MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/psst.1686

Published online 11 May 2015 in Wiley Online Library

A Phase I/II adaptive design for heterogeneous groups with application to a stereotactic body radiation therapy trial

Nolan A. Wages,[a]* Paul W. Read,[b] and Gina R. Petroni[a]

Dose-finding studies that aim to evaluate the safety of single agents are becoming less common, and advances in clinical research have complicated the paradigm of dose finding in oncology. A class of more complex problems, such as targeted agents, combination therapies and stratification of patients by clinical or genetic characteristics, has created the need to adapt early-phase trial design to the specific type of drug being investigated and the corresponding endpoints. In this article, we describe the implementation of an adaptive design based on a continual reassessment method for heterogeneous groups, modified to coincide with the objectives of a Phase I/II trial of stereotactic body radiation therapy in patients with painful osseous metastatic disease. Operating characteristics of the Institutional Review Board approved design are demonstrated under various possible true scenarios via simulation studies.

Keywords: Phase I; dose-finding; Phase I/II; heterogeneity; adaptive design; optimal dose

1. INTRODUCTION

Historically, the primary objective of Phase I clinical trials in oncology is to identify the maximum tolerated dose (MTD) of the agent being investigated from a discrete set of available doses. Numerous Phase I designs have been proposed for identifying the MTD from a set of I doses, $\mathcal{D} = \{d_1, \ldots, d_I\}$, in which toxicity is measured as a binary outcome; a thorough review of which is given in a recent manuscript by Braun [1]. In oncology trials of a single cytotoxic agent, identification of the MTD is often determined by considering dose-limiting toxicity (DLT) information only, with the assumption that the MTD is the highest dose that satisfies some safety requirement and therefore provides the most promising outlook for efficacy. In a subsequent Phase II trial, the therapy is evaluated for efficacy at the recommended dose (estimated MTD or one dose level below). Phase I trials evaluating the toxicity of single agents are becoming less common, giving way to more complex studies including those involving combinations of agents, targeted therapies or multiple treatment schedules. This class of more complex dose-finding studies has created the need to adapt early-phase trial design to the specific type of drug being investigated and the corresponding endpoints. Frequently, Phase I and Phase II trials are performed independently, without formally sharing information across the separate phases. More recently, there has been a shift in the paradigm of drug development in oncology to integrate Phase I and Phase II trials to accelerate the drug development process, while potentially reducing costs. To this end, several recent Phase I/II methods have extended dose-finding methodology to allow for the modeling of both toxicity and efficacy, some within the context of the more complex problems mentioned earlier [2–4].

Another complexity that may arise in early-phase dose-finding studies is encountering some patient heterogeneity, in which patients can be categorized into two or more prognostic groups. A straightforward example of the two group structure is seen in the degree of previous treatment (heavily pretreated vs. no/lightly pretreated). Given the varying amounts of prior treatment, it is not unreasonable to assume that there could be a difference in the way the patients in each group tolerate and respond to a new treatment (or treatment regimen), and it may be important to incorporate this heterogeneity into the design of the trial. Most studies do not incorporate any group structure into the design. The resulting recommended dose is weighted in favour of the appropriate dose for the most frequently occurring group, and effectively moves away from 'personalized' dosing for patients not in the dominant group. Within this scenario of heavily pretreated versus no/lightly pretreated example, it is reasonable to assume that a heavily pretreated patient may be more likely to experience a DLT than a no/lightly pretreated patient. Consequently, an objective of the trial may be to find an MTD within each group, and it would be beneficial for the trial design to reflect this goal. A simple solution would be to conduct parallel dose-finding studies, one for each group. This approach would not make use of any information common between the groups and, thus, may not be very efficient. There are many examples of these designs, including Ramanathan *et al*. [5], which stratifies patients into 'none', 'mild', 'moderate' or 'severe' liver dysfunction at baseline. A similar classification is used by LoRusso *et al*. [6].

[a] *Division of Translational Research & Applied Statistics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA*

[b] *Department of Radiation Oncology, University of Virginia, Charlottesville, VA 22904-4135, USA*

**Correspondence to: Nolan A.Wages, Division of Translational Research & Applied Statistics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA 22908, USA.*
E-mail: nwages@hsc.edu

Dasari *et al*. [7] defined groups in terms of the type of cancer, while Prados *et al*. [8] stratifies patients by prior therapies. Ura *et al*. [9] and Kim *et al*. [10] conduct Phase I trials in groups defined by patient genetic characteristics. These examples use different dose-finding designs within groups; Ramanathan *et al*. [5] use the traditional 3 + 3 design while Ura *et al*. [9] use the more efficient continual reassessment method (CRM; [11]), but they share the common feature that the MTD in each group is determined only from the data obtained within that patient group.

Given the typically small samples used in early-phase trials, a more efficient design that can borrow information in order to identify the most appropriate dose level for both groups is desirable. Several advanced design methods have been proposed in order to address the problem of patient heterogeneity in dose-finding. O'Quigley, Shen and Gamst [12] introduced a two-sample CRM, which allowed for the identification of the appropriate MTD's for two groups simultaneously. Legedza and Ibrahim [13] proposed a related method, augmenting the dose-toxicity model for a vector of patient characteristics and putting a prior on the coefficient in the dose-toxicity model. In O'Quigley *et al*. [12], no assumption was made regarding the order of tolerance towards the treatments between the two groups. O'Quigley and Paoletti [14] proposed a two-parameter CRM for ordered groups that utilizes known differences between the groups. Yuan and Chappell [15] describe a generalization of up-and-down, isotonic and CRM designs for trials carried out in ordered groups. Ivanova and Wang [16] also incorporate isotonic estimates into designs for ordered groups that take into account both toxicity and efficacy endpoints. Thall, Nguyen, and Estey [17] introduced a Bayesian sequential Phase I/II method that accounts for the interaction between patient and dose covariates. Morita [18] presented an application of the CRM that utilized information from Caucasian patients in order to design a Phase I dose-finding study for Japanese patients.

