k}_i, \boldsymbol{\gamma})\} f(\mathbf{Z}_i|\boldsymbol{\theta}, x_i), \quad (4)$$

where $f(\boldsymbol{\theta})$ denotes the prior of $\boldsymbol{\theta}$. Under this representation, the latent variables are used to ease the computational burden of computing the posterior $f(\boldsymbol{\theta}|\mathbf{Y})$. By using a Markov chain Monte Carlo (MCMC) algorithm, values of $(\mathbf{Z}^{(n)}, \boldsymbol{\theta})$ generated from (4) will yield the desired posterior. The following algorithm is similar to those given by Chen and Dey (2000) and Cowles (1996). Let $\mathbf{Z}_j^{(n)} = (Z_{j,1}, \ldots, Z_{j,n})$ be the independent latent variables associated with Y_j, and let $\mathbf{Z}_{-j}^{(n)}$ be the subvector of $\mathbf{Z}^{(n)}$ obtained by deleting $Z_{i,j}$ from $\mathbf{Z}_i$ for each $i = 1, \ldots, n$.

The MCMC algorithm proceeds as follows:

1. For each $j = 1, \ldots, J$, start with $f(\boldsymbol{\gamma}_j, \mathbf{Z}_j^{(n)}|\mathbf{Z}_{-j}^{(n)}, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho})$, integrate out $\mathbf{Z}_j^{(n)}$ to obtain $f(\boldsymbol{\gamma}_j|\mathbf{Z}_{-j}^{(n)}, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho})$, generate $\boldsymbol{\gamma}_j$ from this distribution, and generate $\mathbf{Z}_j^{(n)}$ from $f(\mathbf{Z}_j^{(n)}|\mathbf{Z}_{-j}^{(n)}, \boldsymbol{\gamma}_j, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho})$.
2. Generate $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\mathbf{Z}^{(n)}, \mathbf{Y}^{(n)}, \boldsymbol{\gamma}, \boldsymbol{\rho})$.
3. Generate $\boldsymbol{\rho}$ from $f(\boldsymbol{\rho}|\mathbf{Z}^{(n)}, \mathbf{Y}^{(n)}, \boldsymbol{\gamma}, \boldsymbol{\beta})$.

Step 1 uses the fact that $f(\boldsymbol{\gamma}_j|\mathbf{Z}_{-j}^{(n)}, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho}) = f(\boldsymbol{\gamma}_j|\mathbf{Z}_{-j}^{(n)}, \boldsymbol{\gamma}_{-j}, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho})$ due to the conditional independence of $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J$. It alternates between this distribution and $f(\mathbf{Z}_j^{(n)}|\mathbf{Z}_{-j}^{(n)}, \boldsymbol{\gamma}_j, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho})$ because these are much more tractable than $f(\boldsymbol{\gamma}_j|\mathbf{Z}^{(n)}, \mathbf{Y}^{(n)}, \boldsymbol{\beta}, \boldsymbol{\rho})$. Because steps 1–3 generate $f(\boldsymbol{\theta}, \mathbf{Z}^{(n)}|\mathbf{Y}^{(n)})$, the posterior $f(\boldsymbol{\theta}|\mathbf{Y}^{(n)})$ is obtained as a natural consequence.

For the simulation study, we evaluated the MCMC algorithm's performance using standard convergence diagnostics. A burn-in of 1,000 and a chain of length 30,000, retaining every 15th sample, provided adequate convergence. Although the posterior sample size is constrained by computing resources due to the need for many replications in the simulation study, in the actual trial conduct we base all inferences on a much larger MCMC posterior sample size.

