Published in final edited form as:

Clin Trials. 2011 August ; 8(4): 370–379. doi:10.1177/1740774511410732.

Incorporating lower grade toxicity information into dose finding designs

Alexia Iasonos¹, Sarah Zohar², and John O'Quigley³

¹Department of Biostatistics, Memorial Sloan Kettering Cancer Center, New York, USA

²INSERM, U717, Biostatistics Department, F75010 Paris, France

³INSERM, Université Paris VI, Place Jussieu, 75005 Paris, France

Abstract

Background—Toxicity grades underlie the definition of a dose limiting toxicity (DLT) but in the majority of phase I designs, the information contained in the individual grades is not used. Some authors have argued that it may be more appropriate to consider a polytomous rather than dichotomous response.

Purpose—We investigate whether the added information on individual grades can improve the operating characteristics of the Continual Reassessment Method (CRM).

Methods—We compare the original CRM design for a binary response with two stage CRM designs which make different use of lower-grade toxicity information via simulations. Specifically we study; a two-stage design that utilizes lower-grade toxicities in the first stage only, during the initial non model-based escalation, and two-stage designs where lower grades are used throughout the trial via explicit models. We postulate a model relating the rates of lower grade toxicities to the rate of DLTs, or assume the relative rates of low to high grade toxicities is unknown. The designs were compared in terms of accuracy, patient allocation and precision.

Results—Significant gains can be achieved when using grades in the first stage of a two-stage design. Otherwise, only modest improvements are seen when the information on grades is exploited via the use of explicit models, where the parameters are known precisely. CRM with some use of grade information, increases the number of patients treated at the MTD by approximately 5%. The additional information from lower grades can lead to a small increase in the precision of our estimate of the MTD.

Limitations—Our comparisons are not exhaustive and it would be worth studying other models and situations.

Conclusions—Although, the gains in performance were not as great as we had hoped, we observed no cases where the performance of CRM was poorer. Our recommendation is that investigators might consider using graded toxicities at the design stage.

Keywords

Dose-finding; Phase I; Toxicity Grades; Dose Limiting Toxicity

1 Introduction

The majority of phase I designs use the presence of severe toxicities as a guide to finding the maximum tolerated dose (MTD) which is the aim of Phase I trials. Dose limiting toxicities (DLT) are defined as pre-specified severe adverse events (AE) of grade 3 or higher, based on Common Toxicity Criteria (CTCAE v4.0) [1]. CTCAE are international guidelines that measure the severity of an AE from mild (grade 1) to death-related (grade 5). Most phase I trials [2] use a binary response for each patient which indicates the presence or absence of DLTs. Many statistical papers have explored the use of individual grades either by combining various toxicities to a summary measure and assuming a continuous response [3, 4], or by fitting multivariate models for ordinal response, when the outcome is any toxicity grade in the scale of 1-5 [5, 6, 7]. A number of authors, [4, 8] have proposed methodology that transforms the observed toxicities per patient, from mild to severe, into a single summary measure of “equivalent toxicity score” or “toxicity burden”. This summary measure takes the form of a linear combination of weights resulting in a single, continuous or quasi-continuous outcome, whose expected value is a weighted sum of rates of different toxicity types and grades. Chen et al [9] have extended designs with escalation with overdose control (EWOC [10]) to use a quasi-continuous variable as a toxicity response. Specifically, they proposed a novel toxicity score system that quantitatively maps the multiple toxicities per patient to a normalized toxicity score. Some of these designs have been shown via simulations to be superior to designs that use a binary response under certain conditions. However, there are practical implications that one should take into account. These designs are mapping the summary scores into a scale that it is not easily interpretable by clinical investigators, and in addition the designs are targeting an acceptable toxicity level based on an arbitrary cutoff that it is not directly related to the original 1-5 scale. Moreover, the weighting scheme of certain configurations of toxicities can result in an acceptable “summary score” based on a pre-selected cutoff, when in fact individually these toxicities could be deemed alarming or unsafe by the clinicians. More recently Lee et al [11] proposed a toxicity burden score that summarizes different types and grades of toxicities into a single outcome per patient, and uses this within the Continual Reassessment framework. The approach of Lee et al has a lot of promise since, in actual clinical studies, the type of information on toxicities is always of such a nature. Their idea is to address this specificity directly so that the different types of toxicity, if not necessarily equally weighted, appear on an equal footing when the problem is considered. Other authors [12, 13] have argued that the usual binary information on toxicity (DLT yes/no) can be refined by taking explicit account of lower and intermediary toxicity grades. Our focus here is to still consider the single binary outcome (DLT yes/no) as the main outcome variable of interest, but to see to what extent auxiliary information on lower and intermediary grades can help us in that endeavor. Our development is close to that of Wang et al.[12] but we pay particular attention to the use of information on intermediary grades in the context of the two stage CRM designs.

