Check for updates

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2018, Vol. 27(11) 3447–3459
© The Author(s) 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280217703812
journals.sagepub.com/home/smm
SAGE

Hierarchical models for sharing information across populations in phase I dose-escalation studies

Kristen M Cunanan[1] and Joseph S Koopmeiners[2]

Abstract

The primary goal of a phase I clinical trial in oncology is to evaluate the safety of a novel treatment and identify the maximum tolerated dose, defined as the maximum dose with a toxicity rate below some pre-specified threshold. Researchers are often interested in evaluating the performance of a novel treatment in multiple patient populations, which may require multiple phase I trials if the treatment is to be used with background standard-of-care that varies by population. An alternate approach is to run parallel trials but combine the data through a hierarchical model that allows for a different maximum tolerated dose in each population but shares information across populations to achieve a more accurate estimate of the maximum tolerated dose. In this manuscript, we discuss hierarchical extensions of three commonly used models for the dose–toxicity relationship in phase I oncology trials. We then propose three dose-finding guidelines for phase I oncology trials using hierarchical modeling. The proposed guidelines allow us to fully realize the benefits of hierarchical modeling while achieving a similar toxicity profile to standard phase I designs. Finally, we evaluate the operating characteristics of a phase I clinical trial using the proposed hierarchical models and dose-finding guidelines by simulation. Our simulation results suggest that incorporating hierarchical modeling in phase I dose-escalation studies will increase the probability of correctly identifying the maximum tolerated dose and the number of patients treated at the maximum tolerated dose, while decreasing the rate of dose-limiting toxicities and number of patients treated above the maximum tolerated dose, in most cases.

Keywords

Concurrent studies, continual reassessment method, hierarchical modeling, multiple cancer populations, phase I

1 Introduction

The primary goal of a phase I clinical trial in oncology is to evaluate the safety of a novel treatment and identify the maximum tolerated dose (MTD), defined as the maximum dose with a toxicity rate below some pre-specified threshold. The primary outcome in these trials is a binary indicator for the presence or absence of a dose-limiting toxicity (DLT). A DLT is an adverse event, or toxic side-effect, that is severe enough to prevent increasing dosage. In phase I oncology trials, adverse events are defined and their severity classified using the common terminology criteria for adverse events.

Phase I oncology trials typically take the form of dose-escalation studies, where initial subjects are treated at the lowest dose level and subsequent subjects are treated at progressively higher doses until the MTD is identified. A wide array of phase I dose-escalation designs have been proposed in the literature.[1–5] Historically, the most commonly used design is the 3 + 3.[1] The 3 + 3 is a simple, algorithmic design that bases dose escalation or de-escalation on the presence or absence of DLTs in the previous cohort. An alternate approach is to specify a fully parametric model for the dose–toxicity relationship and allow dose escalation or de-escalation to depend on the current estimate of the MTD. An advantage of this approach is that escalation or de-escalation is based on all available data, unlike the traditional 3 + 3 design which only uses the outcomes from the previous cohort.

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA

Corresponding author:
Kristen M Cunanan, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA.
Email: kristenmay206@gmail.com

The most commonly used model-based design is the continual reassessment method (CRM).[2] In the CRM, the dose–toxicity curve is represented by a simple, parametric model, such as the logistic, power, or hyperbolic tangent model, and subjects are treated at the current estimate of the MTD based on all available data. The MTD is defined as the dose level with estimated probability of DLT closest to the pre-specified target probability of toxicity at trial completion. Modifications and extensions to the CRM have been proposed to better ensure patient safety and improve estimation.[3,5–9]

1.1 Hierarchical modeling in phase I oncology trials

Researchers are often interested in evaluating the performance of a novel treatment in multiple patient populations. In this case, investigators may be required to complete multiple phase I trials to determine the MTD for each population if the novel treatment is to be used in combination with background standard-of-care that differs by population. For example, the results have recently been reported for phase I clinical trials of Veliparib, a novel PARP inhibitor, in combination with cyclophosphamide for patients with solid tumors and lymphomas,[10] whole abdominal radiation for patients with advanced solid malignancies and peritoneal carcinomas,[11] whole brain radiation for patients with brain metastases[12] and cisplatin and etoposide in patients with small cell lung cancer,[13] each trial resulting in a different estimated MTD.

Completing separate phase I trials for each patient population is expensive and time-consuming, while collapsing across populations into a single trial would not be scientifically justified. Furthermore, while it is reasonable to assume that the MTD will vary by population, it is also likely that the results of a phase I trial in one population would provide information about the MTD in the other populations. This motivates a hierarchical modeling (HM) approach that allows each patient population to have a separate MTD but shares information across populations during dose finding and identifying the population-specific MTD.

HM is a widely used statistical approach for sharing information across populations that has been applied to the analysis of clinical trials in a number of settings, including meta-analysis, multi-center trials, multiple comparisons, variable selection, and subgroup analysis.[14] While HM has been used extensively in phase III clinical trials, the statistical literature relating to HM in early phase clinical trials (i.e. phases I and II) is limited. In phase I dose-finding trials, HM has been used to pool information across pharmacokinetic data from healthy volunteers,[15] to improve estimation of the probability of a DLT for combinations of doses of two therapeutic agents[16] and for bridging multiple phase I clinical trials.[17] HM has been used more extensively in phase II clinical trials as an approach to properly model treatment response in the presence of disease subpopulations[18,19] and to identify personalized cancer treatments in genetically defined subgroups, most notably in the BATTLE trial.[20]

In this manuscript, we discuss HM in the context of phase I oncology trials. HM has been used in phase II and III clinical trials to facilitate the borrowing of information across patient populations. Applying HM to phase I oncology trials will allow the borrowing of information across populations but provide flexibility by specifying a different MTD for each population. In addition, we also propose novel dose-finding guidelines (DFGs) for phase I clinical trials using HM. Standard dose-finding approaches that escalate in cohorts of three will not fully utilize the advantage of sharing information across populations, while dose adaptation after every patient will be too aggressive and jeopardize patient safety. The DFGs proposed in Section 3 allow us to fully realize the potential of HM in phase I clinical trials, while achieving safety profiles that are similar to standard phase I trial designs.

The remaining sections of this manuscript proceed as follows. First, we discuss hierarchical extensions of three commonly used dose–toxicity models for the CRM in Section 2. In Section 3, we propose DFGs and present our dose-finding algorithm for phase I oncology trials using HM. In Section 4, we present simulation results evaluating the operating characteristics of the proposed design. Finally, we conclude with a brief discussion of our findings and the potential for implementing the proposed methods in practice in Section 5.