5.3 Simulation Results

Table 4 summarizes the simulation results. Under scenarios 1 and 2, because the target $\psi^* = 3.04$ is achieved at 200 mg/m^2, the starting dose of 400 mg/m^2 is unacceptably toxic. Under scenario 1, most of the toxicity is due to LS events, such as L_2, F, or N. In contrast, most of the toxicity burden under scenario 2 is due to HS toxicities, such as M^+ or L_4. Under either of these two scenarios, the method chooses the best dose, 200 mg/m^2, more than 90% of the time, and on average treats 22 of 36 patients at this dose. The algorithm thus appears to perform well regardless of whether most of the toxicity burden arises from LS or HS toxicities. For scenarios 3 and 4, the target TTB is achieved at 500 mg/m^2, with most of the TTB due to LS toxicities under scenario 3 and to HS toxicities under scenario 4. Again, the method is insensitive to the source of the toxicities, with an 85–87% correct selection rate and most of the 36 patients treated at or near the selected MTD. For scenarios 5 and 6, where $\psi^* = 3.04$ at 800 mg/m^2, the correct selection rate is about 80%, slightly lower than the other cases. This is due primarily to the "do-not-skip" rule, which requires that at least one cohort be treated at each dose level when escalating, so that at least 16 of the 36 patients must be treated at doses below 800 mg/m^2 and few patients are available for evaluation at the higher dose levels.

5.4 Sensitivity Analyses

A fundamental issue is the method's robustness to the severity weights, because other physicians might specify substantively different weights. To address this, we performed a sensitivity analysis by randomly perturbing the elicited weight vector, $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_J)$, to obtain hypothetical weights as follows. For each toxicity type $j = 1, \ldots, J$, temporarily suppressing j for simplicity, we first replaced the maximum weight, w_C, by $w_C^{(h)} = U\, w_C$ for $U \sim$ uniform[.5,1.5]. Thus $w_C^{(h)}$ is obtained by randomly decreasing or increasing w_C by up to half its value. Next, we randomly perturbed the lower severity weights $w_1 < w_2 < \cdots < w_{C-1}$ while maintaining their ordering by first generating a C-category Dirichlet random vector $\mathbf{q} = (q_1, \ldots, q_{C-1})$ with parameter vector $\mathbf{a} = (a_1, \ldots, a_C)$ determined by dispersion parameter $a_+ = a_1 + \cdots + a_c = 3$ and mean vector $(a_1, \ldots, a_C)/a_+ = (w_1, w_2 - w_1, \ldots, w_C - w_{C-1})/w_C$. The $C - 1$ hypothetical lower severity weights were then given by $(w_1^{(h)}, \ldots, w_{C-1}^{(h)}) = (q_1, q_1 + q_2, \ldots, \sum_{r=1}^{C-1} q_r) w_C^{(h)}$. Repeating this for each j yielded $\mathbf{w}^{(h)} = (\mathbf{w}_1^{(h)}, \ldots, \mathbf{w}_J^{(h)})$. The hypothetical $\mathbf{w}^{(h)}$ determines a TTB target, $\psi^{(*,h)}$, and given fixed outcome probabilites $\{p_{j,k,d}\}$ as defined in Section 5.1, $\mathbf{w}^{(h)}$ determines the best dose, $d^{(*,h)}$, among the 10 doses $\{100, \ldots, 1{,}000\}$, having a TTB closest to $\psi^{(*,h)}$. Independently generating 10,000 such $\mathbf{w}^{(h)}$ vectors, the distribution of the corresponding $\psi^{(*,h)}$ values has (2.5, 5.0, 50, 95, 97.5)th percentiles (1.76, 1.92, 2.97, 4.37, 4.67). Examining the distribution of the corresponding $d^{(*,h)}$ values chosen under scenario 4, for example, where $d = 500$ has TTB closest to the elicited $\psi^* = 3.04$, we found that $d^{(*,h)} = 400$, 500, and 600 mg/m^2 with probabilities .051, .875, and .071. Thus, despite the fact that $\mathbf{w}^{(h)}$ is obtained as a rather severe perturbation of $\mathbf{w}$, under scenario 4 the targeted dose under $\mathbf{w}^{(h)}$ is nearly certain to be within one dose level of

Table 4. Simulation Results for the Sarcoma Trial Under the Six Dose-Toxicity Scenarios

