Continual Reassessment Method vs. Traditional Empirically-Based Design: Modifications Motivated by Phase I Trials in Pediatric Oncology by the Pediatric Brain Tumor Consortium

Arzu Onar[*], **Mehmet Kocak**, and **James M. Boyett**
Biostatistics Department, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Mail Stop 768 Memphis TN 38105

Abstract

In this article we provide additional support for the use of a model based design in pediatric Phase I trials, and present our modifications to the continual reassessment method (CRM), which were largely motivated by specific challenges we encountered in the context of the Pediatric Brain Tumor Consortium trials. We also summarize the results of our extensive simulations studying the operating characteristics of our modified approach and contrasting it to the empirically based traditional method (TM). Compared to the TM, our simulations indicate that the modified version of CRM is more accurate; exposes fewer patients to potentially toxic doses; and tends to require fewer patients. Further, the CRM based MTD has a consistent definition across trials, which is important, especially in a consortium setting where multiple agents are being tested in studies that are often running simultaneously and accruing from the same patient population.

Keywords

Continual Reassessment Method; Up-and-down studies; Dose-toxicity model; simulation; dose finding

1. INTRODUCTION

In recent years, adaptive designs have gained popularity in the context of clinical trials due to the flexibility they offer in utilizing the emerging information throughout the study to guide pre-determined adjustments in an effort to gain efficiency and improve predictability. The continual reassessment method (CRM) is one such approach which has been a common choice for dose-finding trials in various disease areas but most notably in adult oncology. In contrast, the pediatric oncology dose-finding (Phase I) trials have been dominated by the so-called traditional method (TM) also known as the 3+3 up-and-down design.

The most common primary objective of oncology Phase I trials is to estimate the 'maximum tolerated dose' (MTD) of a new agent. The MTD is the dose level associated with a target probability of typically reversible toxic responses, usually in the 20–35% range. Once the MTD is estimated, if the agent seems promising based on biologic, pharmacokinetic and/or clinical information, it is subsequently investigated in Phase II trials for safety and early indications of disease specific efficacy, either as a single agent or in combination with other agents. Thus, it is important that the MTD is estimated as accurately and as reliably as possible. Furthermore,

[*]Corresponding author: Arzu Onar Biostatistics Department, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Mail Stop 768 Memphis TN 38105. arzu.onar@stjude.org. Tel: 901 595 5499. Fax: 901 544 8843.

ideal Phase I designs minimize the total number of patients treated on the trial; and aim to maximize the number of patients assigned to the higher yet safe dose levels, while limiting the number of patients treated at dose levels associated with dose-limiting toxicity (DLT) probabilities exceeding the targeted level. Satisfying all these criteria simultaneously is a tall order for any design, but the magnitude of difficulty in pediatric neuro-oncology Phase I trials is further amplified since they accrue from very heterogeneous and limited populations of patients, mostly with advanced disease.

The intent of this manuscript is to share our modifications to the CRM which were motivated by challenges we faced in the context of pediatric Phase I trials conducted by the Pediatric Brain Tumor Consortium (PBTC). The PBTC (www.pbtc.org) is a multidisciplinary cooperative research organization founded by the National Cancer Institute (NCI) in 1999. The consortium is charged with conducting early phase trials for primary CNS tumors of childhood and is devoted to the study of novel therapies and correlative tumor biology. The PBTC has member institutions across the United States with an Operations and Biostatistics Center housed at St Jude Children's Research Hospital in Memphis, TN. To date, the PBTC has completed eight Phase I trials using a modified CRM design to estimate an MTD. Three additional trials are currently ongoing and 2 more are in development. In the sections below, we briefly introduce the CRM and the TM. We then describe our modifications to the CRM and illustrate the operating characteristics of our algorithm via simulation results before closing by a brief discussion of their implications in a pediatric setting.

2. THE TRADITIONAL METHOD VS. THE CONTINUAL REASSESSMENT METHOD

The TM, also called the 'up-and-down scheme' (Storer, 1989), is an empirical approach to dose-finding that is favored by many clinical investigators in pediatric oncology due to its simplicity and long history in pediatric Phase I trials. A variety of versions have been employed, which tend to differ from each other by cohort size, stopping criteria and/or rules that determine the dose that is ultimately declared as the MTD. Regardless of the variant however, all TM based approaches use a pre-determined set of doses and initiate escalation from the lowest proposed dose level. The version of the traditional design widely used in pediatric oncology which has also been employed for some PBTC trials uses cohorts of 3 patients and proceeds in an empirical manner via the following rules: Start with the lowest proposed dose level. If 0 out of 3 patients treated experience DLT, escalate to the next dose level. If 1 out of the initial 3 patients experiences a DLT, treat three more patients at that dose. If two or more patients (out of 3 or 6 patients) experience DLTs, declare that dose too toxic and de-escalate. Six patients must be treated at the dose to be declared the MTD, and the dose above this level must be too toxic. If the highest dose investigated is safe or the lowest dose is too toxic, the MTD will not have been estimated.

