Published in final edited form as:

Clin Trials. 2018 August ; 15(4): 386–397. doi:10.1177/1740774518772309.

Modeling adverse event counts in phase I clinical trials of a cytotoxic agent

Daniel G Muenz, Thomas M Braun, and Jeremy MG Taylor

Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Abstract

Background/aims—The goal of phase I clinical trials for cytotoxic agents is to find the maximum dose with an acceptable risk of severe toxicity. The most common designs for these dose-finding trials use a binary outcome indicating whether a patient had a dose-limiting toxicity. However, a patient may experience multiple toxicities, with each toxicity assigned an ordinal severity score. The binary response is then obtained by dichotomizing a patient's richer set of data. We contribute to the growing literature on new models to exploit this richer toxicity data, with the goal of improving the efficiency in estimating the maximum tolerated dose.

Methods—We develop three new, related models that make use of the total number of dose-limiting and low-level toxicities a patient experiences. We use these models to estimate the probability of having at least one dose-limiting toxicity as a function of dose. In a simulation study, we evaluate how often our models select the true maximum tolerated dose, and we compare our models with the continual reassessment method, which uses binary data.

Results—Across a variety of simulation settings, we find that our models compare well against the continual reassessment method in terms of selecting the true optimal dose. In particular, one of our models which uses dose-limiting and low-level toxicity counts beats or ties the other models, including the continual reassessment method, in all scenarios except the one in which the true optimal dose is the highest dose available. We also find that our models, when not selecting the true optimal dose, tend to err by picking lower, safer doses, while the continual reassessment method errs more toward toxic doses.

Conclusions—Using dose-limiting and low-level toxicity counts, which are easily obtained from data already routinely collected, is a promising way to improve the efficiency in finding the true maximum tolerated dose in phase I trials.

Keywords

Adaptive design; Dose finding; Dose-limiting toxicity; Maximum tolerated dose; Bayesian method

Introduction

Phase I clinical trials are small studies whose aim is to identify the maximum tolerated dose (MTD) of a treatment, i.e., the highest dose with an acceptable probability of severe dose-limiting toxicity (DLT). The selected dose is then recommended for additional, larger scale testing in a phase II trial. For cytotoxic agents such as chemotherapy, a key feature of phase I trials is that they measure only toxicity and ignore efficacy.¹ The omission of efficacy data can be justified by assuming that as dose increases, so do the probabilities of both DLT and therapeutic effect. Therefore, the MTD, based solely on toxicity data, is optimal in the sense that increasing the dose beyond it would lead to unacceptable risk of toxicity, while decreasing the dose would decrease the efficacy.

Phase I trials typically enroll patients sequentially in small cohorts, adaptively changing what dose to assign to the next cohort. At the end of the trial, one of the pre-specified doses will be selected as the MTD. Two competing methods for implementing this framework are the 3+3 design,² which uses a rule-based algorithm for escalating the dose towards the MTD, and the continual reassessment method (CRM),^{3–6} which adaptively updates a statistical model of DLT rates for each dose and uses model-based estimates of those rates to select the next dose. Both the 3+3 design and the CRM use a binary indicator of whether a patient had a DLT.

While most phase I trials have used binary data,^{7;8} richer data are routinely collected. We refer to the “full data” as a vector of event grades for each patient, capturing all events a patient has across multiple event types, e.g., infections, gastrointestinal disorders, or hematologic events. Each grade is an integer from 0 to 5, with higher grades indicating greater severity, according to the Common Terminology Criteria for Adverse Events.⁹ A DLT is commonly taken to be an event of grade 3 or higher, although this may vary by protocol.

A number of authors have explored using this richer toxicity data in different ways. Bekele and Thall¹⁰ developed a method to incorporate the full data, adopting a multivariate ordinal probit model. The model estimates, for each dose level, the probabilities of observing each grade for every event type. Rather than base their adaptive dose-finding algorithm on DLT rates, Bekele and Thall defined a new metric, the “total toxicity burden,” as a sum of weighted event grade probabilities, with weights elicited from physicians and probabilities estimated by the model. The algorithm then seeks a pre-specified target total toxicity burden to define the MTD. The appeal of this approach is that it never dichotomizes the toxicity information, either before modeling or after in selecting the MTD, and it provides a different target than the DLT rate. A drawback, however, is that it is not straightforward to think in the total toxicity burden scale, and thus choosing and explaining the target is challenging.

Yuan et al.¹¹ developed a quasi-continuous toxicity score to reflect the highest-grade event a patient experiences. A quasi-Bernoulli likelihood is used to carry out estimation and inform dose escalation decisions. This method worked well in simulation studies, but similar to the CRM it only uses one event per patient. Thus, Chen et al.¹² extended Yuan et al. to the case where patients may have multiple toxicities with varying grades. They used isotonic

regression to estimate the average toxicity score at each dose level. Alternative quasi-continuous toxicity scores have been proposed by Ezzalfani et al.¹³ and Wang and Ivanova.¹⁴ Lee et al.¹⁵ proposed using multiple constraints on the percentiles of a toxicity score, as opposed to the typical single constraint of the score being less than a threshold. Their multi-constraint framework can work with any of the aforementioned scores. As with Bekele et al., using these methods with novel scores requires redefining the MTD.

Van Meter et al.¹⁶ proposed using the original ordinal adverse event grades (0 to 5) in a proportional odds model, finding similar performance to the CRM in how often the correct MTD was selected. As with Yuan et al., the response for each patient was a single grade, which could be taken as the maximum grade across all the patient's events.

We propose three new models that make use of phase I data in a way that we have not previously seen in the literature. Specifically, one model requires all of a patient's events to be categorized as DLTs or not, and models the total count of DLTs per patient. Our two other models require that events be categorized as DLTs or low-level toxicities, which we define as events of medical significance but which should not prevent a patient from continuing treatment, and we model the total number of DLTs and low-level toxicities per patient. As a motivation for our model, we note that in a review of 54 single-agent phase I trials, Postel-Viney et al.¹⁷ found a total of 2,084 patients experienced 24,918 adverse events, including 189 patients who experienced 300 DLTs. These numbers imply that, on average, patients experienced 12 events, and that 9% of patients had an average of 1.6 DLTs. Thus, data with multiple events and in particular multiple DLTs per patient are available. Our method uses the same adaptive dose-finding framework as the CRM, but incorporates these richer data in an attempt to better estimate the DLT rates.

Data and notation

Let N_i be the DLT count, and M_i the total event count (DLTs and low-level toxicities) for patient i , so the patient experiences a DLT if $N_i > 0$. These counts are obtained from the full data, with a clinician categorizing each adverse event as a DLT or low-level toxicity. A patient may experience multiple adverse events, both across and within event types, so in principle N_i and M_i are unbounded above. With data from n patients, let $N_+ = \sum_{i=1}^n N_i$ and $M_+ = \sum_{i=1}^n M_i$ be the dose level assigned to patient i among J pre-specified levels. We define $r_+ = \sum_{i=1}^n r_{X_i}$ as the total scaled dose given to patients. The scaled dose values $r_1, \dots, r_J$, will be defined in the next section.

A common feature in phase I designs is the need for investigators to specify, *a priori*, a so-called probability skeleton. This represents the best prior guess for the DLT rate for each candidate dose, which we denote π_{0j} , $j = 1, \dots, J$. To ensure that the probability of DLT increases monotonically with dose, we require $\pi_{01} < \dots < \pi_{0J}$.