If we decide to retain as our single outcome criterion the presence of DLT, then there is still an important question concerning the information which may be contained in lower grade toxicities. We might like to know whether information on individual grades can provide a more accurate or more efficient estimate of the MTD. If so, then this information could be used in improving the operating characteristics of a Phase I design. For example, a design might utilize the occurrence of a DLT as its essential outcome and, as a secondary, or auxiliary outcome variable, the occurrence of a lower grade toxicity. In this paper, we compare designs based on the Continual Reassessment Method (CRM) that use information on grades in various ways. For example, one can use individual grades only at the beginning of a two-stage design when limited data is available in order to enrich the dose escalation early on based on the information obtained from non-DLTs and possibly reach the MTD faster. Another design can utilize the grades throughout the trial by modifying the CRM

algorithm for an ordinal response. In this situation, we assume a model so that the rate of occurrence of lower grade toxicities can be related to the rate of occurrence of DLTs. As the data are sequentially gathered, information can be obtained about the relative rates of lower grade toxicities and DLTs. If this knowledge can help in identifying the MTD more accurately then, of course, it might be made use of in dose finding studies.

The definition of a DLT itself combines a lot of information on grades. This is done in as part of a complex, if informal, procedure to determine which kind of toxicities are, broadly, unacceptable. Taking into account of lower grade toxicities is unlikely to make very big changes to our ability to accurately locate the MTD. However, even intuitively, we can see that there are cases where lower grade toxicities will be providing some information, if only be their absence. If there is no indication of any kind of reaction to treatment, or drug levels, then the chance must be quite high that we are still experimenting too low to be in the neighborhood of the MTD. It can also be of interest in its own right to learn something of the relationship between the occurrence of lower grade toxicity to the occurrence of DLT. In the comparative work that follows, we consider two stage CRM designs which make different use of lower-grade toxicity information. Specifically we study; a two-stage design that utilizes lower-grade toxicities in the first stage only, and two-stage designs where lower grades are used throughout the trial via explicit models. For purposes of reference within this article, we call these CRMG(1,0), CRMG(1,1) and CRMG(1,2) as shown in Table 1.

2 Methods

2.1 CRM background

When the response is binary, i.e., the presence or absence of a DLT, we use the CRM-based design as described by O'Quigley et al. [14, 15]. Using the same notation as O'Quigley and Shen [15], we assume the trial consists of k ordered dose levels, $d_1, d_2, \dots, d_k$, and a total of n patients. The visited dose level for patient j is denoted as x_j , and the binary toxicity outcome is denoted as y_j , where $y_j = 1$ indicates a dose-limiting toxicity (DLT) for patient j , and 0 indicates absence of a DLT. O'Quigley and Shen used a one-parameter working model for the dose toxicity relation of the form, $\psi(d_i, a) = \beta_i^a$, where $a \in (0, 1)$ is the unknown parameter, and β_i are the standardized-units representing the discrete dose levels d_i . Since drugs are assumed to be more toxic at higher dose levels, $\psi(d_i, a)$ is an increasing function of d_i . The parameter estimate $\hat{a}$ can be obtained through a Bayesian framework [14] or maximum likelihood estimation [15]. The first design in our comparison is the LCRM as described by O'Quigley and Shen [15]. This is a two stage design where the first stage can be based on any rule, usually one constructed by the clinicians. This allows for great flexibility. For this first stage there is no model and no real statistical considerations. Once a toxicity has been observed (assuming we have at least one non-toxicity) then the model can be fit and we embark on the second stage based on a usual CRM model with binary response. Although not enjoying any particular advantage in terms of efficiency, a first stage based on 3+3 inclusions has a feature attractive to some clinicians in that it looks exactly like the standard design until the first toxicity is observed, at which time the CRM allocation and estimation cycle kick in. For this reason, here, we use a first stage based on 3+3 inclusion until heterogeneity among the responses is observed and then the parameter can be estimated using the likelihood approach. The heterogeneity among the responses guarantees a unique solution to the likelihood equation. Once the current estimate of $\hat{a}$ is calculated, the MTD is defined to be the dose $d_0 \in \{d_1, \dots, d_k\}$ such that some distance measure is minimized. The choice of measure allows a lot of flexibility and we may, for example, prefer to give greater weight to the lower of two doses on either side of the running estimate, such as in EWOC designs. Here we use the simplest distance; $|\psi(d_0, \hat{a}) - \theta|$. The parameter is a pre-specified acceptable probability of toxicity (also known as the target rate).