It is clear that such an empirical algorithm cannot be used to estimate a dose associated with a pre-specified toxicity probability_ a sensible definition of the MTD_ since it cannot accommodate a target toxicity level. Furthermore, the dose selected by the TM is more vulnerable to the high variability in small samples sizes of three or six patients, as information from distant doses are quantitatively or statistically irrelevant for the assessment of results from the current dose. In contrast, the CRM is a sequential sampling procedure that utilizes a mathematical model relating dose levels to binary responses (DLT vs. not) in order to estimate the dose at which the desired toxicity probability can be expected. Though the original version proposed by O'Quigley et al. (1990) used a Bayesian algorithm, frequentist procedures have also been developed (O'Quigley and Shen, 1996). Following the considerable amount of criticism received by the original version of the CRM, the algorithm has since been substantially modified and refined both in its theoretical development (O'Quigley and Shen,

1996; Shen and O'Quigley, 1996; O'Quigley, 2002) and in its practical implementation (Goodman et al., 1995; Piantadosi et al. 1998). As a result, the CRM has been successfully utilized in adult trials (see Schoffski et al., 2004; Kasahara et al., 2002; Royce et al., 2001 for examples) but to date, its use has been quite limited in pediatric trials. An informal search in PubMed resulted in 35 references since 1995 that cite the use of CRM in trials, almost all of which are in adult trials; in contrast, 50 references involving statistical work on the CRM are listed for the past 15 years!

The modeling approach behind the CRM is quite simple: Let the dose for the jth patient, X_j, be randomly chosen from a pre-determined discrete set of doses, $x_1, \ldots, x_k$. Note however that the original version of the CRM as well as many of the modified versions assume a continuous dose range, rather than pre-determined levels. Define Y_j is the binary response variable for patient j which takes the value 1 in the event of a DLT and is 0 otherwise, where $j=1, \ldots, n$. Most published versions of the dose-toxicity models utilized for the CRM assume that the relationship between dose and toxicity is monotonically increasing. O'Quigley et al. (1990) modeled the probability of a toxic response at dose level $X_j = x_j$ via $P(Y_j = 1 \mid X_j = x_j) = E(Y_j \mid x_j) = \psi(x_j, a_0)$, where $\psi(x_j, a) = \{(\tanh x_i + 1)/2\}^a$ for $i=1, \ldots, k$ was the one parameter-model utilized. O'Quigley and Shen (1996) indicate that 1-parameter models were preferred from an identifiability perspective. Shen and O'Quigley (1996) provide alternatives to this model as well as a set of conditions under which the CRM is expected to perform well. They also caution that some familiar functions such as the one-parameter logistic function of the form $\psi(x, a) = \exp(ax)/\{1+ \exp(ax)\}$, which was the model used by Korn et al. (1994), do not satisfy these conditions. Interestingly Goodman et al. (1995) also used the one-parameter logistic model; however as a result of their modifications, most of which we also implement, they were able to circumvent the problems encountered by Korn et al. (1994).

Due to the fact that the TM often utilizes cohorts of 3 patients and reacts when 1 or more DLTs are observed, a misconception has formed among many clinicians, that the toxicity probability associated with the MTD chosen through the TM is approximately 33%. In addition to simulation results reported by a variety of authors (Korn et al., 1994; Goodman et al. 1995 and others), Lin and Shih (2001) showed methodologically that this percentage is largely dependent upon the underlying (unobserved) dose-toxicity relationship. Further they showed that the number of patients needed to determine the MTD and the number of patients who will be treated at, above and below the MTD are also affected by the unknown dose-toxicity relationship.

If it is of interest to have a consistent interpretation of the "MTD" across trials, the TM cannot provide one. The "MTD" determined via the TM for one agent may be associated with an appreciably different toxicity probability than the TM "MTD" for another agent, since TM does not attempt to identify a target toxicity level. In a setting like the PBTC, where several Phase I trials for CNS tumors are conducted simultaneously or in close succession targeting the same population of patients using agents with similar DLTs, a consistent definition of the MTD across current trials as well as in relation to completed and future studies is important for selecting agents for subsequent later-phase studies. In the case of cytotoxic drugs in particular, the conventional wisdom is that the higher the dose, the higher the likelihood of efficacy for the agent. Thus in the presence of a fixed toxicity target, identifying the doses with consistent toxicity probabilities across agents is desirable. Such a practice would contribute to improved Phase II designs as well, since it would facilitate more accurate anticipation of toxicities and would allow a better assessment of the agent's potential for combination therapies.

3. OUR MODIFICATIONS TO THE CRM

Pediatric trials are typically preceded by the corresponding adult trial (Smith et al., 1998) and consequently when a pediatric Phase I trial is being developed, an adult MTD is usually available on which the starting pediatric dose level can be based. Though many exceptions exist, a common strategy is to use 80% of the adult MTD as the starting dose (Marsoni et al., 1985). Whether one uses this exact approach or not, the knowledge of the adult MTD, if present, is almost always utilized in perhaps an arbitrary but agreed upon manner in determining the starting dose for the pediatric trial, which substantially reduces the probability that the doses studied during the pediatric trial will be biologically ineffective (Lee et al., 2005). Once the starting dose is set, typically 30% increments are used to determine higher doses to be investigated (Lee et al., 2005).

Though some versions of the CRM assume the availability of doses in a continuous manner within a given range, our experience indicates that having pre-set levels is not only more acceptable to clinicians but is easier to manage operationally, especially in multi-institutional settings. Unlike in adult trials where dosing is typically in terms of mg., patients in pediatric trials are often dosed by body surface area (BSA) and hence the dose is defined in terms of mg/m^2. For oral agents, our protocols include specific instructions for each dose level with respect to BSA adjustments etc. in order to ensure consistency and avoid dosing errors.