Models and methods

In this section we specify our three models, in increasing order of complexity, and we show how each model is used to estimate DLT rates.

Model 1: Using DLT counts

Our first model uses only DLT counts, not low-level toxicity counts. We model the DLT count as an increasing function of dose, $N_i | \beta, X_i = j \sim \text{Poisson}(\beta r_j)$, where r_j is an increasing function of j that rescales the doses to agree with the skeleton. We will show later how to calculate r_j . As with the CRM, our model has only one parameter, β , which we give a Bayesian prior, $\beta \sim \text{Gamma}(\sigma_\beta^{-2}, \sigma_\beta^{-2})$, such that, *a priori*, $E(\beta) = 1$ and $sd(\beta) = \sigma_\beta$.

According to our model, the probability of a patient having any DLTs at dose j is

$$\begin{aligned}\pi_j(\beta) &= \Pr(N_i > 0 | \beta, j) \\ &= 1 - \Pr(N_i = 0 | \beta, j) \\ &= 1 - e^{-\beta r_j}.\end{aligned}$$

Thus, we equate the prior mean of $\pi_j(\beta)$ with the skeleton,

$$\begin{aligned}\pi_{0j} &= E[\pi_j(\beta)] = 1 - (1 + r_j \sigma_\beta^2)^{-\sigma_\beta^{-2}}. \text{ Solving this equation for } r_j \text{ gives} \\ r_j &= \sigma_\beta^{-2} \left[(1 - \pi_{0j})^{-\sigma_\beta^{-2}} - 1 \right]. \text{ The hyperparameter } \sigma_\beta \text{ controls the amount of prior uncertainty} \\ &\text{around the skeleton.}\end{aligned}$$

Model 1: Likelihood and estimation

The likelihood for data from n patients is $L_\beta = \prod_{i=1}^n e^{-\beta r_{X_i}} \frac{(\beta r_{X_i})^{N_i}}{N_i!} \propto e^{-\beta r_+ N_+}$. Using

a gamma prior $p(\beta) \propto e^{-\sigma_\beta^{-2} \beta} \beta^{\sigma_\beta^{-2} - 1}$, the posterior distribution of β is $\beta | \text{Data} \sim \text{Gamma}(\tilde{a}_\beta, \tilde{b}_\beta)$, where $\tilde{a}_\beta = \sigma_\beta^{-2} + N_+$ and $\tilde{b}_\beta = \sigma_\beta^{-2} + r_+$. We estimate the DLT rate, specifically the probability of having any DLTs at dose j , denoted $\hat{\pi}_j$, as the posterior mean of $\pi_j(\beta)$: $\hat{\pi}_j = E[\pi_j(\beta) | \text{Data}] = 1 - (1 + r_j \tilde{b}_\beta)^{-\tilde{a}_\beta}$. Thus, we have a closed-form Bayesian estimator for DLT rates, and no numerical integration or Markov chain Monte Carlo is necessary.

Model 2: Using DLT and low-level toxicity counts

Our second model incorporates more information than Model 1. First, we use a Poisson distribution to model a patient's total event count M_i , such that $M_i | \theta, X_i = j \sim \text{Poisson}(\theta r_j)$, $\theta > 0$, where again r_j rescales the dose to agree with the skeleton. Note that we reuse the

symbol r_j to emphasize the similarity with Model 1, but the calculation of r_j is separate for the two models.

Next, conditional on M_j , the number of DLTs N_j is given a binomial distribution, $N_j|q, M_j, X_j \sim \text{Binomial}(M_j, q)$. The parameter q is the probability that a given event is a DLT rather than a low-level toxicity. Note that, although we condition on dose X_j , we do not include them on the right-hand side. Thus we are assuming that the proportion of DLTs is constant across doses; we will relax this assumption in Model 3. We assume that θ and q are *a priori* independent with marginal priors $\theta \sim \text{Gamma}(\sigma_\theta^{-2}, \sigma_\theta^{-2})$ and $q \sim \text{Beta}(a_q, b_q)$.

The modeled probability of a patient having any DLTs at dose j is

$$\begin{aligned}\pi_j(\theta, q) &= \Pr(N_j > 0 | \theta, q, j) \\ &= 1 - \sum_{m=0}^{\infty} \Pr(N_j = 0 | N_j = 0 | q, m, j) \Pr(M_j = m | \theta, j) \\ &= 1 - \sum_{m=0}^{\infty} (1-q)^m e^{-\theta r_j} (\theta r_j)^m / m! \\ &= 1 - e^{-\theta q r_j}.\end{aligned}$$

Thus, we calculate r_j by solving the equation

$$\begin{aligned}\pi_{0j} &= E[\pi_j(\theta, q)] = 1 - E[e^{-\theta q r_j}] \\ &= 1 - \int_0^1 \left[\int_0^\infty e^{-\theta q r_j} p(\theta) d\theta \right] p(q) dq \\ &= 1 - \int_0^1 (1 + q r_j \sigma_\theta^2)^{-\sigma_\theta^{-2}} p(q) dq \\ &= 1 - \frac{{}_2F_1(b_q, \sigma_\theta^{-2}; a_q + b_q; r_j / [r_j + \sigma_\theta^{-2}])}{(1 + r_j \sigma_\theta^2)^{\sigma_\theta^{-2}}}\end{aligned}$$

where ${}_2F_1$ is the Gaussian hypergeometric function¹⁸, and $p(\cdot)$ denotes a density. This equation can be solved numerically for r_j for any choice of skeleton π_{0j} and hyperparameters σ_θ , a_q , and b_q .

Model 2: Likelihood and estimation

With data for n patients, the full likelihood is

$$\begin{aligned}
L_{\theta, p} &= \prod_{i=1}^n \Pr(M_i, N_i | \theta, q, X_i) \\
&= \prod_{i=1}^n \Pr(M_i | \theta, X_i) \prod_{i=1}^n \Pr(N_i | q, M_i, X_i) \\
&= L_{\theta} L_q
\end{aligned}$$

where $L_{\theta} = \prod_{i=1}^n e^{-\theta r_{X_i}} \frac{M_i!}{(\theta r_{X_i})^{M_i}} \propto e^{-\theta r_+} \theta^{M_+}$ and

$L_q \propto \prod_{i=1}^n q^{N_i} (1-q)^{M_i - N_i} = q^{N_+} (1-q)^{M_+ - N_+}$. Due to the separability of the likelihood and the prior independence of the parameters, we can estimate θ separately from q . Using a gamma prior for θ , we have $\theta | \text{Data} \sim \text{Gamma}(\tilde{a}_{\theta}, \tilde{b}_{\theta})$, where $\tilde{a}_{\theta} = \sigma_{\theta}^{-2} + M_+$ and $\tilde{b}_{\theta} = \sigma_{\theta}^{-2} + r_+$. For q , using a beta prior gives us $q | \text{Data} \sim \text{Beta}(\tilde{a}_q, \tilde{b}_q)$, where $\tilde{a}_q = a_q + N_+$ and $\tilde{b}_q = b_q + M_+ - N_+$.

