BIOMETRICS 58, 671–674
September 2002

A Simple Technique to Evaluate Model Sensitivity in the Continual Reassessment Method

Ying Kuen Cheung

Department of Biostatistics, Mailman School of Public Health,
Columbia University, 630 West 168th Street, New York, New York 10032, U.S.A.
email: cheung@biostat.columbia.edu

and

Rick Chappell

Department of Statistics and Department of Biostatistics and Medical Informatics,
University of Wisconsin, 1210 Dayton Street, Madison, Wisconsin 53706, U.S.A.

SUMMARY. The continual reassessment method (CRM) is a sequential design used in phase I cancer trials to determine the maximal dose with acceptable toxicity. It has been established that the CRM is consistent under model misspecification but not generally. When the method does not converge to the target percentile, some dose–response models will be more sensitive than others in terms of how close the converged recommendation is to the target. In this article, we interpret the main condition under which the CRM is consistent and apply it to evaluate the sensitivity of the model used with the CRM. The technique presented is found to be a useful supplement to simulation when planning a phase I trial.

KEY WORDS: Continual reassessment method; Indifference interval; Model sensitivity.

1. Introduction

The primary objective of phase I clinical trials in cancer is to determine a maximal dose that does not exceed an acceptable level of toxicity, i.e., the maximum tolerated dose (MTD). The continual reassessment method (CRM), a sequential design for phase I studies, has received much attention since its first proposal (O'Quigley, Pepe, and Fisher, 1990). The method uses a simple model to describe the curve that delineates the stochastic relationship between dose and probability of toxicity. The model parameter is estimated repeatedly throughout the trial as binary toxicity observations are accrued. The CRM then assigns the next patient to a dose whose toxicity probability is estimated to be closest to the target probability. The two most frequently used dose–response models, a one-parameter logistic function with a known intercept and a power function, are examined by Chevret (1993) via simulation. Since then, not much, if any, effort has been put on the choice of dose–response model. It is probably because the performance of the CRM is believed to be robust against model misspecification and therefore practitioners deem the conventional choice adequate.

In this article, we illustrate that some models are less sensitive than others in terms of how close the converged recommendation is to the desired dose. A numerical technique will be presented to evaluate the model sensitivity in the CRM. The technique exploits a condition that suffices the method's consistency. While the consistency results were established (Shen and O'Quigley, 1996), few insights had been derived, partly due to the opaqueness of the condition. This article interprets the main condition in Section 3 and then presents the technique in Section 4. To start, we review the CRM in the following section.

2. The CRM

In a phase I trial with doses $d_1, \ldots, d_K$ on trial, the CRM models toxicity probability via a single-parameter model $F(d; \beta)$. This dose–response model F should be monotone increasing in d and monotone in the parameter β; also, the model should be flexible enough to reproduce the target probability p_T at any dose levels. Other than these, choice of F should respect a few regularity conditions (see Appendix A). Let Y_i be the indication of toxic response for the ith patient who receives dose $d_{[i]}$. With the first n observations, we estimate the model parameter β by maximizing the likelihood

$$L(\beta) = \prod_{i=1}^{n} \left\{F(d_{[i]}; \beta)\right\}^{Y_i} \left\{1 - F(d_{[i]}; \beta)\right\}^{1-Y_i}$$

and assigning the next dose level $[n+1]$ such that

$$|F(d_{[n+1]}; \hat{\beta}_n) - p_T| \leq |F(d_k; \hat{\beta}_n) - p_T| \quad \text{for } k = 1, \ldots, K,$$

where $\hat{\beta}_n$ is the maximizer of L. This maximum-likelihood approach of the CRM is proposed by O'Quigley and Shen (1996)

and is shown to be consistent under model misspecification by Shen and O'Quigley (1996).

3. Sufficient Conditions for Consistency

In this section, we state and interpret the main condition under which the CRM is consistent. We refer readers to Shen and O'Quigley (1996) for proof of consistency. First, let $F_k(\beta) = F(d_k;\beta)$ for $k = 1,\ldots,K$ and define for level j

$$H_j = \{\beta \in \Theta : |F_j(\beta) - p_T| < |F_k(\beta) - p_T| \text{ for } k \neq j\},$$

where the parameter space Θ is assumed to be a closed finite interval $[b_1, b_{K+1}]$. We show in Appendix B that $H_1 = [b_1, b_2)$, $H_k = (b_k, b_{k+1})$ for $k = 2,\ldots,K-1$ and $H_K = (b_K, b_{K+1}]$, where b_k solves

$$F_{k-1}(b_k) + F_k(b_k) = 2p_T \quad \text{for } k = 2,\ldots,K. \tag{3.1}$$

Note that the CRM will recommend dose level j if and only if $\hat{\beta}_n \in H_j$.