2.2 Two-stage designs using information on grades during the first stage

Two-stage designs often follow dose escalation rules based on clinical criteria at the beginning of the trial when limited data is observed [16, 17]. Thus, the second design we included for comparison utilizes information on individual grades at the first stage of a two-stage design in order to guide the dose-escalation early on before a DLT is observed. The hypothesis is that individual grades might improve the performance of two-stage CRM-based designs since low grade toxicities, although non-DLTs, can be indicative of an increased probability of encountering a DLT at the respective dose levels. For simplicity we assume that the response for the j th patient Y_j takes three values denoting 0: no toxicity; 1: mild and or moderate toxicity (grade 1-2 based on CTCAE criteria); and 2= DLT (grade 3-5 by CTCAE). Table 2 describes the above severities.

Throughout the paper, we assume that multiple toxicities have already been combined by the clinicians into one single outcome per patient which takes three levels as described above. The design at the first stage allows for dose-escalation when the sum of toxicities is less than or equal to two, whereas it stays at the same dose level if the maximum level has been reached. This design is similar to accelerated designs that use grades during the first stage. However, if the sum of toxicities is greater than two (assuming we also have heterogeneity), then the second stage of the design uses the CRM-based algorithm with a binary response of presence/absence of DLT. Patients are accrued one at a time. This scheme means that in practice, as long as we see toxicities of severities 0, or two or fewer moderate toxicities, we escalate. The first DLT coded 2 necessitates a further inclusion at this same level and only a 0 severity for this inclusion allows an escalation. A severity of 1 or 2 at this inclusion or 3 moderate toxicities among the accrued patients would initiate the second stage. Note that we define escalation rules during the first stage that do not allow an escalation after observing a DLT, so that the design is coherent. CRM designs are shown to be coherent during the modeling stage [18], and Cheung [19] has proposed ways so that two-stage designs are coherent at the point where the first stage ends and the second stage initiates. That is, a patient cannot be treated at a higher dose level than the level for the patient who just had a DLT. It is important to enforce these restrictions in practice in order to make the investigators more comfortable with two-stage designs. The attractive aspect of two-stage designs is that the algorithms and database infrastructure used for binary response can be used without modification to obtain the dose assignment since the second stage of two-stage CRM designs remains unchanged. On the other hand the escalation in the first stage is guided by clinical information rather than by some arguably arbitrary algorithm. For example Figure 1 shows the dose-escalation for a 25-patient trial that followed a first stage design based on the 3+3 algorithm versus using information on individual grades to test six dose levels. If one compares CRM and CRMG(1,0) (top panel), both trials have used CRM with binary response at the second stage and have successfully found the true MTD which is level four, but the trial that utilizes grades during the first stage, CRMG (1,0), is able to reach the MTD faster and treat 11/25 patients at the MTD.

2.3 CRM design with ordinal response

The third design assumes an ordinal response in order to incorporate the grade information. This information is used via a model so that we postulate a relationship between the rates of DLTs and the rates of moderate toxicities that are non-DLTs. Specifically, assume that x_j denotes the visited dose level for the j th patient, Y_j is the response outcome for that patient taking three values 0, 1 and 2, corresponding to no-toxicity; mild or moderate toxicity; and a DLT respectively. Assuming Y_j follows multinomial distribution, the likelihood $L = (a, b, Y, x)$ will be updated after the j th patient is observed by:

$$L_j = L_{j-1} \varphi(x_j, Y_j)$$

where $\varphi(x_j, Y_j)$ is the contribution of the j th patient to the likelihood given by:

$$\varphi(x_j, Y_j) = p_0^{Y_{j0}} p_1^{Y_{j1}} (1 - p_0 - p_1)^{1 - Y_{j0} - Y_{j1}}$$

where Y_{j0} and Y_{j1} are two binary indicators taking the value 1 if patient j has an outcome of $Y = 0$ and $Y = 1$ respectively; 0 otherwise. Now p_0, p_1 can be expressed through the following working models which are a function of the respective dose level x_j and the two parameters a, b .

$$\begin{cases} p_0 = P(Y=0) = 1 - \psi_2(x_j, a, b) = 1 - (x_j^a)^b \\ p_1 = P(Y=1) = \psi_2(x_j, a, b) - \psi_1(x_j, a) = (x_j^a)^b - x_j^a \\ p_2 = P(Y=2) = \psi_1(x_j, a) = x_j^a \quad \text{and} \quad P(Y=1 \text{ or } Y=2) = \psi_2(x_j, a, b) \end{cases}$$