Finally, we estimate the posterior DLT rate at dose j as

$\hat{\pi}_j = E[\pi_j(\theta, q) | \text{Data}] = 1 - {}_2F_1(\tilde{b}_q, \tilde{a}_q; \tilde{a}_q + \tilde{b}_q; r_j / [r_j + \tilde{b}_{\theta}]) / (1 + r_j / \tilde{b}_{\theta})^{\tilde{a}_{\theta}}$. In the R programming language,¹⁹ the function `{}_2F_1` is available in the `gsl` package²⁰ under the name `hyperg_2F1`.

Model 3: Using DLT and low-level toxicity counts, version two

Our third model is very similar to Model 2. The key difference is that q , the probability of an event being a DLT rather than a low-level toxicity, is assumed to vary with dose. Specifically, we model $q = q_j(\alpha) = \text{expit}(-3 + e^{\alpha} r_j)$. The requirement that the coefficient of r_j is positive ensures that the log-odds of an event being a DLT increases with increasing dose level j . We now place a prior on α instead of q : $\alpha \sim \text{Normal}(\alpha_0, \sigma_{\alpha}^2)$.

According to Model 3, the probability of a patient having any DLTs at dose j is

$\pi_j(\theta, \alpha) = 1 - e^{-\theta q_j(\alpha) r_j}$. The rescaled doses r_j are found by numerically solving

$\pi_{0j} = E[\pi_j(\theta, \alpha)] = 1 - E[e^{-\theta q_j(\alpha) r_j}] = 1 - E[(1 + q_j(\alpha) r_j \sigma_{\theta}^2)^{-\sigma_{\theta}^{-2}}]$, where the last expectation is with respect to the distribution of α only.

Model 3: Likelihood and estimation

The likelihood factors into separate functions of the two parameters, $L_{\theta, \alpha} = L_{\theta} L_{\alpha}$. The function L_{θ} is the same as in Model 2; the function L_{α} is

$$\begin{aligned}
L_{\alpha} &= \prod_{i=1}^n \Pr(N_i | q, M_i, X_i) \\
&\propto \prod_{i=1}^n \left(\frac{e}{-3 + e^{\alpha} r_{X_i}} \right)^{N_i} \left(\frac{1}{-3 + e^{\alpha} r_{X_i}} \right)^{M_i - N_i} \\
&\propto \exp \left(e^{\alpha} \sum_{i=1}^n r_{X_i} N_i \right) \prod_{i=1}^n \left(1 + e^{-3 + e^{\alpha} r_{X_i}} \right)^{-M_i}.
\end{aligned}$$

Due to the separability of the likelihood and the prior independence of the parameters, we can estimate θ separately from α . The posterior for θ is the same as in Model 2:

$\theta | \text{Data} \sim \text{Gamma}(\tilde{a}_{\theta}, \tilde{b}_{\theta})$, where $\tilde{a}_{\theta} = \sigma_{\theta}^{-2} + M_{+}$ and $\tilde{b}_{\theta} = \sigma_{\theta}^{-2} + r_{+}$. For α , however, a closed-form posterior is not available. To obtain the posterior, we use a simple discrete grid-based approximation, which is computationally efficient for a single parameter.²¹

Finally, we estimate the posterior DLT rate at dose j as

$\hat{\pi}_j = E[\pi_j(\theta, \alpha) | \text{Data}] = 1 - E[e^{-\theta q_j(\alpha) r_j} | \text{Data}] = 1 - E[(1 + q_j(\alpha) r_j / \tilde{b}_{\theta})^{-\tilde{a}_{\theta}} | \text{Data}]$, where the final expectation is with respect to the posterior distribution of α .

Dose-finding algorithm

As a trial is run, patients are enrolled sequentially, and with data from each new patient we update our estimates of the DLT rates. We use the following algorithm to estimate the MTD. First, we choose a set of J dose levels, and we set a target DLT rate τ , denoting the maximum acceptable probability of a patient having any DLTs. We enroll the first patient at the lowest dose, level 1, and follow the patient for a desired time. Using the patient's data and one of our models, we estimate the DLT rates $\hat{\pi}_1, \dots, \hat{\pi}_J$. Our updated estimate of the MTD, denoted $\hat{MTD}$, is the dose j that minimizes $|\hat{\pi}_j - \tau|$. The next patient is then assigned either to $\hat{MTD}$ or to the lowest dose not yet assigned, whichever is smaller. This rule prevents the sequence of dose assignments from skipping over untested dose levels. We continue to enroll new patients and update $\hat{MTD}$ using all the patients' data. Once we have observed a desired number of patients, the trial concludes and we obtain a final estimate of the MTD.

Simulation study and comparison with CRM

We simulate adaptive trials using each of our proposed models, and we record how often the true MTD is selected. We include a comparison with a CRM model, described in the next section.

CRM model for comparison

We will use a CRM model that is related to all of our proposed models. Specifically, let $Y_i = \mathbb{I}(N_i > 0)$ be a binary indicator of whether patient i had a DLT. Recall that, according to our three models, the probability of DLT is, respectively, $\pi_j(\beta) = 1 - e^{-\beta r_j}$, $\pi_j(\theta, q) = 1 - e^{-\theta q r_j}$, and $\pi_j(\theta, \alpha) = 1 - e^{-\theta q_j(\alpha) r_j}$. For the CRM, with a single parameter ϕ , we have to specify a model for $\Pr(Y_i = 1 | \phi, j) = \Pr(N_i > 0 | \phi, j)$ directly. We use $\Pr(Y_i = 1 | \phi, j) = 1 - e^{-\phi r_j}$, $\phi > 0$, due to its functional similarity to our models, which will facilitate jointly calibrating the hyperparameters of each model. Our prior on ϕ is $\phi \sim \text{Gamma}(\sigma_\phi^{-2}, \sigma_\phi^{-2})$, and the rescaled dose is $r_j = \sigma_\phi^{-2} \left[(1 - \pi_{0j})^{-\sigma_\phi^2} - 1 \right]$. The posterior of ϕ is obtained with a discrete grid-based approximation, which is then used to calculate the posterior DLT rates $\hat{\pi}_j = 1 - E[e^{-\phi r_j} | \text{Data}]$.

Simulating a patient's toxicity data

In our simulation study, for the data-generating mechanism we assume that there are $K = 15$ distinct event types, each of which can experience either no event, a low-level toxicity, or a DLT, and there is a patient-specific frailty influencing the probabilities of those outcomes. The observed data for a patient are the total number of DLTs and low-level toxicities across the K event types. We chose to use $K = 15$ as a middle ground among the number of event types we have seen in published trials,^{10;17;22} however we will also conduct a small sensitivity analysis with $K = 7$.

Letting $Y_{ik} \in \{0 = \text{no event}, 1 = \text{low-level toxicity}, 2 = \text{DLT}\}$ be the response for patient i in event type $k = 1, \dots, K$, we use a cumulative logit model,

$$\text{logitPr}(Y_{ik} \geq y | X_i) = \eta_{iy} = \alpha_y + \gamma_i + r_{X_i} \quad \text{for } y = 1, 2, \text{ where } \gamma_i \sim \text{Normal}(0, \gamma_{\text{sd}}^2) \text{ is a random}$$

patient-specific frailty; X_i is the dose; and r_{X_i} rescales the dose to obtain a desired

probability of DLT. The multinomial probabilities $p_{iy} = \Pr(Y_{ik} = y | X_i)$ implied by this model are $p_{i0} = 1 - \text{expit}(\eta_{i1})$, $p_{i1} = \text{expit}(\eta_{i1}) - \text{expit}(\eta_{i2})$, and $p_{i2} = \text{expit}(\eta_{i2})$. Note that the linear predictor η_{iy} does not depend on the event type k , that is, for simplicity we make all event types have the same multinomial probabilities.