Scenario	Main toxicities		Gemcitabine dose (mg/m^2) 100	200	300	400	500	600	700	800	900	1,000
ψ^* = 3.04 at 200 mg/m^2												
1	LS	Psel	1.6	[93.7]	4.7	0	0	0	0	0	0	0
		Npats	2.2	21.8	5.3	5.6	1.1	0	0	0	0	0
2	HS	Psel	4.1	[92.0]	3.8	0	0	0	0	0	0	0
		Npats	4.5	22.0	4.5	4.6	.4	0	0	0	0	0
ψ^* = 3.04 at 500 mg/m^2												
3	LS	Psel	0	0	0	5.6	[85.4]	9.0	0	0	0	0
		Npats	0	0	0	5.6	18.5	9.7	2.1	.1	0	0
4	HS	Psel	0	.1	0	5.2	[86.9]	7.8	0	0	0	0
		Npats	0	0	.1	6.6	20.6	7.8	1.0	0	0	0
ψ^* = 3.04 at 800 mg/m^2												
5	LS	Psel	0	0	0	0	0	.3	10.4	[80.7]	8.3	.2
		Npats	0	0	0	4.0	4.0	4.0	5.6	11.4	5.9	1.0
6	HS	Psel	0	0	0	0	0	.1	13.3	[80.3]	6.2	0
		Npats	0	0	0	4.0	4.0	4.0	5.7	11.7	5.8	.7

NOTE: Psel, % selected; Npats, number of patients treated.

the dose (500 mg/m^2) targeted by the elicited **w**, and 87.5% of the time the targeted doses are the same. We obtained similar results under the other scenarios. Thus the targeted dose appears to be robust to the elicited weights.

To examine the sensitivity of the dose-finding method itself to **w**, we simulated the trial under scenario 4 using each of 16 hypothetical weight vectors, $\mathbf{w}^{(h,1)}, \ldots, \mathbf{w}^{(h,16)}$, given in Table 5. We chose these weight vectors to reflect the distribution of $d^{(*,h)}$ noted earlier, because $1/16 = .063$, so that the targeted doses were $d^{(*,h,1)} = 400$, $d^{(*,h,2)} = \cdots = d^{(*,h,15)} = 500$, and $d^{(*,h,16)} = 600$. We chose $\mathbf{w}^{(h,1)}$ and $\mathbf{w}^{(h,16)}$ so that $\psi^{(*,h,1)}$ and $\psi^{(*,h,16)}$ were the medians of the hypothetical **w** with $d^{(*,h)} = 400$ and $d^{(*,h)} = 600$. We chose $\mathbf{w}^{(h,2)}, \ldots, \mathbf{w}^{(h,15)}$ so that $\psi^{(*,h,2)}, \ldots, \psi^{(*,h,15)}$ were equally spaced percentiles, between the 10th and 90th, of the distribution of $\psi^{(*,h)}$ values for which $d^{(*,h)} = 500$. For the rth hypothetical weight vector, $\mathbf{w}^{(h,r)}$, we denoted the percent absolute deviation of the selected dose, d_{sel}, from $d^{(*,h,r)}$ by $\text{dev}_r = 100|d_{sel} - d^{(*,h,r)}|/d^{(*,h,r)}$. In all 16 cases, d_{sel} was within one level of $d^{(*,h,r)}$ more than 99.9% of the time. For $d^{(*,h,1)} = 400$, doses (300, 400, 500) were selected (1.7, 60.6, 37.7)% of the time, and on average, dev_1 was 10.4%. In 13 of the 14 cases where $d^{(*,h,r)} = 500$, this target dose was selected between 80.6% and 86.5% of the time, and in one case 500 was selected 73.2% of the time. The mean values of $\text{dev}_2, \ldots, \text{dev}_{15}$ varied from 2.9% to 5.5%. In the 16th case, (500, 600, 700) were selected (43, 54, 3)% of the time, and on average dev_{16} was 12.1%. Thus the method appears to be robust to the elicited weights in terms of changes in the targeted dose, correct selection percentage, and deviation of the selected dose from the targeted dose.