Further define $\phi_k = F_k^{-1}(\mu_k)$ for $k = 1,\ldots,K$, where μ_k is the true toxicity probability associated with d_k. If F is a correct model, $\phi_k = \beta_0$ for all k for some true parameter β_0. Consistency requires a less restrictive condition,

(C1) $\phi_k \in H_l$ for all k, where l is the correct dose level.

For example, suppose we set the target probability p_T to be .20 and use the power model $F(d;\beta) = d^\beta$ for $\beta \in [0, 10]$ at doses .05, .10, .20, .30, .50, and .70 (after rescaling). The upper limit 10 of the parameter space is chosen arbitrarily to avoid technical difficulty; any large number shall suffice. Then we can calculate the sets

$$\begin{aligned} H_1 &= [0.00, 0.62), \\ H_2 &= (0.62, 0.84), \\ H_3 &= (0.84, 1.16), \\ H_4 &= (1.16, 1.80), \\ H_5 &= (1.80, 3.35), \\ H_6 &= (3.35, 10.0] \end{aligned}$$

according to equation (3.1). Now suppose the correct level is level 3. Condition (C1) requires $\phi_k \in (0.84, 1.16)$ for all k or, equivalently, $\mu_k \in \{F(d_k;\beta) : \beta \in (0.84, 1.16)\}$, which is an interval for each k. A graphical representation is illustrated in the left panel of Figure 1: The condition is satisfied if the true dose–response curve crosses all six vertical bars in the shaded area.

It is intuitive that a reasonable design will perform well when the true curve is steep around the MTD (Storer, 1989). Therefore, we further postulate that consistency will hold under a more relaxed condition,

(C2) $\phi_l \in H_l$, $\phi_k \in \cup_{j=k+1}^{K} H_j$ for $k = 1,\ldots,l-1$, $\phi_k \in \cup_{j=1}^{k-1} H_j$ for $k = l+1,\ldots,K$.

Graphically, condition (C2) is satisfied for $l = 3$ if the true monotone increasing curve crosses all six vertical bars in the right panel of Figure 1.

The above design setup for the CRM was considered by O'Quigley et al. (1990) and Cheung and Chappell (2000). Both reported that the CRM had mediocre performance under the scenario where toxicity probabilities at the six doses were .00, .00, .03, .05, .11, .22; in addition, the method showed little improvement in accuracy as sample size grew from 25 to 48 (Cheung and Chappell, 2000, Table 2). Shen and O'Quigley (1996) conduct simulations to verify failure in convergence to the MTD (dose level 6), positing that such failure is due to the violation of condition (C1). We here observe that condition (C2) is also violated: In this case, both conditions do not hold because $\phi_5 = \log(.11)/\log(.50) = 3.18 \notin (3.35, 10.0] = H_6$.

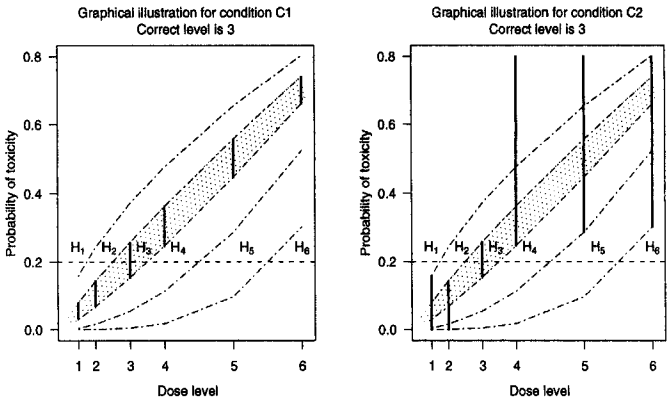

Figure 1. If the true monotone increasing dose–response curve crosses all six vertical bars in the left and right panels, conditions (C1) and (C2) are, respectively, satisfied with the power model $F(d;\beta) = d^\beta$ at doses .05, .10, .20, .30, .50, .70 and $p_T = .20$.

Consider yet another scenario with toxicity probabilities .06, .08, .12, .18, .40, .71 (d_4 is the MTD) where condition (C1) is not obeyed because $\phi_6 = 0.96 \notin (1.16, 1.80) = H_4$. However, convergence did occur, as with simulations of larger trials. In this case, condition (C2), requiring $\phi_6 \in (0, 3.35)$, is satisfied and thus appears to be a more reliable indication for consistency of the method.