The number of DLTs for patient i is $N_i = \sum_{k=1}^K \mathbb{I}(Y_{ik} = 2) \sim \text{Binomial}(K, p_{i2})$ and the total event count is $M_i = \sum_{k=1}^K \mathbb{I}(Y_{ik} \in \{1, 2\}) \sim \text{Binomial}(K, p_{i1} + p_{i2})$. For dose j to have a population average DLT rate of π_j , we select r_j such that $\pi_j = E[\Pr(N_j > 0)] = E[1 - (1 - \text{expit}(\eta_{j2}))^K]$, where the expectation is with respect to γ_i . Since we rescale the doses, shifting the intercepts α_1 and α_2 by some constant δ has no effect on the probabilities. Thus, without loss of generality, we fix $\alpha_2 = 0$, and we require $\alpha_1 > 0$ to ensure positive probabilities. Once

calibrated, we draw $Y_{ik} \sim \text{Multinomial}(1; p_{k0}, p_{k1}, p_{k2})$ for $k = 1, \dots, K$. Then $N_i = \sum_{k=1}^K I(Y_{ik} = 2)$ and $M_i = \sum_{k=1}^K I(Y_{ik} \in \{1, 2\})$.

Simulation scenarios

Using our data-generating mechanism and dose-finding algorithm, we can simulate full trials. In each simulated trial, there are $J = 5$ candidate dose levels, we will enroll n patients one at a time, and the target DLT rate is $\tau = 30\%$. We use a skeleton of (10%, 20%, 30%, 40%, 50%). We must also specify some model-specific tuning parameters: For Model 1, we use $\sigma_\beta = 1$. For Model 2, we use $\sigma_\theta = 0.8$, $a_q = 2$, and $b_q = 8$. For Model 3, we use $\sigma_\theta = 0.8$, $\alpha_0 = 0.8$, and $\sigma_\alpha = 0.7$. For the CRM model, we use $\sigma_\phi = 1$. These hyperparameter values were chosen so that the prior DLT rate distributions at each dose are close across the four models, as shown in Figure 1.

We consider a variety of scenarios by varying three factors: (1) the sample size n , with either 30 or 60 patients; (2) the true DLT rates for all J doses, thereby determining which dose is the true MTD; and (3) the parameters for our simulation model, α_1 and γ_{sd} . All scenarios use $K = 15$ event types. For the scenarios we call s -A, $s = 1, \dots, J$, we set $\alpha_1 = 2$ and $\gamma_{sd} = 0.5$. For scenarios s -B, we set $\alpha_1 = 2.5$ and $\gamma_{sd} = 0.75$. The relevance of these parameters is that, in the A scenarios, the simulated event counts M_j are typically small, while in the B scenarios M_j is more likely to reach its upper bound of 15 (i.e., there is an event in all 15 event types). Thus the B scenarios present more of a challenge for our Models 1, 2, and 3, which have no concept of an upper bound for the event count. For example, in scenario 3-A, with true DLT rates (10%, 20%, 30%, 40%, 50%), the probability of a patient having at least 10 events at the different doses are (0%, 0.004%, 0.06%, 0.33%, 1.2%). In contrast, for scenario 3-B, with the same DLT rates, the probabilities of having at least 10 events are (0.02%, 0.38%, 1.8%, 4.9%, 11%). Table 1 shows the true DLT rates for each scenario. Note that in scenarios s -A and s -B, the true MTD is dose s .

Simulation results

For each simulation scenario and statistical model, we simulated 10,000 adaptive trials. The results are reported in Tables 2 and 3, for 30 and 60 patients, respectively, with the following summary statistics: the percentage of trials in which each dose is selected as the MTD; the average number of patients (out of 30 or 60) assigned to each dose during a trial; and a score that measures how far the true DLT rate of the selected MTD is from the target τ . We actually consider two scores, defined as $\text{score}_k = \sum_{j=1}^J \hat{s}_j \text{loss}_k(\pi_j, \tau)$, where $\hat{s}_j$ is the percentage of trials in which dose j is selected as the MTD, π_j is the true DLT rate at dose j , $\text{loss}_1(\pi, \tau) = |\pi - \tau|$, and $\text{loss}_2(\pi, \tau) = |\pi - \tau|I(\pi - \tau) + 2|\pi - \tau|I(\pi > \tau)$. Note that loss_2 implies a higher penalty than loss_1 for picking dose levels above the target rather than below. For both scores, smaller values are better, and a score of 0 is ideal.

In Table 2, we see that, in the A scenarios, our models consistently do at least as well and often better than the CRM, picking the correct MTD more often and treating more patients at the correct MTD. In the B scenarios, the results are more mixed. For scenarios 2-B through 4-B, our models typically do around as well as the CRM in selecting the correct

MTD. However, when not picking the correct MTD our models have a clear tendency to select doses lower than the MTD, while the CRM appears more likely to pick higher doses. This is reflected in the score_2 column, where our models consistently have lower scores. This trend is more pronounced in scenarios 1-B and 5-B, where the true MTD is the lowest or highest dose, respectively. In 1-B, our models do much better than the CRM, but this is likely because our models prefer lower doses. Similarly, in 5-B the CRM does considerably better, but this is likely because the CRM prefers higher doses.

It is difficult to identify a clear winner among the models across all scenarios. If we ignore scenario 5-B, then model 3, which incorporates low-level toxicities and allows the probability of DLT versus low-level toxicity to vary with dose, is the preferred model. This validates the idea that counting events and using low-level toxicities can improve estimation of the MTD. However, model 3 is clearly not preferred in scenario 5-B, in which the highest dose is the MTD and patients commonly experience many events, sometimes reaching the maximum of 15 events. We consider this scenario somewhat unlikely, however, as a well-designed trial will typically have the true MTD among the middle dose levels.

Example trial

We illustrate the behavior of Model 3 by presenting a full simulated trial. We simulated the trial data using scenario 4-A described previously, where the true DLT rates are (7%, 14%, 21%, 30%, 40%). The target DLT rate is 30%, so that dose 4 is the true MTD. Our skeleton is (10%, 20%, 30%, 40%, 50%), implying that *a priori* the MTD is believed to be dose 3. The results from simulating 30 patients are shown in Table 4. First we note that, looking at the last row of the table, the trial correctly concludes that dose 4 is the MTD. Indeed, although there is still some movement, the trial has mostly fixed on dose 4 by patient 12. Based on the results in Table 2, Model 3 will pick the correct dose in scenario 4-A 53.1% of the time, so this example trial is in line with expectations. We also note that only one patient has multiple DLTs in this trial, although 18 patients have more than one total event, so our model does have data to facilitate parameter estimation. At the end of the trial, beyond selecting the correct MTD, we have also estimated the DLT rates well: (7%, 12%, 18%, 27%, 38%).