In the usual single-group CRM, an under-parameterized (one-parameter) working model is often utilized in modeling toxicity. In O'Quigley *et al*. [12] and O'Quigley and Paoletti [14], a two-parameter CRM model was introduced, with the additional parameter essentially measuring differences between the groups. The regression models described in O'Quigley *et al*. [12] and O'Quigley and Paoletti [14] allow for a large, possibly infinite, number of possible values for the second parameter. O'Quigley and Iasonos [19] briefly outlined the potential for being more restrictive on the range of potential values for the 'group' parameter and suggested using a very small, discrete set of values, which has been termed the 'shift' model [20]. In this parameterization, suppose Group 1 is recommended dose level d_ν, based on the minimization of the distance between its estimated DLT probability, $\widehat{\pi}$, and some target DLT rate, ϕ_T; that is, $\nu = \arg\min_i |\widehat{\pi}_i - \phi_T|$. Then, Group 2 will be recommended to receive either the same dose or a dose that is 'shifted' one, two or more levels away from d_ν. The investigators involved in the study can specify constraints on both the direction and magnitude of the difference between the two groups. A study of the operating characteristics of the shift model in identifying the MTD in each of several groups can be found in Shu [20]. Recently, O'Quigley and Iasonos [21] outlined the use of the shift model in the area of bridging, a problem closely linked to that of group heterogeneity. The purpose of this article is to demonstrate a practical extension of the shift model as part of a Phase I/II design for a trial of stereotactic body radiation therapy (SBRT) in patients with painful osseous metastatic disease classified into two prognostic groups. The work of Shu [20] is a method for handling MTD identification in each group, based on a single toxicity endpoint. A key extension made here is that we are incorporating an additional endpoint to guide allocation within each group. Therefore, the safety objective in this current trial deviates from one of identifying the MTD in each group according a target rate ϕ_T, to one of estimating an acceptable set of doses defining a set of acceptable doses with regards to safety. To this end, ϕ_T is redefined from a target DLT probability to a maximum acceptable toxicity tolerance that will guide the definition of safe doses. This set of acceptable doses is used to drive allocation towards doses that optimize a response endpoint.

2. MOTIVATING TRIAL

The trial design described in this article is an IRB-approved, Phase I/II trial of rapid helical radiation therapy for patients with painful osseous metastatic disease that was designed, and is currently open to accrual, at the University of Virginia (UVA) Cancer Center.

2.1. Two prognostic groups

Patients are grouped into good and poor prognosis subjects, with poor prognosis patients having a shorter life expectancy. The poor prognosis patients are unlikely to require radiation re-treatment for recurrent pain after treatment, and therefore may not require higher dose escalation for palliation, which is associated with more toxicity. At study entry, patients will be classified as having a good prognosis and stratified into Group 1 if they meet all of the following criteria: (1) limited metastatic disease in 2 or fewer organ systems especially bone only disease; (2) have not completed second line chemotherapy or are on hormonal therapy only; (3) indolent disease process with relatively stable disease on serial imaging over the past 3-6 months or reduced tumor burden due to systemic therapy; or (4) stable or improving performance status (PS) over the past 6 months with a current Eastern Cooperative Oncology Group (ECOG) PS of 0–2. Patients are classified as having a poor prognosis and stratified into Group 2 if they meet any of the following criteria: (1) diffuse metastatic disease in three or more organ systems; (2) completed second line chemotherapy with persistent disseminated disease; (3) rapid progression of disease over the past 3–6 months or rapid decline in performance status over the last 6 months with a current ECOG PS of 3 or 4. It was anticipated that 75% of the patient population will be from Group 1 and 25% from Group 2.

2.2. Design considerations

The current trial is evaluating $I = 4$ dose levels in each of the two prognostic groups, which are given in Table I. The trial was designed to determine the optimal dose, defined as the dose with acceptable toxicity and that minimizes re-treatment rate, in each group, among the available dose levels. A toxicity endpoint, Y, is defined as a binary random variable (i.e. DLT yes/no). Similarly, we define a 'response' variable, Z, as a binary outcome (i.e. re-treatment yes/no).

Patients receive a radiation treatment via a novel real time radiation oncology workflow for the delivery of high dose single fraction stereotactic body radiation therapy (SBRT) known as the Scan-Plan-QA-Treat STAT-RAD workflow. The overall goal of this workflow is to develop a more rapid, convenient and effective palliative SBRT approach for patients with osseous metastases that is less toxic and less expensive than current clinical treatment

Table I. Available dose levels of STAT RAD-based single fraction SBRT.

Prognosis group	Doses in Gy 8	10	12.5	15
2 - Poor	d_{21}	d_{22}	d_{23}	d_{24}
1 - Good	d_{11}	d_{12}	d_{13}	d_{14}

SBRT, steriotactic body radiation therapy; Gy, gray.

regimen. Therefore, this trial will attempt to establish whether conformal radiation dose escalation above 8 gray (Gy) can improve patient pain scores and reduce re-treatment rates, because this is a major drawback of the standard of care 8 Gy single fraction treatment. Based on pilot data from a previous clinical trial and data from other larger randomized clinical trials, the investigators believe that patients with osseous metastases treated with STAT RAD-based single fraction SBRT with dose escalation beyond 8 Gy will have a target lesion re-treatment rate less than 20% with acceptable toxicity.

2.2.1. Safety. Adverse events are being assessed and acute toxicity graded using the National Cancer Institute Common Terminology Criteria for Adverse Events version 4. Late toxicity is being scored using the Radiation Therapy Oncology Group Late Radiation Toxicity Scoring System. A DLT is defined as any treatment related grade $\geqslant$ 4 toxicity, within 90 days of treatment. For escalation decisions, subjects must be observed for a minimum of 30 days after their last radiation therapy (RT) treatment. However, any DLT observed through 90 days will be used in the modelling stage of this two-stage, model-based design. Acceptability is defined by any estimated DLT probability less than or equal to a maximum toxicity tolerance of $\phi_T = 20\%$, which was chosen based on the expectedness of adverse events. In group g, the trial is investigating a set of $I = 4$ doses, $\mathcal{D} = \{d_{g1}, \ldots, d_{g4}\}$, and the probability of DLT at dose d_{gi} is denoted $\pi_T(d_{gi})$. The primary safety objective is to identify an acceptable set of doses in group g defined by $\mathcal{A}_g = \{d_{gi} : \pi_T(d_{gi}) \leqslant \phi_T\}$. As is typical in Phase I trial design, it is assumed that within each prognosis group that the dose-toxicity relationship is monotone increasing so that $\pi_T(d_{g1}) \leqslant \cdots \leqslant \pi_T(d_{g4})$. It is expected that patients in Group 2 will be more sensitive to the treatment and thus have a higher probability of DLT compared with patients in Group 1. Therefore, the maximum safe dose for Group 2 should be lower than the maximum safe dose for Group 1. It may be that if the highest dose in $\mathcal{A}_1$ in Group 1 is estimated to be $d_{1\nu}$, then the highest dose in $\mathcal{A}_2$ will be one, two or three levels below $d_{1\nu}$. Suppose a random variable M, with possible values $m = 1, \ldots, 3$, for the current trial, accounts for these possible shifts in acceptable doses between the two groups. That is, $M = m$ corresponds to an m-level shift in the set of acceptable doses between the two groups, which is represented by

$$\mathcal{A}_1 = \{d_{11}, \ldots, d_{1\nu}\}; \mathcal{A}_2 = \{d_{11}, \ldots, d_{1\nu^*}\},$$
$$\text{where } \nu^* = \nu - m;\ m = 1, \ldots, 3.$$