2 Dose–toxicity models

In this section, we discuss hierarchical extensions of three commonly used dose–toxicity models in phase I clinical trials. In each case, we define a two-level, Bayesian hierarchical model with population-specific effects that allow borrowing across populations. The following notation will be used throughout this section. Let, Y_{ikj} be a binary indicator for a DLT in patient $i = 1, \ldots, n_{kj}$, in population $k = 1, \ldots, K$, at dose level $j = 1, \ldots, D$. Denote $\pi_{kj}(i) = Pr(Y_{ikj} = 1 | Dose = j, Population = k)$ as the probability of a DLT for patient i in population k treated at dose level j. Throughout the remainder of this section, we drop the notation for patient (i) for simplicity.

2.1 Power model

The first model we consider is an extension of the widely used power model,[21] which is a one-parameter dose–toxicity model that is commonly used in phase I oncology trials. We define our hierarchical power model as follows

$$\begin{aligned} &\pi_{kj} = p_j^{exp(\alpha_k)} \\ &\alpha_k \sim N(A, \sigma^2) \\ &A \sim N(0, 2^2) \quad \log(\sigma) \sim Unif(-1, 1) \end{aligned} \tag{1}$$

Here, $(p_1, \ldots, p_D)$ is a monotonically increasing skeleton that is constant across populations and specified in advance, α_k is the power parameter for the kth population, A represents the shared mean and σ^2 controls the degree of borrowing across populations. From our prior specification, 99% of A's prior mass falls between -6 and 6, which allows the probability of a DLT to range from 0.01 to 0.99 regardless of p_j. This is a commonly used prior specification in traditional dose-finding studies.[5,9] The uniform distribution assumed for $log(\sigma)$ is defined on $(-1, 1)$ with a prior mean of 0.[22] Consequently, σ^2 is defined on (0.14, 7.4). This parameter controls the degree of borrowing across populations. When σ^2 is small, our model suggests homogeneity across populations and encourages borrowing. In contrast, large values for σ^2 suggest heterogeneity across populations, resulting in less borrowing. Finally, we note that the power model has been shown to be equivalent to the hyperbolic-tangent model with a different dose transformation.[21,23]

2.2 Logistic regression and curve-free models

Another popular one-parameter model is a logistic regression model[6] with fixed intercept, $logit(\pi_{kj}) = -3 + \beta_k \times q_j$; $\beta_k \sim N(B, \tau^2)$; $B \sim N(1, 2^2)$; $\log(\tau) \sim Unif(-1, 1)$. Here, β_k is the slope parameter for the kth population, B is the shared mean slope, and τ^2 controls the degree of borrowing across populations. Originally, we also considered a two-parameter logistic regression model that allowed each population's dose–toxicity model to have a random intercept and slope. We found that this model is overly complex for the small sample sizes found in phase I clinical trials and sensitive to prior input values. Furthermore, it did not improve operating characteristics and it has been shown that the one-parameter logistic model has better performance than the two-parameter logistic model in phase I clinical trials.[23] Therefore, the two-parameter logistic regression model was not given further consideration.

An alternate approach is a curve-free method that directly models the probability of DLT at each dose level while imposing a monotonicity constraint on the relationship between dose and the probability of DLT without specifying a formal parametric structure for the dose–toxicity relationship.[24] Gasparini and Eisele[24] reparameterized the probability of DLT at each dose level as follows: $\{\theta_1 = 1 - \pi_1, \theta_2 = (1 - \pi_2)/(1 - \pi_1), \ldots, \theta_D = (1 - \pi_D)/(1 - \pi_{D-1})\}$. We specify a hierarchical model for $\gamma_{kj} = logit(\theta_{kj})$ as $\gamma_{kj} \sim N(\Gamma_j, \nu_j^2)$; $\Gamma_j \sim N(c_j, 3^2)$; $\log(\nu_j) \sim Unif(-1, 1)$. Here, γ_{kj} is the unrestricted model parameter for population k and dose j, Γ_j is the shared population mean for dose j and ν_j^2 controls the amount of borrowing across populations for each dose level j, $j = 1, \ldots, D$.

The hyper-parameters (i.e. q_j and c_j for $j = 1, \ldots, D$) can be specified such that they induce values of π_{kj} equal to the prior skeleton used in Section 2.1. The prior distributions discussed above represent the final priors used to fit these models but other prior distributions, particularly for the hierarchical variance parameter, were also considered. Specifically, we also considered the conditionally conjugate inverse-Gamma prior for the variance and a Uniform(0,b) prior on the standard deviation. However, the former is sensitive to prior input values when the estimated standard deviation is small, which can occur early on in the trial or for homogeneous populations.[22] A Uniform(0,b) prior on the standard deviation places more density on more extreme prior σ^2 values than the Uniform(−a,a) prior on the log standard deviation. As a result, over-borrowing due to underestimation of σ^2 is a concern with the small sample sizes found in phase I clinical trials.

3 Dose-finding algorithm

In this section, we discuss dose finding when using HM to share information across populations in phase I oncology trials. We expect that enrollment will be staggered and randomly distributed across patient populations with several consecutive patients enrolled in one population, while long stretches may occur without enrolling patients in others. As a result, extending standard phase I dose-finding algorithms to our case

is not trivial. The traditional 3 + 3 rule-based design requires dose escalation or de-escalation be done after each cohort of three patients. Typically, phase I dose-escalation studies that use model-based designs such as the CRM use cohorts of three patients based on recommendations by Goodman et al.[6] and the MTD is re-evaluated for each new cohort, other dosing-cohort sizes of two or four patients are also an option and have been evaluated in the literature. There are two natural extensions of this approach but neither is satisfactory. First, we could re-evaluate the MTD after each cohort of three patients, regardless of patient population. This approach would be too aggressive and could result in a patient being treated at a higher dose level before lower dose levels have been tried in that patient's population. The second option would be to use cohorts of three patients within a patient population and only re-evaluate the MTD within a patient population when a new cohort is ready to enroll. This approach would be too conservative, failing to take full advantage of HM and treat too many patients at sub-therapeutic dose levels. Therefore, we propose three DFGs and compare the performance of each through simulation.