Sensitivity analysis

We conduct a small sensitivity analysis, investigating how model performance varies with changing model parameters and with the number of event types K . In particular, starting from the 2-A scenario, we consider the following three variations: “Larger variance,” in which we increase the prior variances of model parameters ($\sigma_\beta = 1.4$ for Model 1, $\sigma_\theta = 1.15$, $a_q = 1$, and $b_q = 4$ for Model 2, $\sigma_\theta = 1.3$, $\alpha_0 = 2$, and $\sigma_a = 1.1$ for Model 3, and $\sigma_\phi = 1.4$ for the CRM); “Smaller variance,” in which we decrease the prior variances of model parameters ($\sigma_\beta = 0.7$ for Model 1, $\sigma_\theta = 0.56$, $a_q = 4$, and $b_q = 16$ for Model 2, $\sigma_\theta = 0.65$, $\alpha_0 = 3$, and $\sigma_a = 0.8$ for Model 3, and $\sigma_\phi = 0.7$ for the CRM); and “ $K = 7$ event types,” in which we reduce the number of events types from 15 to 7. We simulated 10,000 adaptive trials of 30 patients with each model, and for each described variation. Results are in Table 5, and we observe that all models are sensitive to the prior variances of model parameters, with the “smaller variance” models outperforming the “larger variance” models by about

10% in terms of selecting the correct MTD. In contrast, reducing the number of event types made minimal difference compared to the 2-A results in Table 2. These results suggest that the models, including the CRM, should be carefully calibrated given a variety of plausible DLT rates, while the precise number of event types is less important.

Model properties

To better understand the performance of our models, we investigate the bias in estimating DLT rates with each model. We simulate non-adaptive trials of 10 or 20 patients all receiving the same dose level, and we consider all combinations of true DLT rate and prior guess from 0% to 70% in increments of 2.5%. Patient responses are simulated using the A and B scenarios described previously.

Figure 2 shows the bias for both the CRM model and our Models 1, 2, and 3. In each plot, along the diagonal from bottom left to top right, the true DLT rate and prior guess agree, so we would expect minimal bias near this region. Above the diagonal, the prior guess exceeds the truth, so we would expect positive bias, and vice versa below the diagonal. In Table 2, we note that our models tend to choose doses below the true MTD more often than the CRM. As a plausible explanation, we see in Figure 2 that our models tend to be more positively biased than the CRM (particularly Models 2 and 3). That is, our models overestimate the DLT rates more than the CRM, and thus the dose selection algorithm picks lower doses. However, large differences between the truth and the skeleton tend to lead to greater bias for the CRM than for our models.

We note that Models 2 and 3 tend to select the true MTD less often in the B scenarios than in the A scenarios. This is likely due to a difference between the data-generating model and Models 2 and 3. Specifically, the data-generating model creates low-level toxicity and DLT event counts for a patient, and the sum of these two counts cannot exceed 15, thereby making the two counts negatively correlated. However, Models 2 and 3 implicitly force this correlation to be zero. In the A scenarios, both counts are usually low, so they do not influence each other much, and there is minimal conflict with our models. The B scenarios, in contrast, generate higher event counts for a patient, leading to stronger correlations and greater conflict with our models. If we increase the number of event types to $K = 60$ in our generative model, so that the upper bound is almost never realized in scenarios A and B, then the differences in dose selection percentages between scenarios A and B decreases, although using 60 event types would not be realistic.

Additionally, we note that the DLT count N_i either explicitly, as in Model 1, or implicitly, as in Models 2 and 3 after marginalizing over M_i , follows a Poisson distribution in our proposed analysis models, while in the data-generating model N_i is binomial. This discrepancy impacts the performance of our models, yet it has no effect on the CRM, which does not attempt to model event counts. For this reason, the CRM should perform exactly the same in scenarios A and B, within Monte Carlo simulation error.

Discussion

Models 1, 2, and 3 all make use of phase I trial data that are routinely collected but rarely analyzed or incorporated into the dose selection algorithm. Model 1 makes fewer assumptions than Models 2 and 3, placing a Poisson distribution on the number of DLTs, but disregarding the number of low-level toxicities. Models 2 and 3 jointly model a patient's DLT and low-level toxicity counts.

All three models perform comparably to, and sometimes better than, the CRM across a variety of scenarios in our simulation study. Nonetheless, the simulations did suggest that including total number of toxicities and DLTs might lead to conservative dose decisions when the true MTD is at the highest dose. However, we do observe better operating characteristics for some of our models when the MTD is one of the lower doses.

The data-generating model in our simulations was designed both to be biologically plausible and to disagree with our three proposed models, thereby creating fair test scenarios. Although we believe our models generally compare favorably with the CRM, in scenario 5-B, Models 2 and 3 perform poorly relative to both the CRM and to Model 1. The import of this differential depends on how plausible one believes scenario 5-B to be. If the true MTD is rarely the highest dose under consideration, then Model 3 clearly bests both the CRM and our Models 1 and 2. This suggests that we are indeed gaining useful information and efficiency by modeling events counts and incorporating low-level toxicities.

Models 1 to 3 tended to slightly overestimate DLT rates in simulations, likely because these models do not put an upper limit on the number of events a patient may have, while the data-generating model does. It may be possible to compensate for some of this bias by specifying a probability skeleton lower than one believes *a priori*, perhaps by raising the skeleton to a power slightly greater than one. The precise power to use can be chosen via simulation.

In a real trial, it may be desirable to enroll a new patient before the previous patients have finished follow-up. Our models can be extended to handle partial follow-up by scaling the expected number of toxicities in accordance with follow-up, in a manner similar to the time-to-event CRM.²³ Further research is needed to recommend appropriate scaling weights.

Acknowledgments

Funding

This research was partially supported by National Institutes of Health grant T32 CA083654.

References

- 1 Le Tourneau C, Lee J, Siu L. Dose escalation methods in phase I cancer clinical trials. *JNCI Journal of the National Cancer Institute*. 2009; 101(10):708–720. [PubMed: 19436029]
- 2 Storer B. Design and analysis of phase I clinical trials. *Biometrics*. 1989; 45(3):925–937. [PubMed: 2790129]
- 3 O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*. 1990; 46(1):33–48. [PubMed: 2350571]