These values of ν represent the possible shifts between the groups in this particular trial, because it is expected that Group 1 will tolerate the treatment better than Group 2. The 'true shift' could be any of these three values, and we take account of this uncertainty in modelling toxicity, as described in Section 3. Note that it is possible for the number of dose levels in each set to be equal, even if their respective probabilities are shifted, should all doses be safe or too toxic. Our overall strategy is to formulate a class of models corresponding to various shift (m) values, and allow the accumulating information to guide us as to which model is most represented by the data. Throughout the trial duration, we use a CRM-type 'shift' model to continuously monitor safety data in order to adaptively update an estimated acceptable set $\tilde{\mathcal{A}}_g$ in each group, with which we make allocation decisions to doses that minimize re-treatment rate.

2.2.2. Retreatment. The need for target lesion re-treatment will be assessed within a 4 to 8-week window for each patient, making it observable within a reasonably similar time-frame as the toxicity endpoint. In contrast to toxicity, in this current trial, higher doses may not necessarily produce a better (i.e. lower) retreatment rate. Dose-response relationships may exhibit non-monotone patterns, such as (1) decreasing initially and then leveling off at higher levels or (2) decreasing initially, reaching a minimum point, and then increasing at higher doses levels. Suppose that the probability of retreatment at dose d_{gi} is denoted $\pi_R(d_{gi})$. The primary objective, with respect to response, is to find the optimal dose (OD) in group g, $d_{g\nu}$, from the $i^* \leqslant 4$ safe doses in $\mathcal{A}_g$, defined such that

$$\pi_R(d_{g1}) \geqslant \cdots \geqslant \pi_R(d_{g\nu}) \leqslant \cdots \leqslant \pi_R(d_{gi^*}).$$

The probability $\pi_R(d_{g\nu})$ corresponds to the dose $d_{g\nu}$ where the minimum retreatment rate occurs, or where the dose-response curve begins to level off, beyond which point re-treatment rate does not improve and we are simply adding toxicity.

Within each group, more than one dose may satisfy the safety concerns and therefore provide a range of optimal doses to assess secondary study objectives. Specifically, within the range of optimal doses, the secondary objectives are (1) to determine whether or not there is a significant difference in pain score prior to treatment to 4 weeks following within Group 2 and (2) if more than one dose is contained within the range of optimal dose levels in Group 2, to estimate the difference in the mean change in pain score from pre-treatment to 4 weeks after treatment between pairs of dose levels among those in the range of optimal doses. These secondary objectives will be used to guide stopping rules as outlined in Sections 4.3 and 4.4.

3. ESTIMATION

3.1. Toxicity

The motivating trial described in Section 2 is testing $I = 4$ dose levels in each of $G = 2$ prognosis groups, so that there are $2 \times 4 = 8$ group-dose combinations $\{d_{gi}, g = 1, 2; i = 1, \ldots, 4\}$. We specify a working model for the set of DLT probabilities in each of the groups corresponding to the $m = 1, \ldots, 3$ possible shifts. For a particular model, m, we model $\pi_T(d_{gi})$, the true probability of DLT response at group-dose combination d_{gi} by

$$\pi_T(d_{gi}) = \Pr[Y = 1 | d_{gi}] \approx f_m(d_{gi}, \theta_m) = [p_m(d_{gi})]^{\exp(\theta_m)},$$

where the parameter $\theta_m \in (-\infty, \infty)$ is to be estimated from the data. The $p_m(d_{gi})$ values are pre-specified constants, often termed 'skeleton', with values between 0 and 1 which can be chosen based on clinical experience or using the algorithm of Lee and Cheung [22]. Further discussion of skeleton choice within the context of the current trial is given in a subsequent section. We allow the plausibility of each shift to be described by a set of prior probabilities $\{\xi(1), \ldots, \xi(3)\}$, where $\xi(m) \geqslant 0$ and $\sum \xi(m) = 1$. Even though there is no prior information available on the possible models in the current trial, we formally proceed in the same way by specifying a discrete uniform so that $\xi(m) = 1/3$. At any point in the trial, the accumulated toxicity data can be represented by $\mathcal{Y} = (y_{11}, y_{12}, \ldots, y_{24})$, the number of toxicities and $\mathcal{N} = (n_{11}, n_{12}, \ldots, n_{24})$, the number of patients observed on each group-dose combination. For each of the assumed shift models, $m = 1, \ldots, 3$, the likelihood is given by

$$\mathcal{L}(\theta_m) = \prod_{g=1}^{2} \prod_{i=1}^{4} \{f_m(d_{gi}, \theta_m)\}^{y_{gi}} \{1 - f_m(d_{gi}, \theta_m)\}^{n_{gi} - y_{gi}},$$

from which we can obtain the maximum likelihood estimate, $\widehat{\theta}_m$, of the parameter θ_m for each of the three models. We need a value of m so we weight each of the candidate models as we make progress and appeal to sequential model selection techniques to guide decision-making. A plausible model choice is driven by the maximization of

$$\omega(m) = \frac{\mathcal{L}(\widehat{\theta}_m)\, \xi(m)}{\sum_{m=1}^{3} \mathcal{L}(\widehat{\theta}_m)\, \xi(m)}$$

where $\omega(m)$ is considered as the weight of evidence in favour of model m and $\mathcal{L}(\widehat{\theta}_m)$ is the value of the likelihood evaluated at its MLE. At the inclusion of each new patient, we choose a single model, m^*, that maximizes $\omega(m)$ and implement its working model in generating DLT probability estimates, $\widehat{\pi}_T(d_{gi})$, at each group-dose combination so that

$$\widehat{\pi}_T(d_{gi}) \approx f_{m^*}(d_{gi}, \widehat{\theta}_{m^*}), \quad \text{where } m^* = \arg\max_m \omega(m).$$

Based on these estimates, we obtain a set of acceptable (safe) doses in each group to guide allocation decisions as described in the subsequent section in the dose-finding algorithm.