We utilize a frequentist, likelihood-based approach and employ a two-parameter logistic model, also used by Piantadosi et al. (1998) and studied briefly by O'Quigley et al. (1990) to represent the relationship between dose and toxicity, i.e. $\psi(x_j, a) = \exp(\alpha + \beta x_j)\{1 + \exp(\alpha + \beta x_j)\}$. As is well known, the logistic model is monotonically increasing if $\beta > 0$. We favor the two-parameter model over its one-parameter counterpart due to the former's flexibility. Though more information is needed to identify the parameters of a two-parameter model, we believe the flexibility gained in the response curve is beneficial to modeling the dose-toxicity distributions.

Despite the fact that we use a frequentist procedure, some 'prior information' is required in order to be able to fit the model, especially during the very early stages of the trial. We favor the 'prior' suggested by Piantadosi et al. (1998), namely using two dose levels, one representing a very low toxicity level and the other a very high toxicity level. In particular, we use dose levels that we speculate would correspond to 1% and 99% toxicity probability. Ideally one would like to estimate these two levels from either clinical information or from historical data but since neither is usually available, we use half the lowest proposed dose and twice the highest proposed dose as the lower and upper prior information, respectively. Since the logistic model requires data in triplets, i.e. dose, number of patients treated and the number of events at that dose level, we assume that we have treated 5 patients at each of the two 'prior' dose levels and that the expected number of failures have occurred. Note that by incorporating these values into the model we effectively tie the extremes of the logistic curve to these "prior" dose levels and let the observed data determine the shape of the curve between these two points. Note that since we employ a frequentist procedure, this 'prior information' is treated in the same manner as observed data in the calculations. Provided the prior dose levels are extreme enough this approach works quite well. We have performed extensive simulations (results not shown) experimenting with the location, the toxicity probability and the cohort size associated with these prior levels and observed that the priors selected above provide excellent operating characteristics both in terms of estimating the model parameters as well as in moderating their influence on the final outcome.

Once the priors are identified, following Goodman et al. (1995), we start at the first proposed dose, usually 80% of the adult MTD, and escalate one dose level at a time with no limitation

on dose de-escalation. Note however that we pragmatically designate a dose level 0, which is below the starting dose, to accommodate in the protocol the possibility that the starting dose level may be too toxic; thus reducing the risk of having to stop accrual in order to formally amend the protocol. During the trial, at each new dose level we make slots available for three patients and require that at least two patients are assigned before an escalation decision can be made.

As mentioned above, a certain level of heterogeneity is needed for the model parameters of the logistic function to become identifiable. We circumvent this problem to some extent via the use of the priors described above. However we have detected in our simulations that if no DLTs are observed in the first few doses, then convergence problems may still arise. Further, when we observe no DLTs in a group of patients, the estimate of the toxicity probability at that dose level is zero, which is likely an underestimate. Such underestimation could inflate the final MTD estimate. Thus to avoid the above-mentioned convergence problems as well as to counteract the possible overestimation of the target level, we use a "correction factor." Instead of inserting 0 for the number of DLTs, we record 0.1 for a cohort in which we did not observe any DLTs. The rationale for using 0.1 for the correction factor is detailed in Table I below, which lists the true toxicity probabilities that result in zero observed toxicities with probability greater than 0.90, for different sample sizes relevant for the CRM. In each of these cases, the expected number of DLTs in the cohorts are approximately 0.1. This correction is evenly distributed among all patients who are assigned to a given dose level and we use it until the first DLT is observed at <u>any</u> dose level, after which we discontinue its use, but leave the corrections already incorporated into the data untouched.

Another important issue concerns the stopping rule for the algorithm; i.e. when may one consider the MTD estimated? Several approaches have been proposed from using confidence (or Bayesian credibility) intervals to determine the precision of the estimate (O'Quigley et al., 1990; Heyd and Carlin, 1999); to fixing the maximum sample size (Goodman et al. 1995). Alternatively O'Quigley and Reiner (1998) put forward a stopping rule based on predicting the final dose level given the observations already accumulated and stopping the trial when the probability is large enough that the current dose will not change by the time the pre-set sample size has been reached. As a follow-up to this paper O'Quigley (2002), citing Korn et al. (1994), re-proposed using the idea that the trial will end when some fixed number of patients, say *m*, have been treated at any one of the pre-specified dose levels. The rationale for this approach is that the larger the number of patients the algorithm has assigned to a given dose level, the higher the probability that dose level is the MTD. O'Quigley (2002) recommended *m=5* as a reasonable choice. For our version, we combine these two approaches, modify them slightly and utilize the following rule to terminate the trial: at least 6 patients must be treated at the proposed MTD, and treating two more hypothetical patients at that dose level would not lead to escalation (we do not check for the possibility of de-escalation). Following Korn et al. (1994), who justified their choice based on simulations, we also use *m=6*. This value is further desirable since clinicians accustomed to traditional designs are comfortable with *m=6*. Our simulations presented below indicate that collectively our modifications lead to good operating characteristics.