Using the elicited severity weights, we also examined the method's sensitivity to cohort size, sample size, starting dose, and ψ^*. We conducted simulations examining the effects of cohort size and sample size under scenario 4. For cohort sizes (1, 2, 3, 4, 5), with starting dose 400 and sample size 36, the respective correct selection percentages were (89, 86, 89, 87, 86). Because the range of these values is well within what would be expected from simulation variation, the method appears to be insensitive to cohort size. For sample sizes (28, 32, 36, 40, 44), with the cohort size fixed at 4 and a starting dose of 400 mg/m^2, the correct selection percentages were (82, 84, 87, 88, 89). Thus the method's reliability improves with larger sample size. We examined the effect of changing the starting dose from 400 mg/m^2 to 100 mg/m^2 under scenarios 2, 4, and 6, where the target TTB is achieved at 200 mg/m^2, 500 mg/m^2, and 800 mg/m^2. In these cases the correct selection percentages were 92% when the target was 200 mg/m^2, 84% when the target was 500 mg/m^2 and 42% when the target was 800 mg/m^2. The comparatively low value in the last case was as expected, because the "do-not-skip" rule with starting dose 100 mg/m^2 requires that 28 of the 36 patients be treated at doses below 800 mg/m^2, which leaves at most two cohorts to treat at the correct dose. If this rule is dropped, then the correct selection percentage in this case is 80%. To examine the effect of higher correlation among the toxicities on the correct selection rate, we changed the correlations so that there was high correlation between fatigue and nausea/vomiting (.60), low correlation between fatigue and dermatitis (.20), and moderate correlation between fatigue and myelosuppression (.40). Under scenarios 2, 4, and 6, with this correlation structure the correct selection percentages were 92% when the target TTB was achieved at 200 mg/m^2, 86% at 500 mg/m^2, and 79% at 800 mg/m^2. Because these are nearly identical to the values given in Table 4 obtained with the original correlation structure, it appears that this degree of association among the toxicities does not alter the method's behavior, on average.

We assessed the method's sensitivity to ψ^* by simulating the trial with target $\psi^*(\Delta) = \psi^* \pm \Delta$, for $\Delta = \pm.25$ and $\pm.50$—that is, $\psi^*(\Delta) = 2.54, 2.79, 3.29$, and 3.54—under each of scenarios 2, 4 and 6. Defining the "best" dose d among the 10 levels studied as that at which $\sum_j \sum_k w_{j,k} p_{j,k,d}$ is closest $\psi^*(\Delta)$, in each of these cases the best dose was identical to that for which $\Delta = 0$. The percentages of selecting the best dose for $\Delta = (-.50, -.25, 0, +.25, +.50)$ were (68, 76, 92, 75, 65) under scenario 2, (77, 82, 87, 85, 80) under scenario 4, and

Table 5. The 16 Hypothetical Weight Vectors Used to Study the Method's Sensitivity to the Elicited Weights

Toxicity elicited **w**	M_3^- 100	M_4^- 150	M_3^+ 500	M_4^+ 600	D_3 250	D_4 600	L_2 200	L_3 300	L_4 600	N_3 150	N_4 200	F_3 50	F_4 100	ψ^* 304
$\mathbf{w}^{(h,1)}$	10	11	416	497	140	524	52	535	869	187	234	12	71	198
$\mathbf{w}^{(h,2)}$	20	21	96	563	130	545	81	511	524	93	160	40	138	218
$\mathbf{w}^{(h,3)}$	156	158	469	470	87	541	122	395	420	123	176	13	88	234
$\mathbf{w}^{(h,4)}$	20	21	470	491	325	715	11	12	508	207	220	35	52	247
$\mathbf{w}^{(h,5)}$	1	4	645	706	29	387	291	295	304	153	197	70	134	259
$\mathbf{w}^{(h,6)}$	44	151	732	733	168	803	2	41	321	208	209	118	140	270
$\mathbf{w}^{(h,7)}$	33	191	595	614	171	736	143	146	506	91	100	77	116	280
$\mathbf{w}^{(h,8)}$	74	127	845	846	365	407	137	140	373	160	229	86	103	292
$\mathbf{w}^{(h,9)}$	227	228	458	518	94	854	100	367	769	101	190	49	119	302
$\mathbf{w}^{(h,10)}$	283	285	351	367	139	406	377	421	669	129	144	33	62	314
$\mathbf{w}^{(h,11)}$	351	531	771	822	124	425	259	349	620	36	162	17	90	327
$\mathbf{w}^{(h,12)}$	11	12	346	356	486	583	317	352	830	95	217	15	77	341
$\mathbf{w}^{(h,13)}$	401	412	610	619	544	627	41	243	564	80	113	91	145	356
$\mathbf{w}^{(h,14)}$	149	235	581	874	335	846	145	165	838	37	171	72	137	375
$\mathbf{w}^{(h,15)}$	580	707	777	861	360	526	153	162	572	208	217	60	114	400
$\mathbf{w}^{(h,16)}$	129	146	302	413	330	632	370	386	433	168	185	33	84	365