3.2. Retreatment

For retreatment response, we appeal to a simple, nonparametric approach by estimating response probabilities, $\pi_R(d_{gi}) = \Pr[Z = 1 | d_{gi}]$, using observed retreatment rates. The shape of the dose-response curve is not known at every dose within group g, plus it may not be safe to assume that the nature of this relationship is the same in each group. It is possible for dose-response to exhibit unimodal (decreasing then increasing) patterns in one group and plateau (decreasing then leveling off) in the other group. This complexity can lead to an estimation procedure that may be sensitive to strong parametric modeling assumptions. With the aim of greater practical and computational feasibility, estimation of retreatment probabilities does not rely on a parametric dose-response model. At patient inclusion, the accumulated retreatment data can be represented by $\mathcal{Z} = (z_{11}, z_{12}, \ldots, z_{24})$, the number of retreatments, and $\mathcal{N} = (n_{11}, n_{12}, \ldots, n_{24})$, the number of patients observed on each group-dose combination. Using this information, we can compute a set of estimated retreatment probabilities

$$\widehat{\pi}_R(d_{gi}) = \frac{z_{gi}}{n_{gi}}; \quad g = 1, 2; i = 1, \ldots, 4.$$

At each patient inclusion, these retreatment probability estimates are used to make allocation decisions as described in the subsequent section, in the dose-finding algorithm.

4. TRIAL CONDUCT

4.1. Dose allocation in stage 1

In each group, the initial stage will accrue eligible patients in cohorts of two. The first two eligible patients in each group will be entered onto dose level 1 in the patient's respective group.

- If 0/2 patients experience DLT, then the next cohort is treated at the next highest dose level within that group.
- If 1/2 patients experience DLT in either group, then Stage 2 of the study design begins.
- If 2/2 patients experience DLT, accrual is halted, according to the stopping criteria in Section 4.4.
- In the absence of DLT's, escalation will continue until dose level 4 is reached with the same allocation rules specified earlier. After two patients have been treated at each dose level within a group without DLT, the next cohort is treated at the dose with the lowest observed retreatment rate. Once a DLT has been observed in either group, Stage 2 modelling begins.

4.2. Dose allocation in stage 2

After the run-in phase described above, we sequentially allocate each new patient to the dose estimated to minimize re-treatment among those with acceptable toxicity. Upon obtaining DLT probabilities, $\widehat{\pi}_T(d_{gi})$, for each dose using the estimation procedure outlined in the previous section, we restrict our attention to those doses with estimated DLT rates less than the maximum allowable DLT rate $\phi_T = 0.20$. By estimating an acceptable set, $\tilde{\mathcal{A}}_g = \{\widehat{\pi}_T(d_{gi}) \leqslant \phi_T\}$, we exclude overly toxic doses in each group. If, at any point in the trial, $\tilde{\mathcal{A}}_g$ is empty for group g, we recommend the lowest dose in that group for the next entered patient in g, unless one of the stopping criteria described in the succeeding section is triggered, in which case the trial is terminated for safety. After each cohort inclusion the acceptable set is adaptively redefined based on the current DLT probability estimates, so it is possible, once more data have been observed, for $\tilde{\mathcal{A}}_g$ to include a dose that was previously excluded when a limited amount of data existed. The allocation algorithm depends upon the amount of data that have been observed so far in the trial. In the presence of limited data, we rely on a randomization phase to allocate future patients to acceptable doses.

Early in the trial, there may not be enough data to rely entirely on maximization of estimated retreatment probabilities within $\tilde{\mathcal{A}}_g$ to accurately assign patients to the dose with minimum retreatment and acceptable toxicity. There may be doses in $\tilde{\mathcal{A}}_g$ that have never been tried and information on these can only be obtained through experimentation. It is possible for a sequential design relying on an algorithm that minimizes $\widehat{\pi}_R(d_{gi})$ to become 'greedy' and develop a tendency to repeatedly assign

a suboptimal dose [23]. For example, the first assignment could, by chance, result in a retreatment, meaning $\widehat{\pi}_R(d_{gi}) = 1/1$ at that dose. Implementation of a minimization algorithm at this point could result in this dose never being tried again as a result of this one observation. A common practical solution is to randomly assign a small number of patients to suboptimal levels. This added randomization allows for information to be obtained on competing doses and prevents the method from getting 'stuck' on a suboptimal dose that has been tried early in the trial, thus allowing information to be obtained more broadly. The overall strategy is to randomly assign patients in group g to doses in $\tilde{\mathcal{A}}_g$ until enough data have been obtained at competing doses in $\tilde{\mathcal{A}}_g$ in order to feel comfortable implementing a minimization algorithm. The current trial randomizes patients, with equal probability, to safe doses in $\tilde{\mathcal{A}}_g$ until a minimum of *three* patients have been treated at each dose in $\tilde{\mathcal{A}}_g$. If $n_{gi} < 3$ for any $d_{gi} \in \tilde{\mathcal{A}}_g$, randomize the next eligible patient to dose d_{gi} with probability $1/i^*$, where $i^* \leqslant 4$ is the number of doses contained in $\tilde{\mathcal{A}}_g$. After each dose in $\tilde{\mathcal{A}}_g$ has accrued the minimum number of patients, randomization will cease. In the latter portion of the trial, when a sufficient amount of data have been observed, we will utilize a minimization phase in which we allocate according to the treatment that minimizes re-treatment rate, among the set of acceptable doses. Once $\tilde{\mathcal{A}}_g$ is determined for each group, the recommended dose, d_{gi}, in each group is defined as the acceptable dose with the lowest observed retreatment rate so that

$$d_{gi} = \arg \min_{d_{gi} \in \tilde{\mathcal{A}}_g} \widehat{\pi}_R(d_{gi}).$$

Continuing in this way, the optimal dose in group g is estimated to be the recommended dose, d_{gv}, after the maximum sample size has been exhausted, or one of the stopping criteria described below is triggered.

4.3. Sample size and accrual

Maximum sample size is based upon acquiring sufficient information to assess the secondary objective. In a prior prospective STAT RAD clinical trial currently open at UVA, patient pain scores were recorded using the Brief Pain Inventory (BPI) instrument prior to treatment and at 1 week, 4 weeks, 8 weeks, 12 weeks and 6 months after treatment, with a scale of 0 - 10 used to define the patient's worst pain score in the last 24 h. Preliminary data from 11 patients enrolled in this trial yields a mean difference in score from pre-treatment to 4 weeks of approximately 3.82 with a standard deviation of the difference of approximately 2. For this study, we would consider a drop in 1 standard deviation (i.e. approximately 2 points) to be a significant change in mean BPI worst pain score from pre-treatment to 4 weeks.