4. SIMULATION RESULTS

We performed extensive simulation studies in an effort to ensure that our version of the CRM was at least as safe and as accurate as the TM, which is still the norm in pediatric oncology trials. Indeed the superiority of different versions of the CRM over the traditional method has been demonstrated in other studies but as outlined in the previous section our version of the CRM has some unique characteristics. To compare and contrast the performance of the TM to our version of the CRM, we simulated trials with various toxicity profiles. Below we present

an abbreviated version of our results for 6 distributions each with six dose levels that were studied by Goodman et al. (1995). Note that the version of the TM we utilize is different from the one discussed in Goodman et al. (1995), and thus our results vary from the TM results in that paper. The main difference is that we require 6 patients to be treated at the proposed MTD, which is typical for pediatric trials. This does lead to larger sample sizes yet, as a consequence, the accuracy is also favorably affected.

We simulated two versions of the CRM, one that uses complete cohorts and another which issues decisions based on information from incomplete cohorts. Specifically, the *incomplete-cohort* CRM dictates that at least two patients are assigned to a dose level, but once this requirement is satisfied, the DLT information from a single patient can be used to make escalation or de-escalation decisions. The information from the patients who are still on treatment is incorporated into the model as they become available. We experimented with the incomplete cohort design hypothesizing that, pending accrual patterns, some trials would be completed more quickly. This approach is similar to the one studied by Thall et al. (1999) in a Bayesian setting. We made the following simplifying assumption during the simulations: the DLT information from the patients who were still on treatment when the escalation/de-escalation decision was issued became available before any DLT information was observed from the patients assigned to the new dose.

As the target toxicity probability we used 20%, since this probability corresponds to the mean overall toxicity probability for the 6 distributions we consider here under our version of the TM. Note that only two of these six distributions contain a dose matching the 20% target toxicity level, and several distributions were specifically chosen to differ substantially from the logistic distribution and hence be unfavorable to our version of the CRM. The results presented below are based on 10,000 simulated trials for each distribution, and thus the half-width of the 90% simultaneous confidence intervals associated with the toxicity probabilities for the dose levels cannot exceed 1% (Fitzpatrick and Scott, 1987). Tables II–VII display the simulation results, which include the percentage of times each of the assigned dose levels was chosen as the MTD both for the TM and CRM. Note for dose levels 1 and 6, the values in parentheses indicate the percentage of realizations when the algorithm indicated that the associated dose was too high or too low to be the MTD, respectively. We also provide what Goodman et al. (1995) call the *experimentation percentage,* which tracks the proportion of patients treated at each dose level within a trial. The value reported in the tables for each dose is the mean across all trials. Our intent in tracking this value was to assess the concern that the CRM may expose more patients to potentially toxic levels. We also provide the median and range of the sample sizes observed across 10,000 simulations as well as the median and the range of the percent toxicity (percentage of patients experiencing a DLT) that resulted from the various approaches to dose-finding that were of interest. Finally, we provide in the last two rows of the tables, the percentage of trial realizations in which a dose level with ≥ 2 DLTs was revisited and the percentage of realizations that resulted in at least three patients experiencing DLTs in at least one dose level. Since it cannot happen with the TM, revisiting a dose level at which two patients have already experienced a DLT is a concern both for our clinical colleagues and for the regulatory bodies by whom our protocols are reviewed such as the Cancer Therapy Evaluation Program (CTEP) of the NCI.

4.1 Comparisons of the Two Approaches within the Modified CRM: Complete versus Incomplete Cohorts

There does not seem to be an appreciable difference between the complete and the incomplete cohort versions of the CRM with respect to accuracy, overall toxicity, or the median sample size needed to declare the MTD. However, the incomplete cohort version seems to expose slightly more patients to dose levels above the target. Further uniformly across the six toxicity

distributions, this approach led to notably higher proportions of trial realizations that revisited dose levels at which 2 or more DLTs were already observed. Additionally, trials simulated with the incomplete cohort design had considerably higher proportions of dose levels at which ≥ 3 DLTs were observed. Similar observations regarding their version of the incomplete cohort design were also noted by Thall et al. (1999).

4.2 Comparison of Modified CRM with Complete Cohorts versus TM

Given the observations above, in the sequel we only discuss the results which compare the TM with the complete cohort CRM.

For our simulations we defined accuracy as the algorithm choosing the dose level with a toxicity probability closest to the target value. Once again note that only 2 of the 6 distributions presented below contain a level that corresponds to the target toxicity level. As tables II and III clearly indicate, for distributions 1 and 2, which include a dose level with the target toxicity probability (20%), the modified CRM selects the correct dose considerably more frequently than the TM, where the observed difference is 8.5% and 6.8%, respectively. For distributions 3 and 4 for which the target toxicity dose is between dose levels 4 and 5, both the TM and the modified CRM select these doses as the MTD approximately the same proportion of times. For distributions 5 and 6, where all dose levels are either entirely above the target level or are entirely below the target level, the TM seems to make the correct decision more frequently.

With respect to the sample size, with the exception of distribution 5, the modified CRM uses 3–4 fewer patients as measured by the median of the 10,000 simulated trials. For distribution 5 the medians are the same. Since both approaches start from the lowest dose, naturally the sample size is affected by the location of the target dose within the proposed levels. Regardless, the fact that our modified CRM requires 3–4 fewer patients to complete a trial is clinically significant in pediatric oncology.