Throughout the rest of this section, we will define the MTD and terminate the trial for excess toxicity as follows. Let $\bar{\pi}_T$ be a pre-defined target toxicity rate. Dose level j in population k is considered to have acceptable toxicity if

$$Pr(\pi_{kj} < \bar{\pi}_T | Data, Dose) > \xi \tag{2}$$

This is a commonly used criterion in phase I oncology trials and is typically chosen between 0.05 and 0.20. ξ can be thought of as a tuning parameter, which is chosen to achieve the desired operating characteristics for the trial.[14] We have defined $\bar{\pi}_T$ and ξ to be constant across populations but these could be made population specific, if desired. In the context of the DFGs described below, the MTD for a population is defined as the dose level that minimizes the absolute difference between the probability of a DLT and $\bar{\pi}_T$ from among the set of doses with acceptable toxicity. Dose finding for a population terminates if the first dose level has unacceptable toxicity by criterion (2).

3.1 DFGs

We now describe three DFGs for phase I clinical trials with multiple patient populations. In each DFG, initial patients within each population start at the lowest dose level. Escalation *does not* proceed in cohorts, as in a standard phase I dose-escalation study but, rather the DFGs define when it is acceptable to escalate within a population, at which point dose assignments will depend on the estimated population-specific MTD using the hierarchical models described in Section 2.

The first approach we consider is the "m" DFG, which allows escalation within a population when at least 1 patient within the current population and at least m patients have been treated across all populations. Formally, the "m" DFG allows the kth population to potentially escalate to dose level $j + 1$ if:

- **m** patients overall (and at least 1 patient in population k) have been treated at dose level j

We note that a special case of the "m" DFG occurs when all m patients previously treated at dose level j are in population k, which would allow escalation and is consistent with the standard phase I design that escalates in pre-specified cohorts of size m. For example, if $m = 3$ and assuming that the fourth enrolled patient was from population k, the "m" DFG allows the fourth patient to potentially escalate to the second dose level as long as one of the first three patients was in population k. This suggests observing m patients within a population is equivalent to observing m patients overall in estimating each population's dose–response curve.

It is possible that the "m" DFG might be too aggressive and result in an unacceptably high number of toxicities. Therefore, we consider two other DFGs with further restrictions to protect patient safety. First, we consider the "$m(m+1)$" DFG. The "$m(m+1)$" DFG allows possible escalation to dose level $j + 1$ for population $k = 1, \ldots, K$ if:

- **m** patients in population k have been treated at dose level j
- Or **m + 1** patients overall (and at least 1 patient in population k) have been treated at dose level j

This suggests observing m patients within a population is equivalent to observing $m + 1$ patients overall in informing our dose–response models.

Finally, we propose a third DFG, which we refer to as the "321" DFG, which is similar to the "$m(m+1)$" DFG but puts additional restrictions on escalation to protect patient safety and promote reasonable sharing across

populations. The "321" DFG allows potential escalation to dose level $j + 1$ for population $k = 1, \ldots, K$ if:

- **Three** patients in population k have been treated at dose level j
- OR **two** patients in population k and at least two patients not in population k have been treated at dose level j
- OR **one** patient in population k and at least one patient in three other populations have been treated at dose level j

The "321" DFG is the most restrictive of the three DFGs. Examples of scenarios where the "$m(m + 1)$" DFG would allow possible escalation to dose level $j + 1$ but the "321" would not include: only one patient in population k and three patients in population k_2 have been treated at dose level j; and, only one patient in population k, two patients in population k_2 and one patient in population k_3 have been treated at dose level j. The "321" DFG restricts the scenarios where escalation is allowed after only one patient in a population has been treated at the current dose level to reduce the influence of other populations' estimated dose–response curves, should they be different. While this restriction slows dose finding by requiring more patients to enroll, it is incorporated for patient safety.

Online supplemental Web [Figures 1 through 4] present a single simulated trial using the three DFGs described above, along with a trial with no restrictions on escalation other than untried dose levels cannot be skipped when escalating within a population (denoted as no DFG). We note that the DFGs defined above indicate when it is *potentially* acceptable to escalate to an untried dose within a population but the ultimate decision to escalate will be based on the current estimate of the population-specific MTD using all available data. In some sense, the DFGs define a run-in period for each population and dose level to prevent escalation without sufficient evidence that the current dose level is safe within each population.

In summary, our dose-finding algorithm for identifying the MTD_k in a phase I clinical trial using HM proceeds as follows:

(1) Treat the first patient within each population at the lowest dose level.
(2) When a new patient is enrolled, update the posterior distribution of the probability of toxicity for all dose levels and populations using all available data.
(3) Identify the set of acceptable doses for each population using criterion (2).
(4) If all doses are unacceptably toxic and at least three patients have been treated in the current population, then the trial terminates for that population. For each population, if all doses are unacceptably toxic but less than three patients have been treated in that population, then the current patient is treated at dose level 1
(5) Otherwise, treat the next patient at the dose level that minimizes $|E(\pi_{kj}|Data, Population = k, Dose = j) - \bar{\pi}_T|$ from among the acceptable dose levels, under the restriction that escalation is only allowed if the criteria for escalation is satisfied using the pre-specified DFG and escalating more than one dose level within a population at a time is not allowed
(6) Repeat steps 2–5 until the maximum overall sample size is reached. Within each population, the acceptable dose that minimizes $|E(\pi_{kj}|Data, Population, Dose) - \bar{\pi}_T|$ at study completion is considered the MTD_k for $k = 1, \ldots, K$

We note that in Step 4, only the recently enrolled and treated population can terminate for excess toxicity (i.e. a population cannot terminate based on the outcome of another population). In addition, we do not force balance across populations by specifying a maximum number of subjects per population. Enforcing balance would likely improve the operating characteristics of the trial but would also dramatically increase the duration of the trial. Our dose-finding algorithm proposes to update the model after every patient. We also explored an approach where the model is only updated, when the enrolled population is permitted to escalate. The only practical difference between the two approaches is that we allow de-escalation regardless of whether or not a population is allowed to escalate according to the DFG. We feel that this is appropriate because it is preferable to always treat patients at the current estimate of the MTD and only place restrictions on escalation to protect patient safety, which is not a concern when de-escalating. The operating characteristics of this design were very similar to the proposed algorithm.

4 Simulation study

We conducted a simulation study to evaluate the operating characteristics of a phase I clinical trial using the hierarchical models and DFGs described in Sections 2 and 3. We evaluate the performance of each method based

on (i) the probability of correctly identifying the population-specific MTD, (ii) the percent of patients experiencing a DLT, (iii) the percent of patients treated at the population-specific MTD and (iv) the percent of patients treated above the population-specific MTD. Trial parameters were set as follows. We assume a maximum of 45 patients across $K = 5$ populations and assume an equal probability of enrollment to each population (i.e. an average of nine patients/population). The target toxicity rate was set to $\bar{\pi}_T = 0.3$. The threshold for determining if a dose has an acceptable probability of toxicity was set equal to $\xi = 0.2$. We consider $D = 4$ dose levels with dose index $\{1, 2, 3, 4\}$. All simulations were completed in R version 3.1.1. Gibbs and slice sampling were completed in JAGS via R using *rjags*.[25] Posterior inference was completed using 10,000 MCMC samples following a period of 5000 iterations for burn-in; 1000 simulated trials were completed for each scenario.