The parameter a has the same interpretation as in the original CRM design, it controls the estimated rate of DLTs at each dose level. The parameter b can be viewed as the parameter that controls the probability of any toxicity of any grade > 0 , i.e., $P(Y = 1 \text{ or } Y = 2) = (x_j^a)^b$. The rate of DLTs, given we have observed a toxicity of mild or higher grade is given by the conditional probability $P(Y = 2 | Y = 1 \text{ or } 2) = (x_j^a)^{1-b}$. Thus, a higher value of b corresponds to a higher probability of DLT given that some toxicity is observed. Depending on the prior information on b from previous studies, one can assume that b is known and constant, thus only a needs to be estimated by the data; or alternatively b can be assumed to be an unknown parameter and should be estimated by the current data. If no prior information is available then both a, b can be estimated by maximizing the above likelihood. Alternatively, assuming some information is available on the ratio of DLTs to non-DLTs among any observed toxicities and using a prior distribution on b , one can estimate both a, b using the Bayesian framework. In this paper, we compare via simulations four CRM-based designs as described below:

1. A basic two-stage CRM where no use is made of the information on grades and the first stage is based on 3+3 dose escalation as described in [16]. When the first DLT is observed the second stage kicks in. This is denoted in the text, figures and tables as CRM.
2. A two-stage CRM design that uses information on grades to allow more rapid escalation in the first stage when low grade toxicities are encountered. The first stage is completed when the sum of the toxicity scores is greater than two and there is heterogeneity among the responses. The second stage makes no use of information on lower grade toxicities. This is denoted in the text, figures and tables as CRMG(1,0).
3. A two-stage CRM design that is the same as CRMG(1,0) but, in the second stage, has the additional feature of using a model to relate the probability of a higher grade toxicity given information on the observation of lower grade toxicity. We use a simple working model for this based on a single parameter b . We study the situation when b is considered known. The known value might have been obtained from other data although, mostly, our use of a known value is for theoretical purposes, providing us with some kind of a bound when compared with the more

realistic situation in which b is not known precisely. Such imprecise knowledge could be characterized by an appropriate prior. We denote this design CRMG(1,1).

4. A two-stage CRM design that is similar as the design to CRMG(1,1) but it considers the parameter b to be unknown and to be estimated, simultaneously with the parameter a , by maximizing the likelihood. This is called CRMG(1,2) since it is based on a two-parameter model.

In addition to estimating the MTD by each design we compare the confidence interval (CI) width for the $P(Y = 2)$ which depends on the variance of $\hat{a}$. A 95% CI for the probability of DLT at the MTD can be estimated by normal approximation when the response is binary as discussed by O'Quigley et al. [20] where the variance of $\hat{a}$ is given in the respective paper. For designs CRMG(1,1), CRMG(1,2) the variance of $\hat{a}$ is approximated numerically and obtained via the inverse of the information matrix when maximizing the likelihood at the $(n + 1)^{th}$ dose assignment after using the accumulated data from n patients.

3 Simulation study

We simulated 1000 trials testing six dose levels with a fixed sample size of 25 patients. All designs used an equal number of patients in order to properly assess whether any gains in estimation were a result of the added information provided by the lower toxicity grades, and not by an increased sample size. The recommended dose level is the dose assignment after observing the 25th patient as proposed by O'Quigley et al. [14]. The standardized units β_i representing the discrete dose levels d_i were varied to reflect a steeper or flatter initial dose-toxicity curve. In the bayesian framework, these β_i are the initial estimates of the probability of a DLT at each dose level, while within the likelihood approach they are generally referred to as a skeleton, the reason being that no interpretation can be given the values of β_i since, raising to any arbitrary positive power, leaves all operating characteristics unaffected. We present simulations with β_i equal to (0.05, 0.1, 0.15, 0.2, 0.25, 0.3) which we believe reflect a slowly increasing curve which is common in clinical practice. The target rate of acceptable toxicity at the MTD varied between 0.2 and 0.3. A higher rate of 0.3 allows for fewer moderate or non-toxicities, whereas a rate of 0.2 allows for more information in the observed mild or moderate toxicities. The results were similar so we present simulations when $\tau = 0.2$. The MTD was selected as the level with estimated $P(\text{DLT})$ closer to the target rate of 0.2.

For the data generation, we ran simulations under various true toxicity curves that varied the true probabilities of $P(Y = 2)$, $P(Y = 1)$, and $P(Y = 0)$. The value of $P(Y = 2)$ was selected arbitrarily in order to change the location of the true MTD as well as the spacing between the MTD and the surrounding dose levels. We present four scenarios as shown in Table 3 where the location of the MTD varies from the middle to the highest dose level. Once $P(Y = 2)$ was fixed, $P(Y = 0)$ was calculated as $P(Y = 0) = 1 - P(Y = 2)^b$, where b was fixed at 0.32. The value of $b = 0.32$ was chosen so that the $P(Y = 0)$ and $P(Y = 1)$ were both equal to 0.4 at the MTD. For the analysis part, we used the working models as described in the previous section. Designs denoted as CRM and CRMG(1,0) are using a one-parameter model with binary response to estimate $\hat{a}$. For design CRMG(1,1) of section 2.3, we evaluated cases when b is known with various values of b , such as $b = 0.32, 0.4, 0.25, 0.20$. Design CRMG(1,2) estimates both $\hat{a}$ & $\hat{b}$ via a two-parameter working model.