The ranges of the sample sizes observed from the simulated trials indicate that the tail of the sample size distribution can be somewhat long when the CRM is employed. This is mainly due to the constraint that once 6 or more patients are treated at a dose, we do not stop the trial unless treating two more hypothetical patients at that dose level would not lead to escalation. For some realizations, it takes a while to satisfy this constraint. Another contributor to the elevation in sample size is the oscillation that can occur if the target dose is approximately equidistant between two proposed dose levels. Our PBTC trials often incorporate a provision to study a dose level at the midpoint of two proposed levels in such cases. Out of the 10,000 trial realizations for each of the six 6-level distributions, the percentage of realizations for which the complete-cohort CRM sample size was larger than the maximum sample observed for the traditional method was less than 0.2% for distributions 1–5 and less than 2.2% for distribution 6. Thus, although larger sample sizes can be observed for our version of the CRM, the incidence is quite rare.

In our simulations we observed that the modified CRM had the smaller median proportion of patients experiencing DLTs for distributions 1–4 that cover the target dosage, but had a 2.4% higher median toxicity percentage for distribution 5, and was essentially the same for distribution 6. With respect to the percentage of simulations resulting in at least one dose level where ≥ 3 patients experienced a DLT however, the modified CRM was uniformly better than the TM across all six distributions. For the TM, these percentages were 27%, 16.4%, 33.2%, 40.0%, 29.8% and 3.6%, respectively for distributions 1–6; whereas the corresponding values for the modified CRM were 15.4%, 9.7%, 12.5%, 18.7%, 29.1% and 3.2%. For distributions 1–4, where the MTD was within the proposed dose levels, the CRM treated smaller percentages of patients above the estimated MTD as compared with the TM. In the tables, this information

is captured by the 'experimentation percentage' column which gives mean number of patients treated at each dose level across the simulated trials.

The effect of requiring eight rather than 6 patients to be treated at the MTD was also studied via simulations. Although this approach led to slight improvements in the accuracy, the observed differences were not large enough to justify the increase in the sample size as well as in the duration of the trial.

In addition, we also investigated the performance of the algorithm using a target toxicity level of 25% for this is the actual target we utilize in our PBTC trials. As stated above our extensive simulations were based on 20% since this probability is the average overall toxicity probability for the 6 distributions presented here under our version of the TM. Furthermore, 20% was utilized in the literature for these distributions and thus preserving this target level facilitates comparisons with other published results. We were interested in observing the effect of this shift in the target however and the results remained quite favorable for the CRM. The median sample size increased by 1 in 4 out of the 6 cases; stayed the same for distribution 4 and decreased by 1 in distribution 6. The accuracy was as good or better compared to the simulations where the target was chosen to be 20% and the median percent toxicity did not increase by more than 3% in any of the 6 distributions studied.

In their recent paper Lee et al. (2005) observed that for cytotoxic drugs there is a strong correlation between the adult and the pediatric MTDs obtained via traditional designs. As a result, they recommended that no more than 4 dose levels scattered around the adult MTD be studied in pediatric Phase I trials. To explore the operational characteristics of the modified CRM in this case, we repeated the simulations for 4-dose level distributions. In these simulations only the complete-cohort CRM was used due to reasons outlined in section 4.1. The distributions which were studied are given in table VIII, and the simulation results are abbreviated in table IX. As it is evident from these results, for the distributions tested here, the CRM uniformly outperformed the traditional method in terms of accuracy, i.e. choosing the dose level with toxicity probability closest to the target level, and generally used fewer patients. The observed improvement in accuracy was between 3.0% and 17.7%. As was the case for the 6-dose-level distributions, the toxicity percentages favored the CRM and were never worse than the toxicity percentages observed for the TM. We have also simulated 3-dose level distributions with favorable results; however in cases where very few dose levels are studied it is more difficult to justify fitting a 2-parameter model and the influence of the "prior values" we use is more pronounced.

5. DISCUSSION

The intent of this manuscript was to describe our modifications to the CRM and share our simulation results demonstrating the favorable operating characteristics of this version of the algorithm. In line with previously published results, our version of the CRM, which heavily borrows from previous research done in this area, also outperforms the TM. Our experiences with modified CRM designs for Phase I trials in the Pediatric Brain Tumor Consortium and our simulation-based investigations of the operating characteristics of these competing approaches provide additional credence that a shift in this paradigm may result in more accurate and efficient designs.

If one seeks to identify a dose level at which a specified proportion of patients would be expected to experience DLTs, our version of the CRM seems to out-perform the TM in terms of accuracy while also requiring fewer patients. Our complete-cohort version also has better performance as compared to the versions of the CRM presented in Goodman et al. (1995). Only when the true MTD is outside the range of the proposed doses does the TM appear to

perform better than the CRM – and this is not typical in pediatric Phase I investigations since they usually begin at 80% of the adult MTD. Our simulations involving distributions with four pre-specified dose levels also suggest that the advantages of using the CRM over the TM are preserved. There is very little loss regarding the accuracy or median toxicity probabilities of the design when the incomplete cohorts approach is utilized; however this approach tends to be more aggressive and could lead to treating slightly more patients at dose levels above the MTD as well as increase the chance of revisiting dose levels where 2 or more DLTs have already been observed. In the interest of maximizing safety, we prefer the complete cohort CRM, even if this means the duration of the trials may be somewhat longer.

In a few of the PBTC trials, we have had to restrict our CRM design from revisiting a dose level at which two patients had already experienced DLTs. Our clinical colleagues insisted on this restriction arguing that this could not happen in the traditional Phase I design. Understandably, their concern is to limit exposing children to potentially toxic doses, but as our simulations of the CRM free of such a constraint show, there is little justification for this restriction. While the TM may never return to a dose level at which two previous patients experienced a DLT, our version of the CRM only infrequently does that. Further, on average the TM does not afford better protection for patients compared to the CRM in terms of the number of DLTs per dose level.