The skeleton for the power model $(p_1, \ldots, p_4)$ was set equal to $(0.1, 0.2, 0.35, 0.50)$. The scaled dose levels for the logistic regression model $(q_1, \ldots, q_4)$ were set equal to $(logit(0.1) + 3, logit(0.2) + 3, logit(0.35) + 3, logit(0.5) + 3)$ to achieve a probability of toxicity equal to the power model skeleton assuming a slope equal to the prior mean of 1. We similarly set the hyper parameters for the curve-free method $(c_1, \ldots, c_4)$ equal to $(1, 1.5, 2, 2.5)$ which would induce $(\pi_{k1}, \ldots, \pi_{k4})$ equal to the power model skeleton.

In addition to simulating a phase I clinical trial using the models and DFGs described in Sections 2 and 3, for comparison, we also evaluated the operating characteristics of three other designs. First, we considered a phase I clinical trial using HM that treats each patient at the current estimate of the MTD with no restrictions on dose escalation other than untried dose levels within a population cannot be skipped when escalating (no DFG). The DFGs described in Section 3 were proposed, primarily, to protect patient safety and a comparison to a design with no restrictions on dose finding will allow us to isolate the impact of the DFGs on the various operating characteristics of the trial. In addition, we simulated two types of phase I designs that did not use HM and instead fit independent models for each population. For these two designs, we fit the models specified in Section 2 without the second level of the hierarchy and specified a $N(0, 2^2)$ prior for α_k, a $N(1, 2^2)$ for β_k, and $N(c_j, 3^2)$ for the γ_{kj} with c_j as specified in the previous paragraph. For the independent models, we considered a design with a maximum sample size of 45 patients and a design with a maximum sample size of 90 patients, both with dose adaptation occurring after every patient. This will allow us to evaluate the benefit of using HM, as compared to completing designs that treat each population as independent data.

4.1 Scenarios

We simulated data from the six scenarios presented in Figure 1. The true dose–response curves were set by specifying the slope and MTD_k in a logistic regression model, with the exception of Group 5 in Scenario 6, which is taken to be qualitatively different from the other groups. In Scenario 1, all five groups have the same dose–toxicity curve and the optimal design would be to collapse the five groups and complete a single trial assuming a common dose–toxicity curve for all groups. In Scenario 2, all five groups have the same MTD but different dose–toxicity curves. Scenarios 3 and 4 represent scenarios where the dose–toxicity curves vary by population with true population-specific MTDs ranging from dose 1 to dose 4 in Scenario 3 and dose 1 to dose 3 in Scenario 4. Scenario 4 was suggested to us by a researcher in the field as a particularly difficult case that we would expect to see in practice. In Scenario 5, all doses are unacceptably toxic for Group 1, while dose 1 is the MTD for the Groups 2 and 3 and dose 2 is the MTD for Groups 4 and 5. Finally, in Scenario 6, all doses are unacceptably toxic for Group 1, while Groups 2 to 5 have a different population-specific MTD.

4.2 Results

In this section, we present results using a power hierarchical model comparing the "no DFG" design, independence design, and design implementing the "321" DFG. Results for the other models and proposed DFGs are presented in the Supplemental Materials; and we note differences are observed when varying enrollment rates. Figure 2 presents the probability of correctly identifying the population-specific MTD for Scenarios 1 through 6 using the hierarchical power model. In the upper corner of each plot, we present the average probability of correctly identifying the true MTD_k across all $k = 1, \ldots, K$. In Scenario 1, all populations have the same dose–toxicity curve, while in Scenario 2, all populations have the same MTD but different dose–toxicity curves. We see that in both scenarios the HM approach results in a higher probability of correctly identifying the population-specific MTD than the independence design with a maximum sample size of 45 subjects regardless of the DFG. We note that this is a relatively high bar in that the independence design with 45 patients allows adaptation after every subject and does not include the typical restrictions (such as cohorts