A paired t-test with a two-sided 5% level of significance will be used to compare the mean change in BPI worst pain score from pre-treatment to 4 weeks. The ability of this test to detect a 1 standard deviation difference (2 points) in the mean change in BPI worst pain score from pre-treatment to 4 weeks after treatment with at least 95% power requires accrual of 17 patients. Therefore, the study is designed to stop if the recommendation is to assign the next patient in Group 1 to a dose level that already has 17 patients in Group 1 treated at that dose level. Any dose in Group 1 with at least 13 patients accrued will be considered in the range of optimal doses, and a comparison between BPI scores from pre-treatment to 4 weeks will be made with a paired t-test for each dose. With a sample size of 13 patients, this test has 90% power at the 5% significance level to detect a 1 standard deviation difference (2 points) in the mean change in BPI worst pain score from pre-treatment to 4 weeks after treatment. Sample size is estimated from the simulations and will be determined by the stopping rules in the next section. The maximum total sample size is set at 92 patients, based on a maximum of 17 at each dose level in Group 1 and a maximum of six at each dose level in Group 2, because of the expected 75/25% split in group accrual. Although maximum accrual is 92, in the simulation results in the succeeding paragraph, on average, a total of between 27 and 49 patients were required to complete the study. Accrual is estimated at one to two patients per month.

4.4. Stopping rules

Investigators desire some measure by which to terminate the trial in the presence of undesirable toxicity. At any point in the trial, we calculate a binomial confidence interval for toxicity $[\pi_T^-(d_{g1}), \pi_T^+(d_{g1})]$ at the lowest combination, d_{g1}, in each group. The value of $\pi_T^-(d_{g1})$ provides a lower bound for the probability of DLT at d_{g1}, among which we are 95% confident that the true DLT probability for d_{g1} falls. We compare this lower bound to our maximum acceptable toxicity rate ϕ_T. If our lower bound exceeds this maximum tolerance, we can be confident that the new treatment possesses too high of a DLT rate to warrant continuing the trial, leading to the following termination criteria:

(1) The study will stop accruing patients in Group 2 and only treat eligible patients in Group 1 if $\pi_T^-(d_{21}) > \phi_T$.
(2) The study will stop for safety if $\pi_T^-(d_{11}) > \phi_T$.
(3) The study will stop if the recommendation is to assign the next patient in Group 1 to a dose level that already has 17 patients in Group 1 treated at that dose level.

5. STATISTICAL PROPERTIES

5.1. Illustration

In this section, we illustrate the behaviour of the method described in this article under a set of true DLT and re-treatment probabilities, which will serve as Scenario 1 in our simulation studies in the succeeding section. The true (DLT, re-treatment) probabilities, (π_T, π_R), for Group 1 are $\{(0.01, 0.20), (0.05, 0.10), (0.10, 0.20), (0.15, 0.30)\}$, indicating that dose level 2 (i.e. d_{12}) is optimal because it is safe and has the lowest re-treatment rate. For Group 2, the true probabilities are $\{(0.05, 0.30), (0.10, 0.15), (0.15, 0.25), (0.30, 0.30)\}$, indicating that dose level 2 (d_{22}) is optimal. In Group 1, dose levels 1–3 are considered acceptable in terms of safety (DLT rates $\leqslant 20\%$), whereas in Group 2, doses 1 and 2 are considered acceptable. This would indicate that model $m = 1$ is most consistent with the true underlying DLT probabilities, coinciding with a true shift of 1 level between the groups. The method embodies characteristics of the CRM, so we appeal to its features in specifying design parameters. For instance, the skeleton values, $p_m(d_{gi})$, were chosen according to the algorithm of Lee and Cheung [22] and are reflected in the working models given in Table II. It has been demonstrated that CRM designs are robust and efficient with the implementation of 'reasonable' skeletons, defined by adequate spacing between adjacent values [24]. The values in Table II were generated using

Table II. Working models (skeleton) values corresponding to the various possible tolerance shifts between groups.

Model	Prognosis group	Doses in Gy 8	10	12.5	15
$m = 1$	2 - Poor	0.07	0.13	0.20	0.29
	1 - Good	0.03	0.07	0.13	0.20
$m = 2$	2 - Poor	0.13	0.20	0.29	0.38
	1 - Good	0.03	0.07	0.13	0.20
$m = 3$	2 - Poor	0.20	0.29	0.38	0.47
	1 - Good	0.03	0.07	0.13	0.20

the `getprior` function in R [25] package `dfcrm` [26]. We assumed that each model was equally likely at the beginning of the trial and set $\xi(m) = 1/3$.

The data from the entire simulated trial are provided in Table III. Because of the expected 75%/25% split between the groups, the first eight eligible patients are from group 1 and are escalated in cohorts of two on each dose level in the absence of DLT. The first DLT occurs in patient 6 on dose level d_{13}, at which point the modeling stage begins. With this limited amount of data, $m^* = 3$ is estimated to be the true shift model and $\widehat{\theta}_{m^*} = -0.379$. These values are used to calculate DLT probability estimates in each group via $\widehat{\pi}_T(d_{gi}) \approx p_3(d_{gi})^{\exp(-0.379)}$, yielding $\{0.09, 0.16, 0.25, 0.33\}$ in Group 1 and $\{0.33, 0.43, 0.52, 0.60\}$ in Group 2. Patient 7 is in Group 1, within which doses 1 and 2 are indicated to have acceptable toxicity. Because three patients have not yet been observed on each acceptable dose in this group, and we have yet to see an observed retreatment, patient 7 is randomized with probability 1/2 to either dose 1 or dose 2. This randomization yields a recommendation of d_{12} for patient 7, in which he/she does not experience a DLT nor a retreatment. The toxicity and retreatment data are then updated, from which $m^* = 1$ is estimated to be the true shift model and $\widehat{\theta}_{m^*} = -0.290$. The updated DLT estimates become $\{0.07, 0.14, 0.22, 0.30\}$ in Group 1 and $\{0.14, 0.22, 0.30, 0.40\}$ in Group 2, indicating that doses 1 and 2 are acceptable in Group 1 and dose 1 is acceptable in Group 2. It is important to note that the DLT probabilities are updated in Group 2, even though we have yet to observe a patient in this group, illustrating the formal borrowing of information across groups afforded by the model. This notion is also reflected towards the end of the trial when the last patient (pt# 35) treated in Group 2 is recommended d_{21}, yet the final recommendation for this group is d_{22} because of the data accumulated on the final five patients in Group 1. Overall, in this simulated trial, 31 Group 1 patients and 9 Group 2 patients are treated, yielding optimal dose recommendations of d_{12} and d_{22}, respectively.