Another major advantage of model based designs such as the CRM is the fact that they can incorporate the actual dose levels studied. As mentioned previously the pre-specified dose levels are often separated by 30% increments in pediatric trials, which implies larger differences in mg/m^2 (the unit of dosing in pediatric studies) between dose level 5 and 6 compared to dose level 1 and 2. It makes sense in these cases to require more evidence of safely to move from dose level 5 to 6 then to move from dose level 1 to 2. Additionally, incorporating the dose received by a patient in the model provides the opportunity to use data from patients who may have received different doses from what was targeted, for example due to pill-size limitations or incorrect dosing. Similarly, model-based algorithms can be used in more general early-phase trial settings where other patient-specific measures such as peak concentration, area under the curve, etc. are used instead of dose (Piantadosi and Liu, 1996). Designs such as the CRM can easily accommodate such approaches while TM cannot.

In conclusion, aligned with the results previously published in the literature, we believe our modified continual reassessment method is a superior pediatric Phase I design compared to the empirically based traditional method. The CRM provides a well-defined MTD; tends to expose fewer patients to potentially toxic dose levels and tends to require fewer patients.

Acknowledgments

References

Fitzpatrick S, Scott A. Quick simultaneous confidence intervals for multinomial proportions. JASA 1987;82:875–878.

Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for Phase I studies. Statistics in Medicine 1995;14:1149–1161. [PubMed: 7667557]

Heyd J, Carlin B. Adaptive design improvements in the continual reassessment method for phase I studies. Statistics in Medicine 1999;18:1307–1321. [PubMed: 10399198]

Kasahara K, Myo S, Iwasa K, Kimura H, Shirasaki H, Yasuda U, Shibata K, Shintani H, Nishi K, Fujimura M, Nakao S. A phase I study of carboplatin and docetaxel for advanced non-small cell lung cancer using the continual reassessment method. Japanese Journal of Clinical Oncology 2002;32:512–516. [PubMed: 12578899]

Kieran MW, Packer RJ, Onar A, Blaney SM, et al. Phase I and pharmacokinetic study of the oral farnesyltransferase inhibitor lonafarnib administered twice daily to pediatric patients with advanced central nervous system tumors using a modified continuous reassessment method: a pediatric brain tumor consortium study. Journal of Clinical Oncology 2007;25:3137–43. [PubMed: 17634493]

Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A Comparison of two Phase I trial designs. Statistics in Medicine 1994;14:1799–1806. [PubMed: 7997713]

Lee DP, Skolnik JM, Adamson PC. Pediatric Phase I Trials in Oncology: An Analysis of Study Conduct Efficiency. Journal of Clinical Oncology 2005;23:8431–8441. [PubMed: 16293874]

Lin Y, Shih WJ. Statistical properties of the traditional algorithm-based designs for Phase I cancer clinical trials. Biostatistics 2001;2:203–215. [PubMed: 12933550]

Marsoni S, Ungerleider RS, Hurson SB, Simon RM, Hammershaimb LD. Tolerance to antineoplastic agents in children and adults. Cancer Treat Reports 1985;69:1263–1269.

O'Quigley J, Pepe M, Fisher L. Continual Reassessment Method: A practical design for Phase I clinical studies in cancer. Biometrics 1990;46:33–48. [PubMed: 2350571]

O'Quigley J, Shen LZ. Continual Reassessment Method: A likelihood approach. Biometrics 1996;52:673–684. [PubMed: 8672707]

O'Quigley J. Continual reassessment designs with early termination. Biostatistics 2002;3:87–99. [PubMed: 12933626]

O'Quigley J, Rainer E. A stopping rule for the continual reassessment method. Biometrika 1998;85:741–748.

Piantadosi S, Fisher JD, Grossman S. Practical implementation of a modified continual reassessment method for dose finding trials. Cancer Chemotherapy and Pharmacology 1998;41:429–436. [PubMed: 9554585]

Piantadosi S, Liu G. Improved designs for dose escalation studies using pharmacokinetic measurements. Statistics in Medicine 1996;15:1605–18. [PubMed: 8858785]

Royce ME, Hoff PM, Dumas P, Lassere Y, Lee JJ, Coyle J, Ducharme MP, De Jager R, Pazdur R. Phase I and pharmacokinetic study of exatecan mesylate (DX-8951f): a novel camptothecin analog. Journal of Clinical Oncology 2001;19:1493–1500. [PubMed: 11230496]

Schoffski P, Riggert S, Fumoleau P, Campone M, Bolte O, Marreaud S, Lacombe D, Baron B, Herold M, Zwierzina H, Wilhelm-Ogunbiyi K, Lentzen H, Twelves C. Phase I trial of intravenous aviscumine (rViscumin) in patients with solid tumors: a study of the European Organization for Research and Treatment of Cancer New Drug Development Group. Annals of Oncology 2004;15:1816–1824. [PubMed: 15550588]

Shen LZ, O'Quigley J. Consistency of continual reassessment method under model misspecification. Biometrika 1996;83:395–405.

Silvapulle MJ. On the existence of maximum likelihood estimators for the binomial response models. JRSS-B 1981;43:310–313.