4. Application and Discussion

In reality, the true curve is fixed, and we would like to choose F so as to ensure certain nice estimation properties, such as convergence of the recommended dose to the correct dose d_l. However, the true curve is unknown and cannot be used to verify condition (C2). A reasonable approach then is to determine, for a given model F, an interval of probabilities in which the converged recommended dose will fall. The shorter the range, the less sensitive is the model F to the underlying truth and vice versa.

Reconsider the CRM design setup in Section 3. If we knew that level 3 is the correct dose, condition (C2) would have required $\phi_1 \in (0.62, 10]$, $\phi_2 \in (0.84, 10]$, $\phi_3 \in (0.84, 1.16)$, $\phi_4 \in (0, 1.16)$, $\phi_5 \in (0, 1.80)$, and $\phi_6 \in (0, 3.35)$, which is equivalent to

$$\begin{aligned} \mu_1 &\in (0, .16), \\ \mu_2 &\in (0, .14), \\ \mu_3 &\in (.15, .26), \\ \mu_4 &\in (.25, 1), \\ \mu_5 &\in (.29, 1), \\ \mu_6 &\in (.30, 1). \end{aligned} \tag{4.1}$$

Table 1

Indifference intervals for various CRM setups. (a) Section 3: $F(d;\beta) = d^{\beta}$ at doses 0.05, 0.10, 0.20, 0.30, 0.50, 0.70 and $p_T = 0.20$. (b) Chevret (1993): $\text{logit}\{F(d;\beta)\} = 1 + \beta d$ at doses −3.94, −3.20, −2.10, −1.62, −1.00, −0.15 and $p_T = 0.25$. (c) Chevret (1993): $\text{logit}\{F(d;\beta)\} = 3 + \beta d$ at doses −5.94, −5.20, −4.10, −3.62, −3.00, −2.15 and $p_T = 0.25$. (d) Gasparini and Eisele (2000): $F(d;\beta) = d^{\beta}$ at doses 0.05, 0.10, 0.15, 0.25, 0.35, 0.45, 0.60, 0.80 and $p_T = 0.25$.

	Interval			
l	(a) $p_T = .20$	(b) $p_T = .25$	(c) $p_T = .25$	(d) $p_T = .25$
1	[−, .242]	[−, .291]	[−, .301]	[−, .295]
2	[.158, .257]	[.209, .331]	[.199, .339]	[.205, .283]
3	[.143, .246]	[.169, .301]	[.161, .297]	[.217, .304]
4	[.154, .286]	[.199, .342]	[.203, .321]	[.196, .298]
5	[.114, .302]	[.158, .496]	[.179, .371]	[.202, .297]
6	[.098, −]	[.004, −]	[.129, −]	[.203, .326]
7				[.174, .387]
8				[.113, −]

On the one hand, the CRM will converge to the MTD (d_3) if the true toxicity probabilities fall in the intervals specified in (4.1), particularly $\mu_2 \in (0, .14)$ and $\mu_4 \in (.25, 1]$. On the other hand, if μ_2 is close, though not as close as μ_3, to $p_T = .20$, we expect that the CRM may recommend level 2. To be more precise, condition (C2) is violated when $\mu_2 \in [.14, \mu_3)$. Likewise, if $\mu_4 \in (\mu_3, .25]$, the CRM may converge to level 4. We shall call [.14, .25] the indifference interval of model F for $l = 3$ because the CRM with F may fail to distinguish level 3 from its neighbor levels whose toxicity probabilities fall in this interval. Because the CRM tends to choose a level close to the correct level, if not the correct level itself, we may ignore the nonneighbor levels.

Table 1, column (a), shows the indifference intervals of F when we assume $l = 1, \ldots, 6$ under this design setup. From the table, we deduce that the CRM will recommend a dose that is somewhere between the 10th and 30th percentiles eventually and hereby solicit from the clinicians whether this range is acceptable, i.e., not too wide. This notion is analogous to the idea of minimal relevance difference in the context of hypothesis testing. We note that the CRM can recommend a dose outside the [10, 30]th percentile range either when all the doses are way too low or when all are very toxic. Nevertheless, it holds no implication for the model sensitivity. Rather, this issue should be taken care of when selecting doses for experimentation.