The four designs were compared in terms of (i) accuracy, which is the proportion of trials that selected each dose level out of 1000 simulations, (ii) patient allocation, which is the proportion of patients treated at each dose level, (iii) safety, which is measured by the median number of non-toxicities, moderate toxicities, or DLTs, and (iv) precision, which is

the observed and approximated median CI width of the 95% confidence interval for the predicted probability of DLT at the MTD.

4 Results

We have looked at a large number of situations and, for the sake of brevity, we only present a few here. These have not been chosen as a result of any kind of selection process, favorable or otherwise, and can be considered to give a good representation of what takes place across a much larger range of possibilities. We have chosen four scenarios where the dose toxicity curve increases relatively mildly in the vicinity of the MTD (see Table 3). Performance improves of course when these curves are much sharper and becomes poorer when these curves are flatter. The relative performance however does not change much and that is what we are principally interested in here. Nonetheless, there is something a little atypical or special about those situations where the MTD is the highest level among those available and we provide some explanation on this.

Table 4 presents the estimated probability of recommending each dose level for the four designs and for the four scenarios of Table 3. Clearly, it can be argued that when using a CRM design that utilizes grades, the accuracy is higher in locating the MTD and in allocating to levels at the MTD and around it, as opposed to CRM. However, the gains are small in all cases, typically of the order of a few percent. Performance does not seem to improve very much when grades are modeled explicitly through simple models even when the models are known precisely and not prone to misspecification, (as can be judged by comparing the relative performances of CRMG(1,1) and CRMG(1,2)). Most of the gain appears to result from using the grades in the first stage of a two stage design. Given the extra work and unchecked assumptions inherent in using grades in the second stage, we might suggest that the use of grades during the initial escalation, as in CRMG(1,0), is the most valuable. There are no cases where a design which completely ignores information on grades, i.e. CRM, shows better performance than those that do use such information.

Table 5 shows the proportion of patients treated at each dose level. The results are consistent with other findings in the literature that have shown that CRM concentrates the majority of patients at and around the MTD [21]. All methods did not allow skipping dose levels, and CRM without using lower-grades used cohorts of 3 patients per level as long as no DLTs were being observed. This is a conservative initial escalation scheme and slightly more aggressive alternatives might be considered. Designs that use grades provide a compromise while trying to achieve a more aggressive or rapid escalation. Such designs have a safety measure built in, which slows down the dose escalation process once toxicities of moderate grades are encountered. Consistently across all scenarios, CRM with some use of grade information, increases the number of patients treated at the MTD by approximately 5%. This illustrates that the additional information on lower grades, guides to some extent a more informative evaluation of toxicity, by leading the CRM algorithm to the right level faster. However scenario 3 shows an example where possibly the low rate of observed moderate/mild toxicities allows experimentation to higher dose levels as they are deemed safe, whereas the presence of DLT alone would not have allowed this as much. When the MTD is at the highest level (scenario 4), then there is a tendency to push the experimentation to that level and build up a high frequency of allocation at that level. This is particularly true if b is either underestimated, or given an incorrect low value since, the occurrence of lower grade toxicity tends to confirm that the level has a low enough rate of DLTs for the method to stay there. Only the observation of actual DLTs at the highest level would make the method drop back and, when the target is 0.2, as is common, then we do not expect to encounter many actual DLTs. The results of the sensitivity analysis for the parameter b are available by the first author.

Figure 2 shows the median and interquartile range (IQR) for the estimated probability of DLT for the four methods as estimated by the simulations. All methods are very close to the target rate of 0.2 and the confidence intervals as estimated by the IQR range are very similar. Consistently across scenarios, the CI width is narrower under CRM-based methods that use lower grade information (Table 6). Although the improvement is of the order of 3-5%, this gain in precision is obtained without increasing the sample size, which is a major concern in Phase I studies. The additional information from lower grades can potentially increase the precision in our estimate of the MTD as measured by the IQR or CI width, without having to enroll more patients to achieve this. One question of interest is which of the designs provides smaller variability in the estimated probabilities. These results suggest that using lower grades can increase the precision in the estimated value of a and hence the CI width only when b is known. When b is unknown and estimated from the data, the variance of a is very comparable to designs that do not use lower grades at all, depending on the location of true MTD. If the MTD is at the beginning of the dose range, then using lower grades can result in a narrower CI. However, if the true MTD is the last tested level, then using toxicities of lower grades to estimate a two-parameter model might provide more noise and a small increase in the variance of a .