Smith M, Bernstein M, Bleyer WA, Borsi JD, Ho P, Lewis IJ, Pearson A, Pein F, Pratt C, Reaman G, Riccardi R, Seibel N, Trueworthy R, Ungerleider R, Vassal G, Vietti T. Conduct of phase I trials in children with cancer. Journal of Clinical Oncology 1998;16:966–978. [PubMed: 9508179]

Storer B. Design and analysis of Phase I clinical trials. Biometrics 1989;45:925–937. [PubMed: 2790129]

Thall PF, Lee JJ, Tseng C, Etsey EH. Accrual Strategies for Phase I Trials with Delayed Patient Outcome. Statistics in Medicine 1999;18:11551169.

Table I

True toxicity probabilities that result in zero observed toxicities with probability greater than 0.90 for a variety of sample sizes

Number of Patients	Observed DLTs	Probability that gives 0 DLTs with prob. >0.90	Expected # of DLTs
N		P	
2	0	0.051	0.102
3	0	0.035	0.105
4	0	0.026	0.104
5	0	0.021	0.105
6	0	0.017	0.102

Table II

Simulation Results Based on 10,000 Realizations for Distribution 1

Dose	Actual Toxicity Probability	Traditional Method		CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts		CRM with Escalation/De-escalation Decisions Based on *Incomplete* Cohorts	
		Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage
1	5%	9.9% (2.5%)*	24.6%	4.1% (0.63%)*	20.8%	3.5% (0.58%)*	20.0%
2	10%	28.8%	27.2%	24.3%	29.5%	22.9%	28.5%
3	**20%**	**38.3%**	**26.3%**	**46.8%**	**30.5%**	**47.9%**	**30.8%**
4	35%	17.3%	16.2%	22.0%	15.5%	22.9%	16.4%
5	50%	3.2%	5.0%	2.1%	3.4%	2.2%	4.0%
6	70%	0.0%	0.7%	0.0	0.3%	0.04%	0.3%
Median Sample Size (Range)		18 (3–36)		15 (2–37)		16 (2–35)	
Median % Toxicity (Range)		19.1%(5–67)		18.2% (0–100%)		18.8% (0–100%)	
% of Trials Revisiting a Dose with ≥ 2 DLTs		-		14.5%		31.3%	
% of Trials with ≥ 3 DLTs at a Dose		27%		15.4%		19.5%	

Table III

Simulation Results Based on 10,000 Realizations for Distribution 2

Dose	Actual Toxicity Probability	Traditional Method		CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts		CRM with Escalation/De-escalation Decisions Based on *Incomplete* Cohorts	
		Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage
1	5%	9.5% (2.8%) *	22.8%	3.8% (0.6%) *	19.3%	3.4% (0.5%) *	18.5%
2	10%	16.6%	22.8%	17.9%	25.8%	15.4%	24.1%
3	15%	21.6%	20.8%	28.4%	24.9%	27.8%	24.8%
4	**20%**	**20.6%**	**16.1%**	**27.4%**	**17.7%**	**28.9%**	**18.8%**
5	25%	19.0%	11.1%	16.9%	9.3%	18.5%	10.5%
6	35%	9.9%	6.4%	2.9%	3.0%	3.4%	3.3%
Median Sample Size (Range)		21 (3–36)		17 (2–43)		17 (2–43)	
Median % Toxicity (Range)		16.7% (0–100)		15.4% (0–100%)		15.4% (0–100%)	
% of Trials Revisiting a Dose with ≥ 2 DLTs				12.2%		28.5%	
% of Trials with ≥ 3 DLTs at a Dose		16.4%		9.7%		13.3%	

*
Values in parentheses represent percentage of cases where the MTD was deemed to be below dose level 1

Table IV

Simulation Results Based on 10,000 Realizations for Distribution 3

Dose	Actual Toxicity Probability	Traditional Method		CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts		CRM with Escalation/De-escalation Decisions Based on *Incomplete* Cohorts	
		Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage
1	10%	8.7% (9.6%)*	27.7%	5.5% (2.9%)*	21.3%	5.5% (2.6%)*	20.8%
2	10%	7.8%	19.0%	12.2%	22.5%	10.8%	21.1%
3	10%	6.7%	15.4%	14.1%	19.2%	13.7%	19.3%
4	10%	30.3%	15.8%	31.3%	19.7%	31.8%	20.6%
5	**25%**	**36.9%**	**16.4%**	**33.8%**	**14.6%**	**35.6%**	**15.3%**
6	80%	0.0%	5.7%	0.1%	2.7%	0.1%	2.9%
Median Sample Size (Range)		21 (3–36)		18 (2–45)		19 (2–47)	
Median % Toxicity (Range)		16.7% (8–100)		14.3% (0–100%)		15% (0–100%)	
% of Trials Revisiting a Dose with ≥ 2 DLTs		-		10.0%		25.5%	
% of Trials with ≥ 3 DLTs at a Dose		33.2%		12.5%		16.5%	

*Values in parentheses represent percentage of cases where the MTD was deemed to be below dose level 1