NOTE: These vectors were chosen to reflect the distribution of $d^{(*,h)}$ under scenario 4, so that $d^{(*,h,1)} = 400$, $d^{(*,h,2)} = \cdots = d^{(*,h,15)} = 500$, and $d^{(*,h,16)} = 600$. Each weight and ψ^* has been multiplied by 100.

(55, 67, 80, 79, 75) under scenario 6. But the method selected a dose within one level, (i.e., within ± 100 mg/m^2, of the best dose more than 99.5% of the time in all of these cases. These results are similar to those obtained in the analogous but simpler case where dose-finding is based on a single binary toxicity using the crm. Assuming that Pr(toxicity$|j$th level) $= p_j^{\exp(\alpha)}$, where $\alpha \sim \mathrm{N}(0, 2)$ for fixed probabilities $(p_1, p_2, \ldots, p_{10}) = (.05, .08, .15, .30, .45, .55, .63, .70, .75, .80)$, we simulated each of the nine trials obtained by the targets $p^* = .20, .30$, or .40 being achieved at 200, 500, or 800 mg/m^2. Defining the best dose as that minimizing $|p_j - p^*|$, over these nine cases, the percentages for selecting the best dose ranged from 44% to 65%, but the crm selected a dose within one level of the best with percentages 87% to 99%.

To assess between-physician variability empirically, repeating the elicitation process with another group of oncologists would be useful but logistically difficult. However, a fourth soft tissue sarcoma oncologist agreed to provide his TTB. We presented him with the hypothetical scenarios developed by the other oncologists and asked which action he would take in each case. Thus, his responses and resulting target TTB were based on the other oncologists' hypothetical scenarios. His chosen actions agreed with those of the other three oncologists for 13 of the 16 cohorts, with the differences being to repeat rather than de-escalate for cohort 4, repeat rather than escalate for cohort 7, and escalate rather than repeat for cohort 16. Because he would repeat the dose for cohorts 1, 4, 7, and 9, his resulting target TTB was $(3.00 + 4.00 + 1.25 + 3.12)/4 = 2.84$, a value only .20 (8.6%) below the value of 3.04 used to conduct the trial.

6. CONCLUDING REMARKS

We have presented a new dose-finding method that accommodates several different toxicities with severity levels of varying clinical importance. Our application to the soft tissue sarcoma trial illustrates the method's practicality, and our simulation study shows that on average the algorithm performs well under a wide variety of circumstances. The method requires substantially more effort to implement than conventional dose-finding methods, including close interaction with the physicians to establish toxicities, severity weights, and target TTBs, as well as the simulation study. We feel that this effort is well warranted by our method's advantages over conventional methods that reduce several toxicities to one binary variable.