Example 1. Chevret (1993) examines via simulation the sensitivity of the CRM with a one-parameter logistic model,

$$F(d;\beta) = \frac{\exp(a_0 + \beta d)}{1 + \exp(a_0 + \beta d)},$$

where a_0 is fixed. Before running a simulation, we could have analyzed the method's sensitivity with the indifference intervals of the model. Tables 1, columns (b) and (c), shows the intervals with $a_0 = 1$ and $a_0 = 3$, respectively. The latter was recommended based on simulations. In contrast, the CRM using the model with $a_0 = 1$ may recommend a virtually nontoxic, and probably ineffective, dose if dose level 6 is the MTD. Thus, we may rule out this model without spending computing time on simulations. Compared with simulations, calculation of indifference intervals is easy to program and much faster to run.

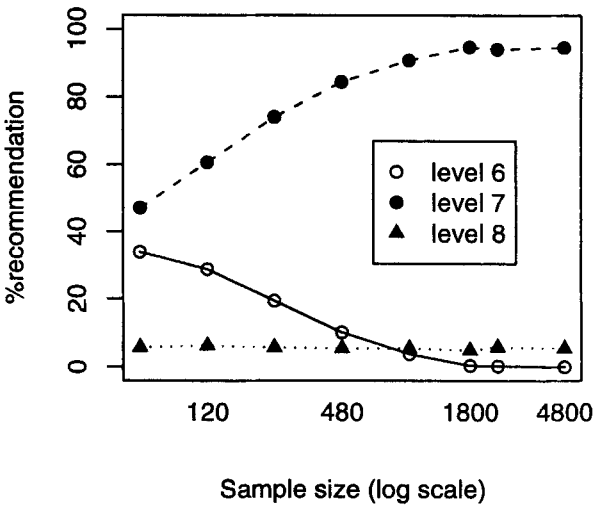

Figure 2. The percentages of recommending levels 6, 7, and 8 by the CRM with the power model used in Gasparini and Eisele (2000) versus sample size. The results are based on 2000 simulated trials. Each trial starts at the lowest level, does not allow skipping levels in escalation, and takes in three patients at a time until a fixed sample size is reached.

Example 2. Gasparini and Eisele (2000) propose an alternative design to the CRM for phase I trials. The authors compared their method to the CRM with the power model via simulation where the target $p_T = .25$. In all six scenarios considered in their article, the CRM was comparable with their method except in scenario 4, where true toxicity probabilities are .01, .05, .10, .10, .15, .15, .20, .20; thus, the correct level is $l = 8$. According to the indifference intervals in Table 1, column (d), we expect the CRM will fail to distinguish level 7 from level 8. The percentages of recommending levels 6, 7, and 8 by the CRM are plotted in Figure 2, which shows that the CRM converges to level 7 as sample size grows. This agrees with the sensitivity analysis: the indifference interval for $l = 7$ not covering $\mu_6 = .15$ implies that the CRM is able to distinguish level 6 from level 7 in scenario 4 in Gasparini and Eisele (2000).

In addition, the indifference intervals suggest that the CRM is likely to recommend a dose that lies in the short range from the 20th percentile to the 30th percentile if the MTD is among the first six doses. It therefore seems that this CRM setup is a promising method if the clinicians have a strong belief a priori that the desired dose is among the first six levels.

Indifference interval, though exact in asymptotics, may not bear relevance in the context of phase I studies, where the sample size is usually small. However, a model with poor asymptotic sensitivity is likely to give poor performance in small-sample settings. In the two examples illustrated above, the sensitivity analyses successfully point out when the CRM may fail.

So far, simulation has been the only tool to evaluate the CRM's operating characteristics in actual trials. With infinite possibility of choices of dose–response models, it is sensible to restrict consideration to those with adequate sensitivity. While the technique presented in this article is not intended to replace simulation studies when planning a trial, it is a simple and useful supplement.

Résumé

La méthode de réévaluation continue (CRM) est une méthode séquentielle utilisée dans les essais de phase I en cancérologie pour déterminer la dose maximale tolérée avec une toxicité acceptable. Il a été démontré que la CRM n'est pas toujours robuste face à des erreurs de spécification du modèle. Quand cette méthode ne converge pas, certains modèles dose-réponse seront plus sensibles que d'autres pour converger vers une dose proche de celle ciblée. Dans cet article, nous interprétons la condition principale sous laquelle la CRM est robuste et l'appliquons pour évaluer la sensibilité du modèle utilisé avec la CRM. La technique présentée est un complément utile aux simulations lorque l'on planifie un essai de phase I.