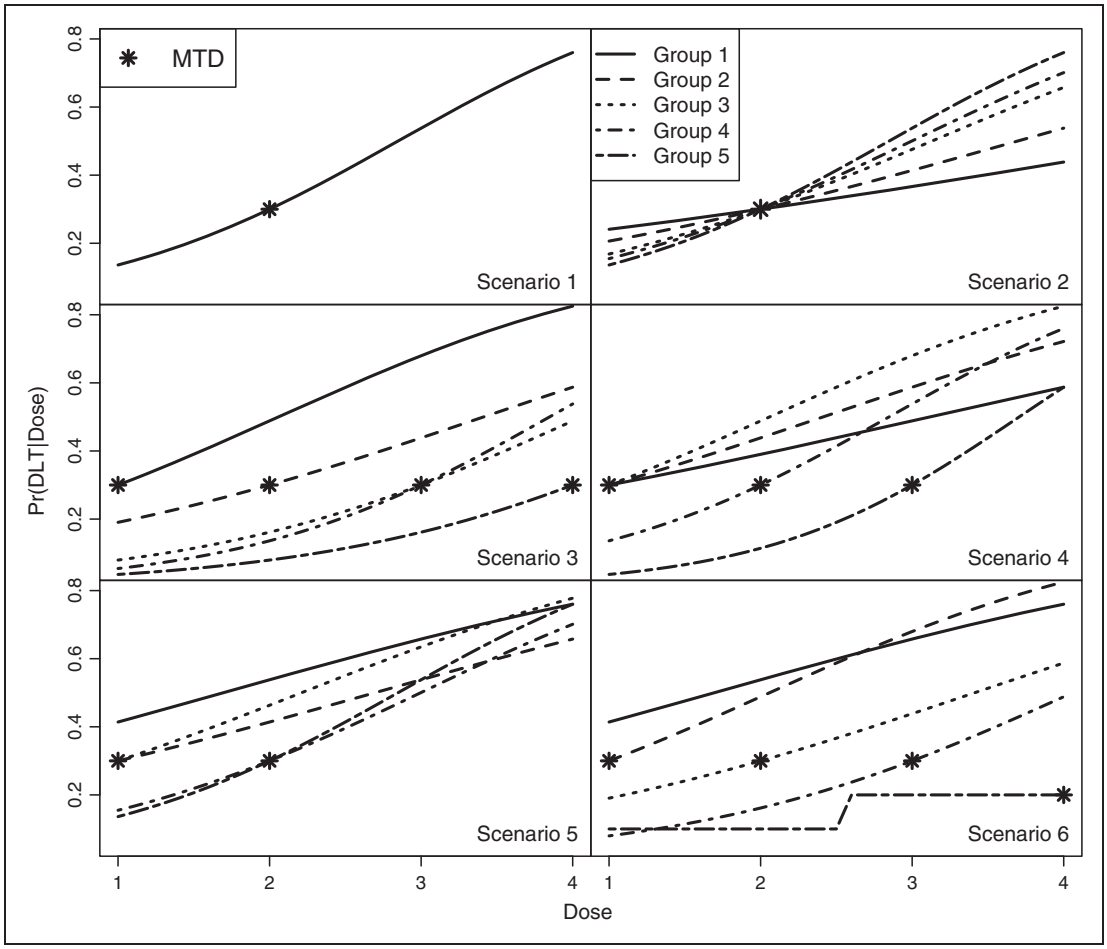

Figure 1. Scenario 1 (top, left): all populations' dose–response curve are equivalent. Scenario 2 (top, right): all populations have the same MTD_k level, but the slope increases as population index increases. Scenario 3 (middle, right): the MTD_k for each population is dispersed across all four dose levels for the five populations. Scenario 4 (bottom, left): the first three populations' $MTD_{1,2,3}$ is dose level 1; the last two populations have $MTD_{4,5}$ at dose levels 2 and 3, respectively. Scenario 5 (middle, left): Group 1 terminates trial; the true $MTD_{2,3}$ for Groups 2 and 3 is dose 1; the true $MTD_{4,5}$ is dose 2. Scenario 6 (bottom, left): similar to Scenario 3, except Group 1 terminates trial. The MTD_k for population $k = 1, \ldots, K$ is identified with an asterisk.

of 3, etc.) that are put in place to protect patient safety but may decrease the probability of correctly identifying the population-specific MTD. In addition, the hierarchical power model does as well (better in some cases), than the independence design with 90 patients. This indicates that in Scenarios 1 and 2 HM was as valuable as doubling the maximum sample size from 45 to 90 patients in the independence designs. We see similar trends for the average percent correct selection across all populations. In Scenarios 3 through 6, where the population-specific MTD varies across populations, the hierarchical approach improves upon the independence design with a maximum sample size of 45 in most cases, with the exception of Group 1 in Scenarios 5 and 6 and Group 5 in Scenarios 3 and 6. These are scenarios where either all of the doses were unacceptably toxic or where dose 4 was the population-specific MTD, suggesting that the hierarchical model was incorrectly shrinking the estimated MTD_k towards intermediate doses due to the results of the other populations. When the population-specific MTDs vary, we see the average percent correct selection across all populations is similar for the "321" DFG and "no DFG" and better than the independence design with a maximum sample size of 45, except when the populations' true MTDs are extremely heterogeneous.

We note that the three hierarchical DFGs correctly identified the MTD_k at a similar rate as the hierarchical model with "no DFG". The goal of implementing the DFGs was to protect patient safety and it is encouraging to see that the three DFGs did not adversely impact the probability of correctly identifying the MTD_k. We found the proposed DFGs, specifically the "321" DFG, were more conservative in escalation. Depending on the scenario, 35–60% of patients' dose assignment was determined by the "321" DFG, rather than by minimizing

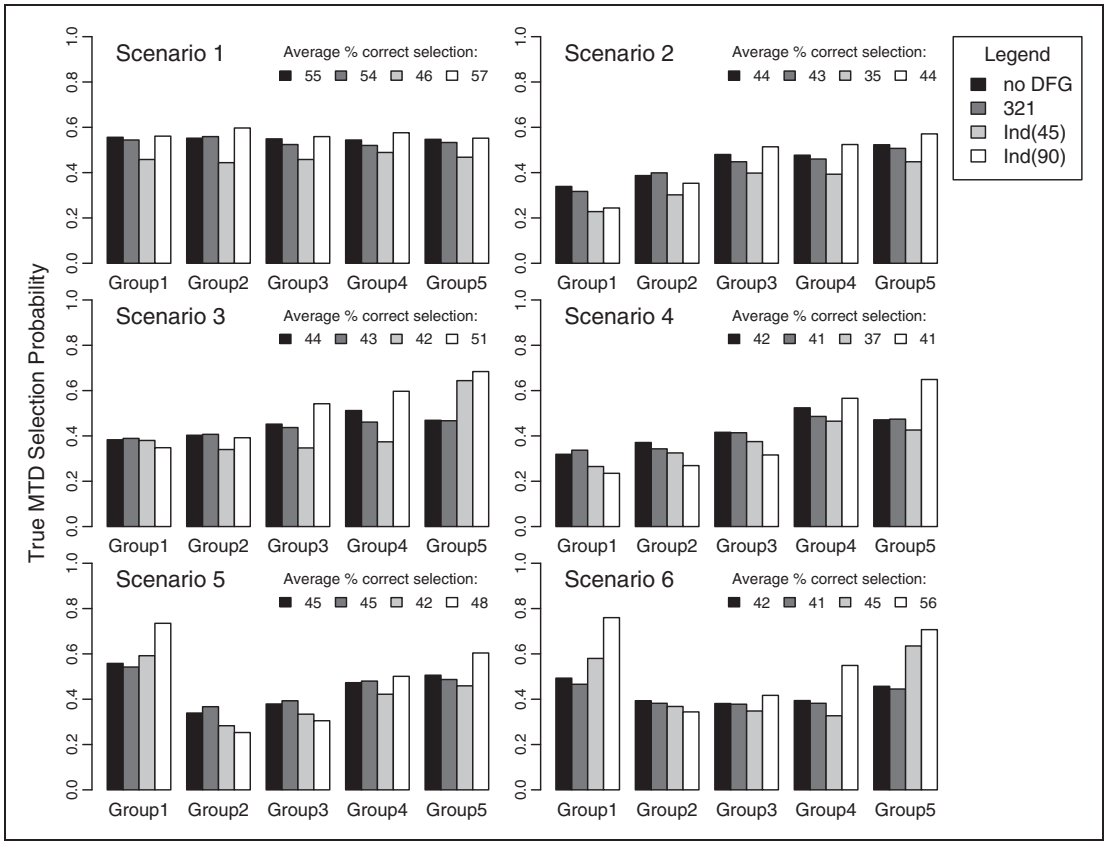

Figure 2. Power Model: probability of correctly identifying the true MTD_k for Scenarios 1 to 6; upper right plot corner: average probability of correctly identifying true MTD_k across all k. Results from 1000 simulated trials are displayed for "no DFG," the proposed "321" DFG (both with max 45 patients), and K independent models assuming different maximum sample sizes (45 and 90 patients overall), displayed in parentheses.