Table III. Simulated sequential trial of 40 patients illustrating the proposed approach.

Pt#	Group	Dose	y	z	m^*	$\widehat{\theta}_{m^*}$
1	1	d_{11}	0	0	—	—
2	1	d_{11}	0	0	—	—
3	1	d_{12}	0	0	—	—
4	1	d_{12}	0	0	—	—
5	1	d_{13}	0	0	—	—
6	1	d_{13}	1	0	3	−0.379
7	1	d_{12}	0	0	1	−0.290
8	1	d_{11}	0	0	3	−0.262
9	2	d_{21}	0	1	1	−0.199
10	1	d_{13}	0	0	1	−0.119
11	2	d_{21}	0	1	1	−0.076
12	2	d_{22}	0	1	1	−0.016
13	1	d_{13}	0	0	1	0.036
14	1	d_{14}	0	0	1	0.098
15	1	d_{14}	1	1	1	−0.125
16	1	d_{12}	1	0	1	−0.376
17	1	d_{12}	0	0	1	−0.343
18	1	d_{12}	0	0	1	−0.313
19	1	d_{11}	0	1	1	−0.302
20	1	d_{12}	0	0	1	−0.275
21	1	d_{12}	0	0	1	−0.251
22	1	d_{12}	0	0	1	−0.228
23	1	d_{12}	0	0	1	−0.207
24	2	d_{22}	0	0	1	−0.178
25	1	d_{13}	1	0	1	−0.294
26	2	d_{21}	0	1	1	−0.274
27	1	d_{12}	0	1	1	−0.256
28	1	d_{12}	1	1	1	−0.383
29	1	d_{12}	0	0	1	−0.364
30	2	d_{21}	0	0	1	−0.346
31	2	d_{21}	0	0	1	−0.328
32	2	d_{21}	0	0	1	−0.312
33	1	d_{12}	0	0	1	−0.296
34	1	d_{12}	0	0	1	−0.281
35	2	d_{21}	0	0	1	−0.267
36	1	d_{12}	0	0	1	−0.253
37	1	d_{12}	0	0	1	−0.241
38	1	d_{13}	0	0	1	−0.235
39	1	d_{13}	0	0	1	−0.210
40	1	d_{13}	0	1	1	−0.193

OD recommendation for Group 1 is d_{12} after $n_1 = 31$ patients.
OD recommendation for Group 2 is d_{22} after $n_2 = 9$ patients.
OD, optimal dose.

5.2. Simulation studies

Simulation results were run to display the performance of the design characteristics and are provided in Table IV. Because of the lack of existing methods for Phase I/II trials of heterogeneous groups, a comparison to an alternative design is difficult. All simulations on all scenarios were carried out using R statistical language. For each scenario, 1000 simulated trials were run. Each table reports the true DLT probability at each dose level (π_T), the true re-treatment rate at each dose level (π_R), the percentage of trials in which each dose level in each group was recommended as the optimal dose (%rec), the percent allocation of patients to each dose level (%exp), the average number of patients treated in each group (Avg n_g) and the percentage of times the trial was stopped in each group for safety (%stopped). Recall that, according to the stopping rules, if Group 2 is deemed too toxic, accrual to this group halts and experimentation is restricted to Group 1 only. If Group 1 is deemed too toxic, the trial is terminated for safety.

Table IV. Results based on 1000 simulated trials. The optimal dose in each scenario is indicated in bold-type. For each scenario, the table reports the proportion of optimal dose recommendation (%rec), the proportion allocation to each dose (%exp), the average number of patients treated in each group (Avg n_g), and the proportion of trials stopped in each group (%stopped) and overall dose-limiting toxicity (DLT) and retreatment rates.

		Group 1				Group 2					
Scenario	Dose	8	10	12.5	15	8	10	12.5	15	%DLT	%Ret
1	π_T	0.01	**0.05**	0.10	0.15	0.05	**0.10**	0.15	0.30	7.3	19.0
	π_R	0.20	**0.10**	0.20	0.30	0.30	**0.15**	0.25	0.30		
	%rec	0.20	**0.61**	0.16	0.03	0.31	**0.44**	0.17	0.08		
	%exp	0.24	**0.39**	0.22	0.14	0.40	**0.35**	0.20	0.06		
		Avg $n_1 = 33.67$ %stopped = 0.0				Avg $n_2 = 11.24$ %stopped = 0.4					
2	π_T	0.01	0.03	0.06	**0.09**	0.03	0.06	0.09	**0.12**	5.9	19.6
	π_R	0.35	0.25	0.15	**0.05**	0.40	0.30	0.20	**0.10**		
	%rec	0.03	0.06	0.22	**0.68**	0.18	0.24	0.24	**0.35**		
	%exp	0.17	0.19	0.26	**0.39**	0.31	0.27	0.28	**0.14**		
		Avg $n_1 = 36.55$ %stopped = 0.1				Avg $n_2 = 12.22$ %stopped = 0.1					
3	π_T	0.01	0.05	**0.10**	0.15	**0.15**	0.30	0.40	0.50	11.6	20.6
	π_R	0.30	0.20	**0.10**	0.10	**0.40**	0.25	0.15	0.15		
	%rec	0.11	0.22	**0.44**	0.23	**0.65**	0.22	0.08	0.00		
	%exp	0.20	0.25	**0.33**	0.23	**0.65**	0.25	0.09	0.01		
		Avg $n_1 = 33.67$ %stopped = 0.0				Avg $n_2 = 10.82$ %stopped = 4.3					
4	π_T	0.02	**0.10**	0.20	0.30	**0.15**	0.25	0.35	0.65	13.5	16.3
	π_R	0.20	**0.10**	0.10	0.10	**0.30**	0.15	0.15	0.15		
	%rec	0.27	**0.52**	0.19	0.03	**0.75**	0.19	0.03	0.00		
	%exp	0.29	**0.38**	0.22	0.10	**0.74**	0.20	0.05	0.00		
		Avg $n_1 = 31.88$ %stopped = 0.1				Avg $n_2 = 10.52$ %stopped = 3.5					
5	π_T	**0.16**	0.22	0.25	0.30	0.45	0.57	0.66	0.80	23.5	19.7
	π_R	**0.20**	0.10	0.10	0.10	0.30	0.15	0.25	0.30		
	%rec	**0.83**	0.10	0.01	0.00	0.54	0.00	0.00	0.00		
	%exp	**0.74**	0.18	0.05	0.02	0.97	0.03	0.00	0.00		
		Avg $n_1 = 21.51$ %stopped = 5.8				Avg $n_2 = 5.68$ %stopped = 46.1					
6	π_T	0.01	**0.03**	0.06	0.10	0.06	**0.10**	0.18	0.25	6.03	19.3
	π_R	0.20	**0.10**	0.20	0.30	0.30	**0.15**	0.25	0.30		
	%rec	0.20	**0.59**	0.17	0.04	0.27	**0.39**	0.19	0.14		
	%exp	0.24	**0.38**	0.23	0.16	0.39	**0.33**	0.21	0.07		
		Avg $n_1 = 34.46$ %stopped = 0.0				Avg $n_2 = 11.36$ %stopped = 0.6					

Therefore, '%stopped' in Group 1 indicates how often the overall trial terminates and '%stopped' in Group 2 indicates how often accrual to Group 2 halts.