References

Cheung, Y. and Chappell, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56,** 1177–1182.

Chevret, S. (1993). The continual reassessment method in cancer phase I clinical trials: A simulation study. *Statistics in Medicine* **12,** 1093–1108.

Gasparini, M. and Eisele, J. (2000). A curve-free method for phase I clinical trials. *Biometrics* **56,** 609–615.

O'Quigley, J. and Shen, L. Z. (1996). Continual reassessment method: A likelihood approach. *Biometrics* **52,** 673–684.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* **46,** 33–48.

Shen, L. Z. and O'Quigley, J. (1996). Consistency of continual reassessment method under model misspecification. *Biometrika* **83,** 395–405.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45,** 925–937.

Received December 2001. *Revised March* 2002. *Accepted April* 2002.

Appendix A

Regularity Conditions for Dose–Response Model F

(M1) $F(d;\beta)$ is strictly increasing in d for all β.

(M2) $F(d;\beta)$ is strictly monotone in β in the same direction for all d.

(M3) Given any $\pi \in (0,1)$, for each d, there should exist β in the interior of Θ such that $F(d;\beta) = \pi$.

(M4) $F_k(\beta)$ is bounded away from zero and one for all k and $\beta \in \Theta$; and $F'(d;\beta) := \partial F(d;\beta)/\partial\beta$ is uniformly bounded in β.

(M5) For each $0 < \mu < 1$ and each d, the function

$$\mu\frac{F'(d;\beta)}{F(d;\beta)} + (1-\mu)\frac{-F'(d;\beta)}{1-F(d;\beta)}$$

is continuous and strictly monotone in β.

Appendix B

Derivation of the Sets $\{H_k\}$ for $k = 1,\ldots,K$

Without loss of generality, assume $F(d;\beta)$ is strictly decreasing in β and let $\Theta = [b_1, b_{K+1}]$, where $-\infty < b_1 < b_{K+1} < \infty$. For $k = 2,\ldots,K$, define b_k such that

$$F_{k-1}(b_k) + F_k(b_k) = 2p_T.$$

Assuming (M2) and (M3), we can always find such b_k. Our goal is to show that $b_2 < \cdots < b_K$ and that $H_1 = [b_1, b_2)$, $H_j = (b_j, b_{j+1})$, $j = 2,\ldots,K-1$, and $H_K = (b_K, b_{K+1}]$ as defined in Section 3.

Definition (3.1) gives $F_{k-1}(b_k)+F_k(b_k) = 2p_T = F_k(b_{k+1}) + F_{k+1}(b_{k+1})$ for $k = 2,\ldots,K-1$. On the other hand, (M1) implies

$$F_{k-1}(b_k) + F_k(b_k) < F_k(b_k) + F_{k+1}(b_k),$$

which leads to

$$F_k(b_{k+1}) + F_{k+1}(b_{k+1}) < F_k(b_k) + F_{k+1}(b_k)$$

and hence $b_k < b_{k+1}$ for $k = 2,\ldots,K-1$ by (M2).

Now, suppose $\beta \in (b_j, b_{j+1})$ for $k = 2,\ldots,K-1$. Definition (3.1) and (M2) together imply

$$F_{j-1}(\beta) + F_j(\beta) < 2p_T$$

and

$$F_j(\beta) + F_{j+1}(\beta) > 2p_T$$

and, in turn together with (M1), implies $F_{j-1}(\beta) < p_T < F_{j+1}(\beta)$. It follows that $\beta \in H_j$.

Suppose $\beta \le b_j$. We further assume $F_{j-1}(\beta) \le p_T \le F_j(\beta)$. Otherwise, $\beta \notin H_j$. Definition (3.1) and (M2) together imply $F_{j-1}(\beta) + F_j(\beta) \ge 2p_T$. It follows that

$$|F_j(\beta) - p_T| = F_j(\beta) - p_T \ge p_T - F_{j-1}(\beta) = |F_{j-1}(\beta) - p_T|$$

and therefore $\beta \notin H_j$. By similar argument, we can show that $\beta \ge b_{j+1}$ implies $\beta \notin H_j$. As a result, $\beta \in H_j$ iff $\beta \in (b_j, b_{j+1})$ for $k = 2,\ldots,K-1$. Using a similar argument gives that $\beta \in [b_1, b_2)$ iff $\beta \in H_1$ and that $\beta \in (b_K, b_{K+1}]$ iff $\beta \in H_K$.