- 4Faries D. Practical modifications of the continual reassessment method for phase I cancer clinical trials. *Journal of Biopharmaceutical Statistics*. 1994; 4(2):147–164. [PubMed: 7951271]
- 5Goodman S, Zahurak M, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine*. 1995; 14(11):1149–1161. [PubMed: 7667557]
- 6O'Quigley J, Shen L. Continual reassessment method: A likelihood approach. *Biometrics*. 1996; 52(2):673–684. [PubMed: 8672707]
- 7Hansen A, Graham D, Pond G, et al. Phase 1 trial design: Is 3 + 3 the best? *Cancer Control*. 2014; 21(3):200–208. [PubMed: 24955703]
- 8Rogatko A, Schoeneck D, Jonas W, et al. Translation of innovative designs into phase I trials. *Journal of Clinical Oncology*. 2007; 25(31):4982–6. [PubMed: 17971597]
- 9CTCAE. Cancer Therapy Evaluation Program Common Terminology Criteria for Adverse Events, Version 3.0 DCTD, NCI, NIH, DHHS: 2003 <https://ctep.cancer.gov>
- 10Bekele BN, Thall P. Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association*. 2004; 99(1):26–35.
- 11Yuan Z, Chappell R, Bailey H. The continual reassessment method for multiple toxicity grades: A Bayesian quasi-likelihood approach. *Biometrics*. 2007; 63(1):173–179. [PubMed: 17447942]
- 12Chen Z, Krailo M, Azen S, et al. A novel toxicity scoring system treating toxicity response as a quasi-continuous variable in phase I clinical trials. *Contemporary Clinical Trials*. 2010; 31(5):473–482. [PubMed: 20609419]
- 13Ezzalfani M, Zohar S, Qin R, et al. Dose-finding designs using a novel quasi-continuous endpoint for multiple toxicities. *Statistics in Medicine*. 2013; 32(16):2728–2746. [PubMed: 23335156]
- 14Wang Y, Ivanova A. Dose finding with continuous outcome in phase I oncology trials. *Pharmaceutical Statistics*. 2015; 14(2):102–107. [PubMed: 25408518]
- 15Lee SM, Cheng B, Cheung YK. Continual reassessment method with multiple toxicity constraints. *Biostatistics*. 2011; 12(2):386–398. [PubMed: 20876664]
- 16Van Meter E, Garrett-Mayer E, Bandyopadhyay D. Proportional odds model for dose-finding clinical trial designs with ordinal toxicity grading. *Statistics in Medicine*. 2011; 30(17):2070–2080. [PubMed: 21344472]
- 17Postel-Vinay S, Collette L, Paoletti X, et al. Towards new methods for the determination of dose limiting toxicities and the assessment of the recommended dose for further studies of molecularly targeted agents—dose-Limiting Toxicity and Toxicity Assessment Recommendation Group for Early Trials of Targeted therapies, an European Organisation for Research and Treatment of Cancer-led study. *European Journal of Cancer*. 2014; 50(12):2040–2049. [PubMed: 24880774]
- 18Andrews G, Askey R, Roy R. *Special Functions* Cambridge: Cambridge University Press; 1999chapter 2
- 19R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing; Vienna, Austria: 2016URL <https://www.R-project.org/>
- 20Hankin R. Special functions in R: Introducing the gsl package. *R News*. 2006; 6(4)
- 21Hoff P. *A First Course in Bayesian Statistical Methods* New York, NY: Springer New York; 2009Springer Texts in Statisticschapters 6 and 10
- 22Yong K, Cavet J, Johnson P, et al. Phase I study of KW-2478, a novel Hsp90 inhibitor, in patients with B-cell malignancies. *British Journal of Cancer*. 2016; 114(1):7–13. [PubMed: 26695442]
- 23Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics*. 2000; 56(4):1177–1182. [PubMed: 11129476]

Figure 1.

Prior distributions of DLT rates at each dose, for all four models. The boxplots show the 10th, 25th, 50th, 75th, and 90th percentiles, and the × marks the mean. By design, the prior means equal the skeleton.

DLT: dose-limiting toxicity; CRM: continual reassessment method.

Figure 2.

Bias in estimating DLT rates for CRM model and Models 1, 2, and 3. Each pixel in each plot is based on 10,000 simulated trials under scenarios A and B of either 10 or 20 patients. In a trial, all patients receive the same dose level, i.e., the trial is not adaptive, and the true DLT rate is the same. The purpose of this figure is to show, for every combination of true DLT rate and prior guess (i.e. skeleton), how well the models do in estimating the DLT rate. DLT: dose-limiting toxicity; CRM: continual reassessment method.

Table 1

True dose-limiting toxicity rates for each simulation scenario.

Scenarios	Dose level				
	1	2	3	4	5
1-A and 1-B	30%	40%	50%	60%	70%
2-A and 2-B	20%	30%	40%	50%	60%
3-A and 3-B	10%	20%	30%	40%	50%
4-A and 4-B	7%	14%	21%	30%	40%
5-A and 5-B	6%	12%	18%	24%	30%

Table 2

Simulation study results. Each line summarizes 10,000 simulated adaptive trials of 30 patients with 5 dose levels. The target DLT rate is 30%. The *italicized* rows show the true DLT rates. The **bold** entries highlight the results at the true MTD.

Scenario	Model	Dose Selection Percentage (Number of Patients Treated)					Score ₁	Score ₂
		1	2	3	4	5		
		30%	40%	50%	60%	70%		
1-A	1	57.3 (13.4)	40.0 (13.0)	2.7 (2.9)	0.0 (0.6)	0.0 (0.2)	4.5	9.1
	2	73.1 (18.9)	25.8 (9.3)	1.1 (1.6)	0.0 (0.2)	0.0 (0.0)	2.8	5.6
	3	63.2 (15.6)	35.6 (12.6)	1.2 (1.7)	0.0 (0.1)	0.0 (0.0)	3.8	7.6
	CRM	51.5 (11.9)	43.4 (13.8)	4.9 (3.5)	0.2 (0.7)	0.0 (0.2)	5.4	10.8
1-B	1	67.2 (15.6)	31.2 (11.3)	1.6 (2.4)	0.1 (0.5)	0.0 (0.2)	3.5	6.9
	2	83.8 (22.2)	15.5 (6.7)	0.6 (1.0)	0.0 (0.1)	0.0 (0.0)	1.7	3.3
	3	83.7 (21.2)	16.2 (8.3)	0.1 (0.6)	0.0 (0.0)	0.0 (0.0)	1.6	3.3
	CRM	50.9 (11.7)	44.2 (13.9)	4.8 (3.5)	0.2 (0.7)	0.0 (0.2)	5.4	10.9
2-A	1	14.4 (5.1)	61.0 (15.2)	22.7 (7.3)	1.9 (1.8)	0.1 (0.6)	4.1	6.8
	2	22.7 (7.7)	58.6 (15.1)	17.6 (5.9)	1.1 (1.1)	0.0 (0.2)	4.3	6.2
	3	14.5 (5.5)	63.9 (16.3)	20.6 (7.3)	1.0 (0.8)	0.0 (0.0)	3.7	6.0
	CRM	14.3 (4.9)	57.1 (14.6)	25.0 (7.7)	3.6 (2.1)	0.1 (0.7)	4.7	7.9
2-B	1	20.7 (6.5)	61.6 (15.3)	16.6 (6.2)	1.0 (1.5)	0.0 (0.5)	4.0	5.8
	2	37.6 (11.9)	51.5 (13.4)	10.3 (3.9)	0.6 (0.7)	0.0 (0.1)	4.9	6.1
	3	29.1 (9.1)	63.2 (17.2)	7.5 (3.6)	0.2 (0.2)	0.0 (0.0)	3.7	4.5
	CRM	13.8 (4.8)	57.6 (14.8)	25.3 (7.7)	3.2 (2.1)	0.2 (0.7)	4.6	7.8
3-A	1	0.5 (1.5)	26.4 (9.1)	53.5 (12.2)	17.5 (5.3)	2.1 (1.9)	4.9	7.1
	2	0.8 (1.7)	29.3 (9.8)	53.2 (12.4)	15.3 (4.9)	1.5 (1.2)	4.9	6.8
	3	0.4 (1.5)	22.4 (8.2)	59.6 (14.5)	17.0 (5.4)	0.6 (0.5)	4.1	6.0
	CRM	0.8 (1.6)	26.3 (9.1)	49.7 (11.6)	20.1 (5.5)	3.2 (2.2)	5.4	8.1