The true scenarios were chosen to reflect a range of situations, with various locations of true optimal doses, involving an assortment of safety profiles within each group. In Scenario 1, all doses in Group 1 are safe, and all but the highest dose are safe in Group 2, which coincides with a true shift of $m = 1$. The retreatment probabilities decrease initially then begin to increase at higher levels, indicating that dose 2 is optimal in each group. In Scenario 2, all doses in each group are safe and the retreatment probabilities are monotone decreasing, making the optimal dose the highest level in each group. In Scenario 3, all doses in Group 1 are acceptable, while only the lowest dose is considered safe in

Group 2, corresponding to a true shift of $m = 3$. The retreatment probabilities decrease initially and then level off beyond level 2 in each group. Most group-dose combinations in Scenarios 4 and 5 are overly toxic in each group, and retreatment rates level off beginning at dose level 2 in each scenario. Scenario 6 is similar to Scenario 1, with the exception that the true shift value is $m = 2$ instead of $m = 1$. In each scenario in Table IV, the target dose is indicated in bold type.

A design parameter that should be addressed is the correlation between toxicity and retreatment responses. The models and inference presented in Section 3 estimate toxicity and retreatment probabilities independently, without regard to their association. Therefore, we generated independent binary responses for toxicity and retreatment for the simulations results presented. We also evaluated the impact of using these independent models in the presence of correlated binary responses. To this end, we fit the independent models using correlated binary data, generated according to various correlation (ρ) values using the `ep` function in `R` [25] package `mvtBinaryep` [27]. We used values of $\rho = \{0.10, 0.25\}$, and performance was relatively unaffected from when we generated independent responses ($\rho = 0$). Overall, the results (not shown) demonstrated that misspecification of the models does not alter the properties of the proposed design, which is consistent with the findings of recent published work in Phase I/II design [28,29].

It is clear from examining the results in Table IV that the proposed design is performing well in terms of recommending optimal doses, as well as allocating patients to these doses. In Scenario 1, the proposed design selects, as the OD in Group 1, the target dose in approximately 61% of simulated trials, while assigning 39% of 33.67 patients to this dose. In Group 2, the recommendation percentage for the optimal dose is 44% after an average of only 11.24 patients accrued in this group. In Scenario 2, recommendation of target doses as the OD occurs in approximately 68% and 35% of simulated trials in Groups 1 and 2, respectively. Nearly 40% of the patients enrolled in Group 1 are treated at the target dose in Scenario 2, whereas most patients were not treated at the OD in Group 2. This result is not unexpected given that only 12 patients on average were accrued in Group 2 and the initial escalation scheme (stage 1) requires that escalation occur slowly through these relatively non-toxic doses.

In Scenario 3, the design identifies the target dose in Group 1 as the OD in approximately 44% of simulated trials, while allocating 33% of patients to that dose. This scenario, more than any other, illustrates the advantage that the shift model can afford a multi-group design. Because of the borrowing of information across groups, the method is able to eliminate the overly toxic doses (1–3) in Group 2 and recommend the target dose in 65% of trials, while treating 65% of patients at this dose. Similar conclusions can be made with regards to Scenario 4. Most notable in Scenario 5 is that all doses in Group 2 are overly toxic. The method correctly terminates accrual to Group 2 in 46.1% of simulated trials and treats 97% of the 5.68 accrued patients at dose level 1. Finally, in Scenario 6, the true OD is recommended in 59% of trials in Group 1. Once again, more patients in Group 2 are treated at dose 1 rather than at the OD (dose 2). Only 11.36 patients, on average, are accrued in this group, yet the method is able to identify the OD in 39% of trials because of the information accumulated in Group 1 that is used by the shift model. Overall, the simulation results indicate that the proposed design is a practical Phase I/II adaptive design for use with multiple risk groups.

6. CONCLUDING REMARKS

In this article, we have outlined a new Phase I/II adaptive design, implemented in an IRB-approved trial of SBRT in patients with painful osseous metastatic disease classified into two prognosis groups. The simulation results demonstrated the method's ability to effectively recommend optimal doses, defined by acceptable toxicity and low re-treatment rates, in a high percentage of trials with manageable sample sizes within each group. The method we outline in this work can be viewed as an extension of the CRM, leaning upon a broader class of working models and model selection techniques, increasing the ability of CRM designs to handle more complex dose-finding problems [30]. Therefore, many of the features associated with CRM design specifications, such as one-parameter models and 'reasonable skeletons' can be employed in practice for implementing these extended designs.

The approach described in this article can be generalized to other settings containing a varying number of dose levels and/or groups. A higher dimensional (i.e. more doses and/or more groups) problem may require an increase in the number of working models to consider. For instance, for trials with more than four doses levels per group, we may have to consider more possible shifts, thus increasing the value of M. Similarly, for trials with more than two groups, we will have to set up possible shift models from one group to the next, for each of the G groups. The model selection techniques of the proposed approach have the ability to handle more models, but the impact of higher dimensional problems on operating characteristics needs further work, and is beyond the scope of this paper. Our method can easily be adjusted in order to maximize a binary efficacy endpoint rather than specifically minimizing retreatment rate.

The development of novel methods in early-phase dose-finding has been rapid in recent years; yet, the use of innovative designs remains infrequent. This can be attributed to several causes, not least of which includes (1) clinician skepticism, and (2) difficulty or assumed difficulty in obtaining approval of entities such as IRB's, pharmaceuticals and the FDA. These complications are likely to be enhanced in the coming years as the recent paradigm of oncology drug development involves a shift to more complex dose-finding problems, such as combination or targeted therapies. This article outlines an IRB-approved design for dealing with the complexity of patient heterogeneity in early-phase dose-finding. The results include the type of simulation information that aid review boards in understanding design performance, such as average sample size, frequency of early trial termination and so on, which we hope will augment early-phase trial design in oncology.