Table V

Simulation Results Based on 10,000 Realizations for Distribution 4

Dose	Actual Toxicity Probability	Traditional Method		CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts		CRM with Escalation/De-escalation Decisions Based on *Incomplete* Cohorts	
		Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage
1	1%	0.0%	14.9%	0.0%	13.2%	0.04%	12.9%
2	1%	2.6%	15.3%	0.8%	14.2%	0.9%	14.0%
3	5%	9.7%	17.8%	7.8%	18.4%	7.4%	18.3%
4	10%	39.1%	21.8%	40.5%	26.8%	40.6%	27.4%
5	**25%**	**48.3%**	**22.5%**	**50.7%**	**22.6%**	**51.0%**	**22.8%**
6	80%	0.0	7.8%	0.1%	4.8%	0.1%	4.8%
Median Sample Size (Range)			21 (6–33)		19 (3–40)		20 (3–42)
Median % Toxicity (Range)			14.3% (5–33%)		13.3% (6–67%)		13.6% (0–67%)
% of Trials Revisiting a Dose with ≥ 2 DLTs			-		10.2%		24.7%
% of Trials with ≥ 3 DLTs at a Dose			40.0%		18.7%		21.2%

Table VI

Simulation Results Based on 10,000 Realizations for Distribution 5

Dose	Actual Toxicity Probability	Traditional Method		CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts		CRM with Escalation/De-escalation Decisions Based on *Incomplete* Cohorts	
		Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage
1	**30%**	**32.5% (55.1%)***	**74.3%**	**43.5% (28.4%)***	**62.9%**	**44.8% (28.0)***	**61.9%**
2	40%	10.8%	20.5%	25.3%	30.9%	24.0%	30.5%
3	52%	1.7%	5.0%	2.6%	5.7%	3.1%	6.8%
4	61%	0%	0.6%	0.2%	0.5%	0.2%	0.8%
5	76%	0%	0.0%	0.0%	0.0%	0.0%	0.0%
6	87%	0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Median Sample Size (Range)	9 (3–27)		9 (2–26)		10 (2–29)	
	Median % Toxicity (Range)	33.3% (11–100%)		35.7% (11–100%)		37.5% (12–100%)	
	% of Trials Revisiting a Dose with ≥ 2 DLTs	-		24.7%		43.1%	
	% of Trials with ≥ 3 DLTs at a Dose	29.8%		29.1%		36.2%	

Table VII

Simulation Results Based on 10,000 Realizations for Distribution 6

Dose	Actual Toxicity Probability	Traditional Method		CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts		CRM with Escalation/De-escalation Decisions Based on *Incomplete* Cohorts	
		Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage	Percent Chosen	Experimentation Percentage
1	5%	2.7% (2.7%) *	18.0%	0.9% (0.5%) *	14.2%	1.1% (0.5%) *	14.2%
2	5%	2.6%	15.7%	4.0%	16.3%	3.2%	15.5%
3	5%	2.4%	14.8%	5.6%	16.2%	4.7%	15.8%
4	5%	8.6%	14.8%	11.2%	17.3%	9.9%	17.2%
5	10%	17.7%	16.3%	27.5%	19.6%	25.7%	20.1%
6	**15%**	**63.6%**	**20.5%**	**17.0% (33.2%)** #	**16.4%**	**17.5% (37.4%)** #	**17.3%**
	Median Sample Size (Range)	24 (3–33)		20 (2–51)		21 (2–69)	
	Median % Toxicity (Range 0	8.3% (0–67%)		8.5% (0–100%)		8.7% (0–100%)	
	% of Trials Revisiting a Dose with ≥ 2 DLTs	-		6.5%		16.1%	
	% of Trials with ≥ 3 DLTs at a Dose	3.6%		3.2%		4.8%	

* Values in parentheses represent percentage of cases where the MTD was deemed to be below dose level 1

Values in parentheses represent percentage of cases where the MTD was deemed to be above dose level 6

Table VIII

Dose-Toxicity Distributions Studied for the 4-Dose Level Cases

	4-Dose Level Distributions					
Dose	D1	D2	D3	D4	D5	D6
150	0.05	0.25	0.15	0.10	0.01	0.05
200	0.25	0.45	0.20	0.25	0.05	0.10
265	0.60	0.70	0.25	0.50	0.12	0.25
350	0.99	0.99	0.30	0.99	0.25	0.40

The values in columns titled D1–D6 represent DLT probabilities for each dose level for the 6 distributions.

Table IX

Abbreviated Simulation Results Based on 10,000 Realizations for 4-Dose Level Distributions

| Dist | Closest Toxicity | Traditional Method | | | | | CRM with Escalation/De-escalation Decisions Based on *Complete* Cohorts | | | |
		Percent Chosen	Sample Size Median (Range)	Percent Toxicity Median (Range)	% of Trials with ≥ 3 DLTs at a dose	Percent Chosen	Sample Size Median (Range)	Percent Toxicity Median (Range)	% of Trials with ≥ 3 DLTs at a Dose
1	25%	52.0%	12 (3–21)	25% (13–100%)	36.62%	69.7%	12 (2–27)	23% (8–100%)	25.13%
2	25%	44.6%	9 (3–21)	33% (13–100%)	32.79%	47.6%	9 (3–23)	33% (8–100%)	19.76%
3	20%	22.7%	15 (3–24)	22% (0–100%)	15.66%	30.6%	11 (3–24)	21% (0–100%)	10.01%
4	25%	42.9%	12 (3–24)	25% (13–100%)	35.15%	57.1%	12 (3–20)	25% (8–100%)	21.22%
5	25%	45.7%	15 (3–24)	11% (0–67%)	9.10%	51.1%	13 (2–30)	10% (0–100%)	6.01%
6	25%	35.5%	15 (3–24)	19% (0–67%)	20.49%	46.1%	13 (2–27)	17% (0–100%)	12.41%