Scenario	Model	Dose Selection Percentage (Number of Patients Treated)					Score ₁	Score ₂
		1	2	3	4	5		
3-B	1	1.1 (1.8)	33.6 (10.3)	50.5 (11.5)	13.6 (4.6)	1.3 (1.7)	5.2	6.8
	2	2.3 (2.9)	40.2 (12.3)	46.2 (10.7)	10.4 (3.4)	0.9 (0.8)	5.7	6.9
	3	0.9 (1.9)	39.0 (12.4)	53.6 (13.3)	6.4 (2.3)	0.0 (0.1)	4.7	5.4
	CRM	0.8 (1.6)	27.2 (9.2)	48.6 (11.3)	20.0 (5.6)	3.3 (2.2)	5.6	8.2
		7%	14%	21%	30%	40%		
4-A	1	0.0 (1.1)	3.7 (3.8)	33.2 (9.4)	44.8 (9.5)	18.2 (6.2)	5.4	7.2
	2	0.0 (1.1)	3.3 (3.4)	33.4 (9.7)	47.0 (10.4)	16.2 (5.4)	5.2	6.8
	3	0.0 (1.1)	2.7 (3.1)	29.3 (9.5)	53.1 (12.1)	14.9 (4.2)	4.6	6.0
	CRM	0.0 (1.2)	4.9 (4.2)	30.1 (9.0)	44.1 (9.1)	20.8 (6.5)	5.6	7.7
4-B	1	0.1 (1.2)	5.9 (4.5)	37.7 (9.9)	41.8 (9.0)	14.5 (5.4)	5.8	7.2
	2	0.1 (1.4)	8.3 (5.5)	41.2 (10.9)	39.4 (8.5)	11.0 (3.7)	6.2	7.3
	3	0.0 (1.2)	5.3 (4.5)	44.7 (12.9)	45.3 (10.0)	4.8 (1.4)	5.3	5.8
	CRM	0.0 (1.2)	5.1 (4.4)	31.6 (9.0)	42.1 (8.9)	21.3 (6.5)	5.8	7.9
		6%	12%	18%	24%	30%		
5-A	1	0.0 (1.1)	1.5 (2.8)	14.1 (6.1)	33.1 (8.1)	51.2 (11.9)	4.0	4.0
	2	0.0 (1.1)	0.9 (2.3)	12.0 (5.7)	35.4 (8.9)	51.7 (12.1)	3.7	3.7
	3	0.0 (1.1)	0.9 (2.2)	11.6 (6.0)	38.8 (10.4)	48.7 (10.3)	3.9	3.9
	CRM	0.0 (1.1)	2.1 (3.2)	15.6 (6.5)	31.8 (7.6)	50.4 (11.5)	4.2	4.2
5-B	1	0.0 (1.1)	2.4 (3.0)	16.9 (6.6)	35.4 (8.2)	45.3 (11.1)	4.6	4.6
	2	0.0 (1.2)	2.7 (3.5)	20.5 (7.7)	37.8 (8.8)	38.8 (8.8)	5.2	5.2
	3	0.0 (1.1)	1.6 (3.0)	21.9 (9.0)	48.5 (11.4)	27.9 (5.5)	5.8	5.8
	CRM	0.0 (1.1)	2.0 (3.2)	15.4 (6.4)	32.1 (7.7)	50.5 (11.6)	4.1	4.1

DLT: dose-limiting toxicity; MTD: maximum tolerated dose; CRM: continual reassessment method.

Table 3

Simulation study results. Each line summarizes 10,000 simulated adaptive trials of 60 patients with 5 dose levels. The target DLT rate is 30%. The *italicized* rows show the true DLT rates. The **bold** entries highlight the results at the true MTD.

Scenario	Model	Dose Selection Percentage (Number of Patients Treated)					Score ₁	Score ₂
		1	2	3	4	5		
		30%	40%	50%	60%	70%		
1-A	1	68.3 (32.3)	31.5 (23.8)	0.2 (3.2)	0.0 (0.6)	0.0 (0.2)	3.2	6.4
	2	80.3 (41.9)	19.5 (16.0)	0.2 (1.9)	0.0 (0.2)	0.0 (0.0)	2.0	4.0
	3	72.7 (36.0)	27.1 (22.0)	0.1 (1.9)	0.0 (0.1)	0.0 (0.0)	2.7	5.5
	CRM	61.4 (28.9)	37.7 (26.0)	0.9 (4.2)	0.0 (0.7)	0.0 (0.2)	4.0	7.9
1-B	1	79.4 (37.6)	20.5 (19.2)	0.1 (2.5)	0.0 (0.5)	0.0 (0.2)	2.1	4.2
	2	90.1 (48.2)	9.8 (10.6)	0.0 (1.1)	0.0 (0.1)	0.0 (0.0)	1.0	2.0
	3	91.1 (47.5)	8.9 (11.9)	0.0 (0.6)	0.0 (0.0)	0.0 (0.0)	0.9	1.8
	CRM	61.0 (28.7)	37.9 (26.2)	1.1 (4.3)	0.0 (0.7)	0.0 (0.2)	4.0	8.0
2-A	1	12.6 (9.3)	72.4 (35.6)	14.9 (12.7)	0.1 (1.9)	0.0 (0.5)	2.8	4.3
	2	19.1 (13.8)	68.8 (34.6)	12.0 (10.2)	0.2 (1.3)	0.0 (0.2)	3.1	4.4
	3	12.4 (9.2)	73.4 (37.2)	14.1 (12.6)	0.1 (0.9)	0.0 (0.0)	2.7	4.1
	CRM	11.8 (8.7)	67.1 (33.6)	20.5 (14.5)	0.6 (2.5)	0.0 (0.7)	3.4	5.5
2-B	1	19.5 (12.7)	72.0 (35.4)	8.5 (9.9)	0.0 (1.6)	0.0 (0.5)	2.8	3.7
	2	32.7 (22.3)	61.6 (30.4)	5.6 (6.4)	0.1 (0.8)	0.0 (0.1)	3.9	4.4
	3	27.1 (17.9)	68.7 (36.5)	4.2 (5.5)	0.0 (0.2)	0.0 (0.0)	3.1	3.6
	CRM	11.8 (8.8)	67.4 (33.6)	20.2 (14.4)	0.6 (2.5)	0.0 (0.7)	3.3	5.5
3-A	1	0.0 (1.6)	20.8 (16.1)	66.6 (30.1)	12.3 (10.0)	0.2 (2.2)	3.4	4.6
	2	0.0 (1.7)	22.2 (17.3)	66.8 (30.6)	10.8 (9.0)	0.1 (1.4)	3.3	4.4
	3	0.1 (1.5)	18.1 (14.3)	70.3 (33.9)	11.4 (9.7)	0.1 (0.6)	3	4.2
	CRM	0.2 (1.7)	20.1 (15.9)	62.4 (28.4)	16.8 (11.3)	0.6 (2.7)	3.8	5.6