5 Discussion

In this paper, we assessed the use of individual grades in the context of CRM designs by using an ordinal response of three levels in an attempt to reduce the complexity of modeling each individual outcome on the scale of 1-5. The proposed methodology models the rate of DLTs separately from the rate of mild/moderate toxicities. We compared the efficiency and accuracy of CRM-based designs that utilize individual grades in various forms as a guide to reaching the MTD. There are clear advantages when we use information from lower grades during the first stage of a two-stage design which makes some intuitive sense. If, at the outset, we are in the immediate vicinity of the MTD, then rather quickly we will observe DLTs, at which point the second stage will initiate. The information for the fine tuning is now provided by the DLTs alone, and little would have been gained since, after all, we were already close to the MTD. If, however, we start the trial at some levels below the MTD, then use of lower grade information can provide important information upon which we can base a decision to accelerate toward the higher levels and, specifically, those levels at and around the MTD. This can clearly result in patient savings when compared with a 3+3 trial design in which many patients are treated before we are even near the neighborhood where we anticipate the first DLTs. The only other case is where our initial assumptions are completely off and, rather than start out near the MTD, or below the MTD, we start out above the MTD. In this case, the first DLTs will be encountered even more quickly and that information will be enough to quickly de-escalate until we reach a more appropriate dose level.

Since CRM is close to being optimal [22], these findings support that further information on grades, concurrently with the information we obtain from the DLTs themselves, will not do a lot to improve precision. The small improvement we have observed is when we assume some information is known about the true rates linking the occurrence of the higher grades to that of the intermediary grades. When these models are not known exactly and have to be estimated from the observations then it turns out that we gain almost nothing, at least in as far as estimation of the MTD itself is concerned. Using information on lower grades as auxiliary information to the estimation of MTD seems to result in smaller variance in the estimated parameter of interest. This smaller variance suggests the potential of identifying the MTD faster compared to other designs. As one referee pointed out, stopping rules can be incorporated to stop the trial earlier so that clinical testing can move forward without treating so many patients in the phase I setting. Stopping rules such as the ones proposed by

O'Quigley and Reiner[23] can be used in this setting to stop the trial when the probability of having settled at a level is large enough.

In addition, the probability of selecting a higher dose level, is slightly smaller with methods that use lower-grade toxicities which is in agreement with the findings of Lee et al. [11] and Wang et al.[12]. We do gain in that we learn some information about the relationship between the rate for the low grade toxicities and the rates for the higher grade toxicities. But, in some way, that rate at which we learn this information, and the orthogonality of this information with that concerning the identification of the MTD, means that it is never precise enough to allow us to sharpen our inference concerning the MTD alone. Nonetheless, we would conclude by recommending that use be made of this graded information, whether in the first stage of a two-stage design or in both the first and second stages. Although the gains were small, we saw no cases where this information led us astray. Since this information is always available then there is no strong argument for it not to be used, and, at the very least, we do learn something concerning the relative rates, at given levels, of the occurrence of the different types of grades.

Acknowledgments

Partial support for this research was provided by the National Cancer Institute (Grant Number 1R01CA142859). We thank the referees and Editor for their useful comments which have strengthened the manuscript.

Glossary

MTD	maximum tolerated dose
DLT	dose limiting toxicities
AE	adverse events
CTCAE	common toxicity criteria for adverse events
EWOC	escalation with overdose control
CRM	continual reassessment method
CRMG	continual reassessment method with the use of intermediary grades

References

- [1]. DCTD, NCI,NIH,DHHS. Common Terminology Criteria for Adverse Events. Version 4.0. August.2010 <http://ctep.cancer.gov>.
- [2]. Rogatko A, Schoeneck D, Jonas W, et al. Translation of innovative designs into phase I trials. *J Clin Oncol*. 2007; 25:4982–6. [PubMed: 17971597]
- [3]. Ivanova A, Kim SH. Dose finding for continuous and ordinal outcomes with a mono-tone objective function: a unified approach. *Biometrics*. 2009; 65:307–15. [PubMed: 18479486]
- [4]. Yuan Z, Chappell R, Bailey H. The continual reassessment method for toxicity grades: a Bayesian quasi-likelihood approach. *Biometrics*. 2007; 63:173–9. [PubMed: 17447942]
- [5]. Ivanova A. Escalation, group and A + B designs for dose-finding trials. *Stat Med*. 2006; 25:3668–78. [PubMed: 16381057]
- [6]. Van Meter EM, Garrett-Mayer E, Bandyopadhyay D. Proportional odds model for dose finding clinical trial designs with ordinal response. under review.
- [7]. Paul, Rk; Rosenberger, WF.; Flournoy, N. Quantile estimation following non-parametric Phase I clinical trials with ordinal response. *Stat Med*. 2004; 23:2483–95. [PubMed: 15287079]
- [8]. Bekele NB, Thall FP. Dose-Finding Based on Multiple Toxicities in a Soft Tissue Sarcoma Trial. *Journal of the American Statistical Association*. 2004; 99:26–35.