$|E(\pi_{kj}|Data, Population, Dose) - \bar{\pi}_T|$ in the dose-finding algorithm (see Online supplemental Web [Table 1]). The "321" DFG treats more patients below the true MTD_k but fewer patients above the true MTD_k compared to "no DFG." These results become even more dramatic when the populations have different MTDs and enrollment rates.

Our initial results indicate that HM increases the probability of correctly identifying the population-specific MTD_k but a potential concern related to the implementation of this approach is that HM would increase the number of DLTs due to sharing information across populations. Online supplemental Web [Figure 5] displays the percent of patients experiencing a DLT for Scenarios 1 through 6 using the hierarchical power model. Our results suggest that HM does not dramatically increase the rate of DLT and in fact, results in a decreased probability of DLT relative to the independence designs, in most cases, with the only exceptions occurring when the lowest dose is the population-specific MTD for one population but the population-specific MTD is higher for other populations. As a result, HM shrinks the estimated MTD_k towards intermediate dose levels due to the results for the other populations. However, even in these cases, the increase in the DLT rate is minor. In addition, we note that our results present the percent of patients experiencing a DLT, rather than the absolute number of DLTs, which implies that the independence design with a maximum sample size of 90 subjects will have a much larger total number of DLTs than the HM designs. Therefore, while the independence design with 90 patients is more likely to accurately identify the population-specific MTD, in some cases, this comes at the expense of a dramatically higher number of DLTs. Finally, while the differences are subtle, we note that the "321" DFG results in less DLTs than the "$m(m+1)$" DFG, which results in less DLTs than the "m" DFG, as expected.

Online supplemental Web [Figures 6 and 7] present the average number of patients treated at the MTD_k and the average number of patients treated above the MTD_k. Our results indicate that, in addition to increasing the probability of correctly identifying the MTD_k and decreasing the rate of DLTs, HM increases the average number of subjects treated at the MTD_k and decreases the number of patients treated at unsafe doses above the MTD_k, in

many cases. Again, exceptions to this rule occur when the lowest dose is the MTD_k or when all doses are excessively toxic. In this case, the independence design treats fewer patients above the true MTD_k due to its propensity to stop the trial early and declare all doses overly toxic. We note that results for the independence design with a maximum sample size of 90 subjects were not included in these results because the larger sample size skews our results with respect to the y-axis and because the larger design results in a substantial increase in DLTs compared to the HM designs.

Online supplemental Web [Figures 8 through 11] present results using the hierarchical logistic regression model. In summary, the logistic regression model performs well, when the populations exhibit homogeneity with regards to the MTD_k (Scenarios 1 and 2), identifying the MTD_k as often as the independence design with a maximum sample size of 90 subjects; however, performed poorly when the MTD_k varied by population, with particularly poor performance in Scenarios 5 and 6, identifying the MTD_k no more often, and in many cases less often, than the designs that assume independence. We observe similar trends for the average percent correct selection across all populations. We note that the logistic regression model is more conservative in exposing fewer patients to excessively toxic doses across most scenarios, except when no MTD exists among dose levels considered, and we found the logistic regression model to escalate slowly, exposing more patients to ineffective doses below the true MTD_k, and in some cases less patients at the true MTD_k than the independence designs.

Finally, the results for the hierarchical curve-free model can be found in Online supplemental Web [Figure 12] through 15. This model is non-parametric and thus richly parameterized compared to the other two models. In Scenarios 1 and 2, where the populations are homogeneous and the model borrows strength across groups, the added flexibility of the curve-free method results in a very high probability of correctly identifying the MTD_k, outperforming the two independence designs and we see a very high average percent correct selection across all k. In contrast, when the populations are heterogeneous and there is little borrowing, the model is over-parameterized and is not able to accurately estimate each MTD_k; here, we see the average percent correct selection for the independence designs are better than both the "no DFG" and "321" DFG.

In summary, our simulation results suggest that implementing HM in phase I oncology trials increases the probability of correctly identifying the MTD_k by borrowing information across populations. In addition, HM increases the number of patients treated at the MTD_k, decreases the percent of patients treated at unsafe dose levels above the MTD_k and decreases the number of toxicities, in most cases. All three models borrowed strength when the MTD_k was constant across populations, resulting in a more precise estimate of the MTD_k, but the hierarchical power model was more flexible and exhibited better performance when heterogeneity existed across populations. In addition, all three DFGs achieve the stated goal of increased patient safety by restricting dose escalation but the "321" exhibited the best performance with limited impact on the probability of correctly identifying the MTD_k.

4.2.1 Varying enrollment rates

In practice, it is important to consider and evaluate a design assuming different enrollment rates across the patient populations due to different prevalence rates. In a small simulation study, we investigate two extreme scenarios: (i) Group 1 enrolls one patient on average/month, Groups 2–3 enroll two patients on average/month, and Groups 4–5 enroll three patients on average/month; (ii) Groups 1–2 enroll three patients on average/month, Groups 3–4 enroll two patients on average/month, and Group 5 enrolls one patient on average/month.

Our results indicate that slow enrollment can limit our ability to correctly identify the population-specific MTD when the true MTD_k is at an extreme dose level; however, the results are otherwise robust to varying enrollment rates (see Online supplemental Web Figures 16–17). For (i), all designs had difficulty correctly declaring all doses excessively toxic, when this was the case; this is to be expected because the populations where all doses were excessively toxic were the populations with the slowest accrual. On the other hand, groups with faster enrollment rates displayed an increase in correctly declaring the higher dose levels as the MTD_k. Conversely for (ii), the slowest enrolling population had difficulty in correctly identifying the highest dose level as the MTD_k, but faster enrolling populations displayed an increase in correctly identifying no dose and/or the lowest dose level. We note the "no DFG" and "m" DFG display more aggressive behavior in dose escalation, resulting in more patients experiencing DLTs, compared to the "$m(m+1)$" and "321" DFGs.