Such a labor-intensive design process may not appeal to some clinicians. The method is much easier to implement in the case of one ordinal toxicity, however, and this may serve as a bridge to more complex settings as the process of physician–statistician collaboration evolves. This special case still provides a substantial advantage over methods based on one binary toxicity. A single ordinal Y takes on one of $C + 1$ severity values $y_0 < y_1 < \cdots < y_C$, there is one latent variable $Z \sim \mathrm{N}(\beta_0 + \beta_1 x, 1)$ with $(Y = y_k) = (\gamma_k \le Z < \gamma_{k+1})$, and $\pi_k(x, \boldsymbol{\theta}) = \Pr(Y = y_k | x, \boldsymbol{\theta}) = \Phi\{\gamma_k - (\beta_0 + \beta_1 x)\} - \Phi\{\gamma_{k+1} - (\beta_0 + \beta_1 x)\}$ for $k = 0, \ldots, C$, where $0 = \gamma_1 < \gamma_2 < \cdots < \gamma_C$. Only one vector of increasing severity weights, $(w_1, \ldots, w_k)$, is elicited, the TTB $= W$, where W is univariate with $\Pr(W = w_k) = \pi_k(x, \boldsymbol{\theta})$ and ψ^* is the elicited target for $\mathrm{E}(W|data) = \sum_{k=1}^{C} w_k \, E\{\pi_k(x, \boldsymbol{\theta})|data\}$. For example, suppose that a single ordinal toxicity is defined in terms of grades 0, 1, 2, 3, and 4 but has elicited severity weights 0, 1, 2, 3, and 6. If $\boldsymbol{\pi}^{(a)}(x) = (.50, .10, .10, .20, .10)$ and $\boldsymbol{\pi}^{(b)}(x) = (.10, .10, .50, .10, .20)$ for a given dose x, then both of these probability vectors yield the same conventionally used criterion $\Pr(Y \ge 3|x) = .30$ for dose-limiting (grade 3 or 4) toxicity, whereas $\boldsymbol{\pi}^{(a)}(x)$ has $\mathrm{E}^{(a)}(\mathrm{TTB}) = 1.5$ whereas $\boldsymbol{\pi}^{(b)}(x)$ has $\mathrm{E}^{(b)}(\mathrm{TTB}) = 2.6$. This illustrates the fact that even with only one toxicity, accounting for multiple severity levels and eliciting severity weights provides a more informative evaluation of toxicity.

[Received October 2002. Revised December 2003.]

REFERENCES

Albert, J. H., and Chib, S., (1993), "Bayesian Analysis of Binary and Polytomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Babb, J., Rogatko, A., and Zacks, S. (1998), "Cancer Phase I Clinical Trials: Efficient Dose Escalation With Overdose Control," *Statistics in Medicine*, 17, 1103–1120.

Chen, M. H., and Dey, D. K. (2000), "Bayesian Analysis for Correlated Ordinal Data Models," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh, and B. K. Mallick, New York: Marcel Dekker, pp. 133–157.

Chib, S., and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361.

Cowles, M. K. (1996), "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models," *Statistics and Computing*, 6, 101–111.

Durham, S. D., and Fluornoy, N. (1994), "Random Walks for Quantile Estimation," in *Statistical Decision Theory and Related Topics*, eds. S. Gupta and J. O. Berger, New York: Springer-Verlag, pp. 467–476.

Gasparini, M., and Eisele, J. (2000), "A Curve-Free Method for Phase I Clinical Trials," *Biometrics*, 56, 609–615; corr. 57, 659–660.

Kruskal, W. (1958), "Ordinal Measures of Association," *Journal of the American Statistical Association*, 53, 814–861.

O'Quigley, J., Pepe, M., and Fisher, L. (1990), "Continual Reassessment Method: A Practical Design for Phase I Clinical Trials in Cancer," *Biometrics*, 46, 33–48.

Piantadosi, S., Fisher, J. D., and Grossman, S. (1998), "Practical Implementation of a Modified Continual Reassessment Method for Dose-Finding Trials," *Cancer Chemotherapy and Pharmacology*, 41, 429–436.

Storer, B. E. (1989), "Design and Analysis of Phase I Clinical Trials," *Biometrics* 45, 925–937.