- [9]. Chen, Z.; Tighiouart, M.; Krailo, MD.; Azen, SP. An extended escalation with overdose control design treats toxicity response as a quasi-continuous variable. The Society of Clinical Trials; 31st Annual Meeting; 2010; Abstract 18
- [10]. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*. 1999; 17:1103–120. [PubMed: 9618772]
- [11]. Lee SM, Cheng B, Cheung YK. Continual Reassessment method with multiple toxicity constraints. *Biostatistics*. 2010; 1:13.
- [12]. Wang C, Chen TT, Tyan I. Designs for Phase I cancer clinical trials with differentiation of graded toxicity. *Commun Statist Theory Meth*. 2000; 29:975–87.
- [13]. Ivanova A, Murphy M. An adaptive first in man dose-escalation study of NGX267: statistical, clinical, and operational considerations. *J. Biopharmac. Statist*. 2009; 19:247–55.
- [14]. 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. [PubMed: 2350571]
- [15]. O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. *Bio-metrics*. 1996; 52:673–84.
- [16]. Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR. A comprehensive comparison of the continual reassessment method to the standard 3 + 3 dose escalation scheme in Phase I dose-finding studies. *Clin Trials*. 2008; 5:465–77. [PubMed: 18827039]
- [17]. Moller S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat Med*. 1995; 14:911–22. [PubMed: 7569510]
- [18]. O'Quigley J. Theoretical study of the continual reassessment method. *J. Statist. Planning. Inference*. 2006; 136:1765–80.
- [19]. Cheung YK. Coherence principles in dose-finding studies. *Biometrika*. 2005; 92:863–873.
- [20]. J, O'Quigley. Continual reassessment designs with early termination. *Biostatistics*. 2002; 3:87–99. [PubMed: 12933626]
- [21]. O'Quigley J, Zohar S. Experimental designs for phase I and phase I/II dose-finding studies. *Br J Cancer*. 2006; 94:609–13. [PubMed: 16434987]
- [22]. O'Quigley J, Paoletti X, Maccario J. Non-parametric optimal design in dose finding studies. *Biostatistics*. 2002; 3:51–6. [PubMed: 12933623]
- [23]. O'Quigley J, Reiner E. A stopping rule for the continual reassessment method. *Biometrika*. 1998; 85:741–48.

Figure 1.

Trial History using 3+3 followed by CRM (top left panel) versus using grade information during the first stage of a two-stage CRM Design, CRMG(1,0) (top right), versus modeling the lower-grade toxicities via a 1-parameter, CRMG(1,1), (bottom left) or 2-parameter model, CRMG(1,2) (bottom right). The second stage initiates at patient number 10, 6, 5, and 5 respectively. Full circle indicates low grade non-DLT whereas x indicates DLT.

Figure 2. Median and Interquartile Range for estimated probabilities of dose-limiting toxicity $P(Y=2)$, where Y is the response outcome that takes values 0,1,2 representing no toxicity, moderate and dose limiting toxicity respectively.

Table 1

Designs

CRM	Usual CRM two-stage design making no use of information on grades.
CRMG(1,0)	CRM two-stage design using information on grades only in first stage.
CRMG(1,1)	CRM two-stage design using grades in both stages. For the second stage, the association parameter between intermediary toxicities and DLTs is known.
CRMG(1,2)	CRM two-stage design using grades in both stages. For the second stage, the association parameter between intermediary toxicities and DLTs is estimated by maximizing the appropriate likelihood.

Table 2

Toxicity grades (severities)

Severity	Degree of Toxicity
0	No toxicity
1	Mild/moderate toxicity (non-dose-limiting)
2	Severe (dose limiting toxicity)

Table 3True Toxicity Rates at each dose level d_i

True Toxicity Rates: $P(Y=2)$ $jP(Y=1)$ $jP(Y=0)$				
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
d_1	.10 .38 .52	.05 .33 .62	.01 .22 .77	.01 .22 .77
d_2	.15 .39 .46	.10 .38 .52	.04 .32 .64	.05 .33 .62
d_3	.20 .40 .40	.15 .39 .46	.08 .37 .55	.07 .36 .57
d_4	.30 .38 .32	.20 .40 .40	.12 .39 .49	.10 .38 .52
d_5	.45 .32 .23	.30 .38 .32	.20 .40 .40	.12 .39 .49
d_6	.50 .30 .20	.40 .35 .35	.30 .38 .32	.20 .40 .40

$Y = 0; 1; 2$ corresponds to no toxicity, moderate, and dose limiting toxicity respectively.