Acknowledgements

Dr Wages is partially supported by Grant Number K25CA181638 from the National Cancer Institute. Dr Petroni is partially supported by Grant Number 1R01CA142859 from the National Cancer Institute. This publication was made possible by Grant Number 1C1CMS331031 from the Department of Health and Human Services, Centers for Medicare & Medicaid Services. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the US Department of Health and Human Services or any of its agencies.

REFERENCES

[1] Braun T. The current design of oncology Phase I clinical trials: progressing from algorithms to statistical models. *Chinese Clinical Oncology* 2014; **3**:2.

[2] Thall PF, Nguyen HQ, Braun T, Qazilbash MH. Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics* 2013; **69**:673–682.

[3] Wages N A, Conaway MR. Phase I/II adaptive design for drug combination oncology trials. *Statistics in Medicine* 2014; **33**:1990–2003.

[4] Zang Y, Lee JJ, Yuan Y. Adaptive designs for identifying optimal biological dose for molecularly targeted agents. *Clinical Trials* 2014; **11**:319–327.

[5] Ramanathan R, Egorin M, Takimoto C, Remick S, Doroshow J, LoRusso P, et al. Phase I and pharmacokinetic study of Imatinib Mesylate in patients with advanced malignancies and varying degrees of liver dysfunction: a study by the National Cancer Institute Organ Dysfunction Working Group. *Journal of Clinical Oncology* 2008; **26**:563–569.

[6] LoRusso P, Venkatakrishnan K, Ramanathan R, Sarantopoulos J, Mulkerin D, Shibata S, et al. Pharmacokinetics and safety of Bortezomib in patients with Advanced malignancies and varying degrees of liver dysfunction: Phase I NCI Organ Dysfunction Working Group Study NCI-6432. *Clinical Cancer Research* 2012; **18**: 1–10.

[7] Dasari A, Gore L, Messersmith W, Diab S, Jimeno A, Weekes C, et al. A Phase I study of sorafenib and vorinostat in patients with advanced solid tumors with expanded cohorts in renal cell carcinoma and non-small cell lung cancer. *Investigational New Drugs* 2013; **31**:115–125.

[8] Prados M, Chang S, Burton E, Kapadia A, Rabbitt J, Page M, et al. Phase I study of OSI-774 alone or with temozolomide in patients with malignant glioma. *Proceedings of the American Society of Clinical Oncology* 2003; **22**, (abstract 394).

[9] Ura T, Satoh T, Tsujinaka T, Sasaki Y, Yamazaki K, Munakata M, et al. Phase I study of irinotecan with individualized dosing based on UGT1A1 polymorphism in Japanese patients with gastrointestinal cancer. (UGT0601). *Journal of Clinical Oncology* 2008; **26**, (May 20 suppl; abstr 14502).

[10] Kim T, Sym S, Lee S, Ryu M, Lee J, Chang H, et al. A UGT1A1 genotype-directed Phase I study of irinotecan (CPT-11) combined with fixed dose of capecitabine in patients with metastatic colorectal cancer (mCRC). *Journal of Clinical Oncology ASCO Annual Meeting Proceedings (Post-Meeting Edition)* 2009, Vol 27, No 15S (May 20 Supplement): 2554.

[11] O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for Phase I clinical trials in cancer. *Biometrics* 1990; **46**:33–48.

[12] O'Quigley J, Shen LZ, Gamst A. Two sample continual reassessment method. *Journal of Biopharmaceutical Statistics* 1999; **9**:17–44.

[13] Legezda A, Ibrahim J. Heterogeneity in Phase I clinical trials: prior elicitation and computation using the continual reassessment method. *Statistics in Medicine* 2001; **20**:867–882.

[14] O'Quigley J, Paoletti X. Continual reassessment method for ordered groups. *Biometrics* 2003; **59**:430–440.

[15] Yuan Z, Chapell R. Isotonic designs for Phase I cancer clinical trials with multiple risk groups. *Clinical Trials* 2004; **1**:499–508.

[16] Ivanova A, Wang K. Bivariate isotonic design for dose-finding with ordered groups. *Statistics in Medicine* 2006; **25**:2018–2026.

[17] Thall PF, Nguyen HQ, Esty EH. Patient-specific dose finding based on bivariate outcomes and covariates. *Biometrics* 2008; **64**:1126–1136.

[18] Morita S. Application of the continual reassessment method to a Phase I dose-finding trial in Japanese patients: East meets West. *Statistics in Medicine* 2011; **30**:2090–2097.

[19] O'Quigley J, Iasonos A. Dose-finding designs based on the continual reassessment method. In *Handbook of Statistics in Clinical Oncology*, (3rd ed), Crowley J, Hoering A (eds). CRC Press. Taylor & Francis Group: Boca Raton, FL; 2006; pp. 21–51.

[20] Shu J. *CRM designs in the presence of patient heterogeneity (Unpublished doctoral dissertation)*. University of Virgina: Charlottesville, VA, 2013.

[21] O'Quigley J, Iasonos A. Bridging solutions in dose finding problems. *Statistics in Biopharmaceutical Research* 2014; **6**:185–197.

[22] Lee SM, Cheung YK. Model calibration in the continual reassessment method. *Clinical Trials* 2009; **6**:227–238.

[23] Thall PF, Nguyen HQ. Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *Journal of Biopharmaceutical Statistics* 2012; **22**:785–801.

[24] O'Quigley J, Zohar S. Retrospective robustness of the continual reassessment method. *Journal of Biopharmaceutical Statistics* 2010; **5**:1013–1025.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013; ISBN 3-900051-07-0, Available at: http://www.R-project.org (accessed 18.6.2014).

[26] Cheung K. dfcrm: Dose-finding by the continual reassessment method. R package version 0.2-2 2013. Available at: http://CRAN.R-project.org/package=dfcrm (accessed 18.6.2014).

[27] By K, Qaqish B. mvtBinaryEP: generates correlated binary data. R package version 1.0.1 2011. Available at: http://CRAN.R-project.org/package=mvtBinaryEP (accessed 18.6.2014).

[28] Cai C, Yuan Y, Ji Y. A bayesian dose finding design for oncology clinical trials of combinational biological agents. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2014; **63**:159–173.

[29] Wages NA, Tait C. Seamless Phase I/II adaptive design for oncology trials of molecularly targeted agents. *Journal of Biopharmaceutical Statistics* 2014, [epub ahead of print], DOI: 10.1080/10543406.2014.920873.

[30] O'Quigley J, Conaway MR. Extended model-based designs for more complex Phase I clinical trials. *Statistics in Medicine* 2011; **30**:2062–69.