4.3 Exploring other K

Initially, we assumed five populations to evaluate the operating characteristics of a phase I clinical trial using HM with the models and DFGs specified in Sections 2 and 3. We now present additional simulation results to evaluate

the impact of varying K and determine the minimum number of populations needed to observe a benefit from using HM. Simulation results are presented for $K=2$, 3, 4, and we only considered the hierarchical power model due to its superior performance in our initial simulation results. The "321" DFG was used for $K=4$ but is inappropriate for $K<4$ and results for $K=2$ and $K=3$ are presented for the "$m(m+1)$" DFG, instead. Simulations were completed using the prior distributions specified in Section 2. However, performance could potentially be improved by reducing the prior domain specified for σ. This is not unreasonable, since we have fewer populations and therefore do not need as large of a domain to motivate the appropriate amount smoothing. Finally, our simulation results assume a maximum sample size of 18, 27, 36 patients for K equal to 2, 3 and 4, respectively, to achieve an average sample size of 9 patients per population, similar to our results with $K=5$.

Simulation results for 1000 simulated trials per scenario are presented in Table 1. Presented are the probability of correctly identifying the MTD_k and the average probability of correctly identifying the MTD_k across all $k=1,\ldots,K$. The true dose–response curves for $K=2$, 3, 4 represent a subset of the dose–response curves used in the corresponding scenario for $K=5$ in Section 4.1. We note that population indices may change across $K=\{2,3,4\}$ but the $K=5$ population index is reported within each scenario and K. For comparison, we simulated an independence design with nine patients per population for each scenario and K. This design allowed dose escalation after each patient under the restriction that no untried dose levels be skipped when escalating, which we reiterate represents a high bar because this design would typically use cohorts of three patients (and a larger sample size), in practice.

We see that in general the probability of correctly identifying the MTD_k increases with K. Furthermore, we see that there is a clear advantage to HM with $K=4$ but that the probability of correctly identifying the MTD_k was lower with the HM design than with the independence design with $K=2$ due to the increased complexity of the hierarchical power model. With $K=3$, HM increased the probability of correctly identifying the MTD_k, in most cases, but performed particularly poorly in Scenario 6, where there was substantial heterogeneity in the MTD_k across populations. We see similar trends in the average probability of correctly identifying the MTD_k across all k. Based on these results, we recommend that HM in phase I clinical trials be implemented with a minimum of three populations but four populations are likely required to fully realize the advantages of HM. This is consistent with the results of Charles Stein,[26] who demonstrated that using HM improved upon the naive approach (of no borrowing) in that the mean squared error is reduced uniformly (regardless of the true values of the group means), when there are at least three groups.

Table 1. Results from 1000 simulated trials for the power model for different K.

Sc		K = 2			K = 3				K = 4				
1		Grp2	Grp2	Avg	Grp2	Grp2	Grp2	Avg	Grp2	Grp2	Grp2	Grp2	Avg
	HM	46	42	44	46	51	51	49	48	48	49	51	49
	Ind	47	45	46	45	46	44	46	45	46	46	47	46
2		Grp2	Grp2	Avg	Grp2	Grp2	Grp2	Avg	Grp2	Grp2	Grp2	Grp2	Avg
	HM	29	44	36	30	41	50	40	33	36	42	50	40
	Ind	24	46	35	24	40	45	36	26	30	37	46	34
3		Grp1	Grp5	Avg	Grp1	Grp3	Grp5	Avg	Grp1	Grp3	Grp4	Grp5	Avg
	HM	59	45	52	59	35	47	47	55	39	44	49	47
	Ind	62	46	54	50	34	46	43	60	34	42	47	46
4		Grp1	Grp4	Avg	Grp1	Grp2	Grp4	Avg	Grp1	Grp2	Grp3	Grp4	Avg
	HM	36	41	39	40	39	41	40	38	42	40	44	41
	Ind	37	37	37	35	33	38	35	35	34	34	37	35
5		Grp1	Grp4	Avg	Grp1	Grp2	Grp4	Avg	Grp1	Grp2	Grp4	Grp5	Avg
	HM	30	46	38	31	32	47	37	31	30	49	45	39
	Ind	26	46	36	27	30	43	33	28	33	45	44	38
6		Grp3	Grp5	Avg	Grp1	Grp3	Grp5	Avg	Grp1	Grp2	Grp3	Grp5	Avg
	HM	34	47	40	51	36	46	44	53	38	39	42	43
	Ind	36	63	50	58	33	60	60	60	39	34	62	49

Note: the selection probabilities for the target dose for each group and average across groups. The results using HM design are in the first row of each scenario, while the results from an independence design are displayed in the next row. $K=5$ group indices are listed for each K within each scenario. HM: hierarchical modeling.

5 Discussion

We discuss HM for sharing information across populations in phase I clinical trials. First, we present hierarchical extensions of three commonly used dose–toxicity models for phase I oncology trials. These models allow for a different population-specific MTD, while borrowing strength across populations, when appropriate, to achieve a more precise estimate of the population-specific MTD. We then proposed three DFGs for phase I clinical trials using HM. The proposed DFGs allow us to fully utilize the advantages of HM, while protecting patient safety by restricting dose escalation until it has been shown that the current dose level has an acceptable toxicity profile at the current dose level. Our simulation results suggest that all three models are able to borrow strength when the population-specific MTD is constant across populations, resulting in a more precise estimate of the population-specific MTD, but the hierarchical power model is more robust when the populations are more heterogeneous. In addition, we found that the "321" DFG provided the best trade-off for estimating the population-specific MTD, while protecting patient safety of the three DFGs considered. Finally, our simulation results suggest that HM would be beneficial with as few as three populations but independence designs are more effective if only two populations exist.

Returning to our motivating example of completing multiple, independent phase I trials to evaluate a single agent in multiple populations with different background standards-of-care, our results are clear that completing independence designs for each population does not represent the optimal approach and that completing parallel designs while using HM to share information across populations is more efficient. Our results indicate that the HM approach results in an increased probability of correctly identifying the population-specific MTD and an increased number of patients treated at the population-specific MTD while decreasing the percent of patients experiencing DLTs and the number of patients treated at unsafe doses above the population-specific MTD, in most cases. This provides strong evidence for pursuing this type of design but additional work is needed to identify the practical and theoretical challenges related to implementing this approach. Specifically, we note that our approach assumes the populations' dose–toxicity profiles are fully exchangeable. Consequently, a limitation of this approach is the potential negative impact when the assumption of exchangeability is violated. In this case, alternative methods that are robust to violations of this assumption may be considered, such as the approach by Neuenschwander et al.[27]

We have presented the results of this manuscript as an extension of the CRM but the methodology discussed in Section 2 can be applied to Bayesian adaptive phase I designs more broadly. Some clinicians remain hesitant to implement the CRM due to concerns about escalating too quickly, resulting in excess DLTs. We proposed three DFGs to protect patient safely, with the "321" algorithm exhibiting the best performance, and our results indicate that HM actually results in less DLTs than independent CRM designs. Nevertheless, other modifications to protect patient safety could also be considered. For example, the escalation with overdose control (EWOC) was proposed as an approach to limit DLTs in phase I clinical trials,[7] while others[8] propose classifying the posterior probability of a DLT into four categories: under-dosing, targeted toxicity, excessive toxicity, and unacceptable toxicity, and using these probabilities to guide dose finding. The aforementioned approaches represent changes to the dose-finding algorithm and not the dose–toxicity model. As a result, these methods could be easily integrated with HM to achieve the benefits of borrowing information across populations, while further limiting DLTs.