Scenario	Model	Dose Selection Percentage (Number of Patients Treated)					Score ₁	Score ₂
		1	2	3	4	5		
3-B	1	0.2 (2.0)	29.6 (19.6)	62.9 (28.8)	7.3 (7.7)	0.0 (1.9)	3.7	4.5
	2	0.4 (3.2)	36.2 (23.9)	57.0 (26.1)	6.3 (5.9)	0.1 (0.9)	4.3	5.0
	3	0.1 (2.0)	31.7 (22.8)	64.1 (31.4)	4.1 (3.8)	0.0 (0.1)	3.6	4.0
	CRM	0.1 (1.7)	19.8 (15.8)	63.2 (28.9)	16.3 (11.0)	0.6 (2.7)	3.8	5.5
		7%	14%	21%	30%	40%		
4-A	1	0.0 (1.1)	0.8 (4.5)	28.1 (18.6)	59.9 (25.4)	11.2 (10.4)	3.8	4.9
	2	0.0 (1.1)	0.8 (4.0)	26.9 (18.3)	61.6 (27.0)	10.7 (9.6)	3.6	4.7
	3	0.0 (1.1)	0.7 (3.5)	24.7 (17.9)	64.6 (29.8)	10.0 (7.8)	3.3	4.3
	CRM	0.0 (1.2)	1.3 (5.1)	25.6 (17.5)	57.4 (24.2)	15.8 (12.0)	4.1	5.7
4-B	1	0.0 (1.2)	2.0 (5.7)	37.5 (21.3)	53.8 (23.3)	6.8 (8.4)	4.4	5.1
	2	0.0 (1.4)	2.2 (6.8)	39.2 (23.0)	52.3 (22.6)	6.3 (6.2)	4.5	5.1
	3	0.0 (1.2)	1.4 (5.4)	40.9 (26.0)	54.3 (24.7)	3.4 (2.7)	4.2	4.6
	CRM	0.0 (1.2)	1.2 (5.1)	26.4 (17.8)	56.3 (23.9)	16.1 (12.0)	4.2	5.8
		6%	12%	18%	24%	30%		
5-A	1	0.0 (1.1)	0.2 (3.0)	7.1 (9.6)	36.6 (18.6)	56.2 (27.8)	3.1	3.1
	2	0.0 (1.1)	0.1 (2.3)	5.9 (8.5)	37.2 (19.7)	56.8 (28.5)	3.0	3.0
	3	0.0 (1.1)	0.1 (2.3)	5.7 (8.5)	39.1 (21.9)	55.1 (26.2)	3.0	3.0
	CRM	0.0 (1.1)	0.3 (3.5)	7.7 (9.7)	34.5 (17.9)	57.5 (27.8)	3.0	3.0
5-B	1	0.0 (1.1)	0.2 (3.4)	10.6 (11.1)	42.1 (19.9)	47.0 (24.5)	3.8	3.8
	2	0.0 (1.2)	0.4 (4.1)	12.5 (12.6)	44.4 (21.0)	42.8 (21.1)	4.2	4.2
	3	0.0 (1.1)	0.3 (3.3)	11.7 (13.9)	52.6 (26.9)	35.4 (14.8)	4.6	4.6
	CRM	0.0 (1.1)	0.2 (3.4)	7.3 (9.7)	34.1 (17.7)	58.4 (28.1)	3.0	3.0

DLT: dose-limiting toxicity; MTD: maximum tolerated dose; CRM: continual reassessment method.

Table 4

Example of a single simulated trial of 30 patients and 5 dose levels, using Model 3 and scenario 4-A. The true DLT rates are (7%, 14%, 21%, 30%, 40%) and the target is 30%, so dose 4 is the true MTD. Our initial guesses for the DLT rates are (10%, 20%, 30%, 40%, 50%), implying that initially we suspect dose 3 is the MTD. However, our trials always start the first patient on dose 1.

Patient	Dose Given	Total Events	DLTs	Estimated DLT Rates (%)					$\hat{M}TD$
				Dose 1	2	3	4	5	
1	1	1	0	10	21	32	44	55	2
2	2	0	0	7	16	25	36	47	3
3	3	4	0	8	16	25	37	51	3
4	3	3	0	8	16	25	37	50	3
5	3	5	1	11	20	31	45	59	3
6	3	1	0	9	18	28	41	55	3
7	3	0	0	8	16	25	37	51	3
8	3	0	0	7	14	23	34	47	4
9	4	6	1	9	16	26	37	52	3
10	3	2	0	8	16	25	36	50	3
11	3	1	0	8	15	23	34	48	4
12	4	3	0	8	14	22	32	45	4
13	4	1	0	7	13	21	30	42	4
14	4	0	0	7	12	19	28	40	4
15	4	2	1	7	13	21	30	43	4
16	4	2	0	7	13	20	29	41	4
17	4	1	0	7	12	19	28	39	4
18	4	5	1	7	13	20	30	43	4
19	4	1	0	7	12	19	29	41	4
20	4	0	0	6	12	19	27	39	4
21	4	4	0	7	12	18	27	38	4
22	4	2	0	6	11	18	26	37	4
23	4	6	0	7	12	18	26	36	4
24	4	2	0	7	11	17	25	35	4
25	4	2	1	7	12	18	26	37	4

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Patient	Dose Given	Total Events	DLTs	Estimated DLT Rates (%)					MTD
				Dose 1	2	3	4	5	
26	4	0	0	6	11	17	25	36	4
27	4	2	0	6	11	17	25	35	5
28	5	6	2	7	13	19	28	40	4
29	4	2	0	7	12	19	28	39	4
30	4	2	0	7	12	18	27	38	4

DLT: dose-limiting toxicity; MTD: maximum tolerated dose.

Simulation study results, investigating sensitivity of model performance to model parameters. Each line summarizes 10,000 simulated adaptive trials of 30 patients with 5 dose levels. The target DLT rate is 30% and the true DLT rates are (20%, 30%, 40%, 50%, 60%), so dose 2 is the true MTD. Results should be compared with the scenario 2-A results in Table 2.

Table 5

Scenario	Model	Dose Selection Percentage (Number of Patients Treated)					Score ₁	Score ₂
		1	2	3	4	5		
		0.2	0.3	0.4	0.5	0.6		
Larger Variance	1	18.8 (7.0)	56.2 (13.8)	22.7 (6.8)	2.2 (1.8)	0.0 (0.6)	4.6	7.3
	2	24.6 (8.9)	53.7 (13.1)	19.7 (6.0)	1.9 (1.5)	0.0 (0.4)	4.8	7.2
	3	22.9 (8.1)	59.9 (16.0)	16.7 (5.5)	0.6 (0.4)	0.0 (0.0)	4.1	5.9
	CRM	20.3 (7.5)	51.9 (12.9)	24.2 (6.8)	3.5 (2.1)	0.2 (0.7)	5.2	8.4
Smaller Variance	1	7.3 (2.6)	67.3 (16.8)	23.8 (8.5)	1.6 (1.8)	0.0 (0.3)	3.4	6.1
	2	17.9 (5.5)	66.1 (17.9)	15.3 (5.9)	0.6 (0.6)	0.0 (0.0)	3.5	5.1
	3	2.6 (1.7)	67.7 (16.7)	29.2 (11.1)	0.4 (0.5)	0.0 (0.0)	3.3	6.3
	CRM	5.9 (2.2)	62.7 (15.9)	28.6 (9.5)	2.7 (2.1)	0.1 (0.3)	4.0	7.4
K = 7 Event Types	1	12.8 (4.7)	61.0 (15.3)	23.7 (7.6)	2.4 (1.8)	0.1 (0.5)	4.2	7.0
	2	17.8 (6.4)	59.2 (15.0)	21.2 (6.9)	1.7 (1.5)	0.1 (0.3)	4.3	6.7
	3	12.4 (4.9)	60.1 (15.2)	25.7 (8.5)	1.8 (1.4)	0.0 (0.1)	4.2	7.1
	CRM	14.4 (4.9)	57.5 (14.8)	24.9 (7.6)	3.0 (2.0)	0.1 (0.6)	4.6	7.7

DLT: dose-limiting toxicity; MTD: maximum tolerated dose; CRM: continual reassessment method.