Table 4

Estimated Probability of Dose selection

Dose Level	d_1	d_2	d_3	d_4	d_5	d_6	NF
Scenario 1							
CRM	0.142	0.277	0.299	0.211	0.054	0.017	
CRMG(1,0)	0.154	0.300	0.320	0.182	0.036	0.008	
CRMG(1,1)	0.122	0.315	0.345	0.171	0.042	0.005	
CRMG(1,2)	0.144	0.323	0.329	0.164	0.032	0.008	
Scenario 2							
CRM	0.022	0.132	0.237	0.300	0.213	0.096	
CRMG(1,0)	0.021	0.148	0.277	0.312	0.183	0.059	
CRMG(1,1)	0.005	0.127	0.291	0.325	0.192	0.060	
CRMG(1,2)	0.020	0.132	0.281	0.328	0.164	0.075	
Scenario 3							
CRM	0.001	0.011	0.065	0.228	0.320	0.368	0.007
CRMG(1,0)	0	0.015	0.059	0.286	0.319	0.321	
CRMG(1,1)	0	0.003	0.055	0.233	0.358	0.351	
CRMG(1,2)	0	0.016	0.057	0.266	0.318	0.343	
Scenario 4							
CRM	0.001	0.007	0.039	0.095	0.187	0.651	0.020
CRMG(1,0)	0.008	0.007	0.021	0.094	0.209	0.661	
CRMG(1,1)	0	0.001	0.017	0.071	0.203	0.708	
CRMG(1,2)	0	0.005	0.024	0.095	0.189	0.678	0.009

NF= not found

Table 5

Patient allocation: proportion of patients treated at each level

Dose Level	d_1	d_2	d_3	d_4	d_5	d_6
Scenario 1						
CRM	0.31	0.26	0.21	0.14	0.05	0.03
CRMG(1,0)	0.25	0.24	0.23	0.16	0.07	0.05
CRMG(1,1)	0.23	0.25	0.24	0.16	0.07	0.04
CRMG(1,2)	0.24	0.25	0.23	0.16	0.07	0.04
Scenario 2						
CRM	0.19	0.21	0.22	0.18	0.12	0.09
CRMG(1,0)	0.12	0.17	0.21	0.20	0.15	0.15
CRMG(1,1)	0.10	0.16	0.21	0.23	0.16	0.13
CRMG(1,2)	0.11	0.17	0.22	0.21	0.15	0.13
Scenario 3						
CRM	0.13	0.14	0.16	0.18	0.17	0.21
CRMG(1,0)	0.05	0.07	0.13	0.20	0.21	0.34
CRMG(1,1)	0.05	0.06	0.11	0.19	0.23	0.35
CRMG(1,2)	0.05	0.07	0.12	0.20	0.21	0.34
Scenario 4						
CRM	0.13	0.14	0.14	0.15	0.14	0.30
CRMG(1,0)	0.05	0.06	0.09	0.11	0.16	0.53
CRMG(1,1)	0.05	0.06	0.08	0.12	0.16	0.53
CRMG(1,2)	0.05	0.06	0.09	0.12	0.16	0.52

Table 695% Confidence Interval (CI) width of the dose-toxicity model parameter $\hat{a}$

	Median 95% CI width	$\hat{a}$	$Var(\hat{a})$
Scenario 1			
CRM	0.371	0.807	0.046
CRMG(1,0)	0.364	0.794	0.042
CRMG(1,1)	0.321	0.794	0.034
CRMG(1,2)	0.358	0.787	0.040
Scenario 2			
CRM	0.344	0.986	0.060
CRMG(1,0)	0.330	0.943	0.051
CRMG(1,1)	0.319	0.958	0.050
CRMG(1,2)	0.358	0.945	0.058
Scenario 3			
CRM	0.325	1.204	0.080
CRMG(1,0)	0.300	1.151	0.065
CRMG(1,1)	0.316	1.178	0.074
CRMG(1,2)	0.356	1.165	0.089
Scenario 4			
CRM	0.312	1.421	0.107
CRMG(1,0)	0.276	1.393	0.087
CRMG(1,1)	0.304	1.390	0.100
CRMG(1,2)	0.350	1.388	0.128