Our simulation results presented in Section 4 are dependent on the prior distributions and prior dose–response skeletons discussed in Section 2. We considered a variety of prior input values and different prior distributions for our variance parameters; however, we found the general trends to be consistent. Increasing the domain $(-a, a)$ for our smoothing parameter resulted in substantially improved performance when the populations are homogeneous but much worse performance when the populations are heterogeneous. We calibrated the hyper-parameters for the hierarchical logistic regression and curve-free method to be consistent with the prior skeleton used in the hierarchical power model. While the hierarchical power model proved to be superior in this case, it is possible that other skeletons might be more appropriate for the other two models. However, we expect that the non-ideal pooling behavior observed in the hierarchical logistic regression and curve-free models would be similar regardless of the skeleton.

Phase I dose-escalation designs like the CRM, attempt to identify the MTD under the assumption that the maximum dose that can be given safely is the best choice for identifying an efficacious dose as well. In practice, it is often the case that the efficacy of a treatment may plateau or even diminish as dose increases, while potential toxicity is expected to increase monotonically with dose. This motivates the use of phase I–II designs that consider both efficacy and toxicity during dose finding. These designs often rely on a parametric model for the dose–response relationship for efficacy and toxicity and it would be worthwhile to investigate if HM would be beneficial in this scenario as well.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is partially funded by the Doctoral Dissertation Fellowship from the Graduate School at the University of Minnesota (KMC) and a gift from Medtronic Inc. (JSK).

Supplemental material

Supplemental material is available for this article online.

References

1. Storer BE. Design and analysis of phase I clinical trials. *Biometrics* 1989; **45**: 925–937.
2. O'Quigley J, Pepe M and Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 1990; **46**: 33–48.
3. Cheung YK and Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* 2000; **56**: 1177–1182.
4. Ivanova A, Montazer-Haghighi A, Mohanty SG, et al. Improved up-and-down designs for phase I trials. *Stat Med* 2003; **22**: 69–82.
5. Liu S, Pan H, Xia J, et al. Bridging continual reassessment method for phase I clinical trials in different ethnic populations. *Stat Med* 2015; **34**: 1681–1694.
6. Goodman SN, Zahurak ML and Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Stat Med* 1995; **14**: 1149–1161.
7. Babb J, Rogatko A and Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat Med* 1998; **17**: 1103–1120.
8. Neuenschwander B, Branson M and Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med* 2008; **27**: 2420–2439.
9. Yin G and Yuan Y. Bayesian model averaging Continual reassessment method in phase I clinical trials. *J Am Stat Assoc* 2009; **104**: 954–968.
10. Kummar S, Ji J, Morgan R, et al. A phase I study of veliparib in combination with metronomic cyclophosphamide in adults with refractory solid tumors and lymphomas. *Clin Cancer Res* 2012; **18**: 1726–1734.
11. Reiss KA, Herman JM, Zahurak M, et al. A phase I study of veliparib (abt-888) in combination with low-dose fractionated whole abdominal radiation therapy in patients with advanced solid malignancies and peritoneal carcinomatosis. *Clin Cancer Res* 2015; **21**: 68–76.
12. Mehta MP, Wang D, Wang F, et al. Veliparib in combination with whole brain radiation therapy in patients with brain metastases: results of a phase 1 study. *J Neuro-Oncol* 2015; **122**: 409–417.
13. Owonikoko TK, Dahlberg SE, Khan SA, et al. A phase 1 safety study of veliparib combined with cisplatin and etoposide in extensive stage small cell lung cancer: a trial of the ecogacrin cancer research group (e2511). *Lung Cancer* 2015; **89**: 66–70.
14. Berry SM, Carlin BP, Lee JJ, et al. *Bayesian Adaptive methods for clinical trials* (vol. 38). Boca Raton, Fl: CRC press, 2010.
15. Patterson S, Francis S, Ireson M, et al. A novel bayesian decision procedure for early-phase dose-finding studies. *J Biopharm Stat* 1999; **9**: 583–597.
16. Braun TM and Wang S. A hierarchical Bayesian design for phase I trials of novel combinations of cancer therapeutic agents. *Biometrics* 2010; **66**: 805–812.
17. O'Quigley J and Iasonos A. Bridging solutions in dose-finding problems. *Stat Biopharm Res* 2014; **6**: 185–197.
18. Berry S, Broglio K, Groshen S, et al. Bayesian hierarchical modeling of patient subpopulations: efficient designs of phase II oncology clinical trials. *Clin Trials* 2013; **10**: 720–734.
19. Thall PF, Wathen JK, Bekele BN, et al. Hierarchical bayesian approaches to phase II trials in diseases with multiple subtypes. *Stat Med* 2003; **22**: 763–780.
20. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery* 2011; **1**: 44–53.
21. O'Quigley J and Shen LZ. Continual reassessment method: a likelihood approach. *Biometrics* 1996; **52**: 673–684.
22. Gelman A, et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal* 2006; **1**: 515–534.
23. Paoletti X and Kramar A. A comparison of model choices for the continual reassessment method in phase I cancer trials. *Stat Med* 2009; **28**: 3012–3028.

24. Gasparini M and Eisele J. A curve-free method for phase I clinical trials. *Biometrics* 2000; **56**: 609–615.
25. Plummer M. *rjags: Bayesian graphical models using MCMC*, 2011. R package version 3-10.
26. Stein C. *Inadmissibility of the usual estimator for the mean of a multivariate normal distribution* published in *Proceedings of the Third Berkeley symposium on mathematical statistics and probability* (vol. 1.) 1956.
27. Neuenschwander B, Wandel S, Roychoudhury S, et al. Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharm Stat* 2015; **15**: 123–134.