STATISTICS IN MEDICINE, VOL. 14, 1149–1161 (1995)

SOME PRACTICAL IMPROVEMENTS IN THE CONTINUAL REASSESSMENT METHOD FOR PHASE I STUDIES

STEVEN N. GOODMAN, MARIANNA L. ZAHURAK AND STEVEN PIANTADOSI

Johns Hopkins University School of Medicine, Division of Biostatistics, Oncology Center, Baltimore, MD 21205, U.S.A.

SUMMARY

The Continual Reassessment Method (CRM) is a Bayesian phase I design whose purpose is to estimate the maximum tolerated dose of a drug that will be used in subsequent phase II and III studies. Its acceptance has been hindered by the greater duration of CRM designs compared to standard methods, as well as by concerns with excessive experimentation at high dosage levels, and with more frequent and severe toxicity. This paper presents the results of a simulation study in which one assigns more than one subject at a time to each dose level, and each dose increase is limited to one level. We show that these modifications address all of the most serious criticisms of the CRM, reducing the duration of the trial by 50–67 per cent, reducing toxicity incidence by 20–35 per cent, and lowering toxicity severity. These are achieved with minimal effects on accuracy. Most important, based on our experience at our institution, such modifications make the CRM acceptable to clinical investigators.

1. INTRODUCTION

Phase I studies are experiments whose main focus is to find a dose of a new therapy or combination of therapies at which there is an optimal therapeutic ratio. Storer focused attention on these designs in his 1988 comparative study that showed that the most commonly used designs had little statistical justification and performed poorly in comparison with several alternatives.[1] A promising alternative to the traditional designs is a Bayesian dose-finding method developed by O'Quigley *et al.*, who called it the 'continual reassessment method' (CRM).[2] While the reported performance of this method in published simulations has been encouraging,[3] it has a number of characteristics that have impeded its acceptance both by clinical investigators and statisticians. The most important concerns involve the required duration of time for completion of the published designs, the possibility of increased toxicity compared with standard designs, and sensitivity to the statistical model upon which it is based. These issues have led at least one group of investigators to recommend against the use of the CRM at this time.[4]

Although there is no single universally accepted standard method, there is one commonly used in the U.S.A. This involves testing three subjects at a time, escalating the dose by one level if there are no toxicities, or if there are two toxicities, terminating and declaring the previous level the MTD. If there is one toxicity, three more patients are added. If there are $\geqslant 2/6$ toxicities, the trial stops and the previous dose is declared the MTD. If 1/6 experience toxicity, three subjects are assigned to the next higher dose and the trial continues.

We present simulation results for a number of modifications of the published CRM that we have found are necessary to make this design acceptable to clinical investigators. These are based on our efforts to introduce such designs at our own institution. We present the basis of our

CCC 0277–6715/95/111149–13
© 1995 by John Wiley & Sons, Ltd.
Received October 1993
Revised December 1993

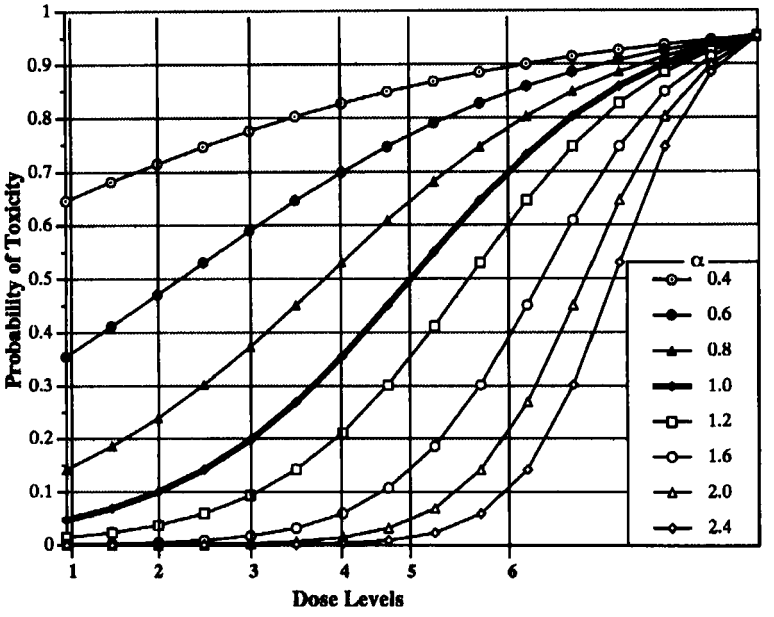

Figure 1. Family of one parameter logistic curves with constant = 3 (see equation (1)). Dose levels are indicated corresponding to toxicity probabilities of 5, 10, 20, 35, 50 and 70 per cent on the $\alpha = 1$ curve (in bold)

conclusion that the modified CRM is preferable to the standard method described above. In addition, we discuss several non-statistical benefits of the CRM that have not received emphasis.

2. THE UNMODIFIED CONTINUAL REASSESSMENT METHOD

The CRM has been described previously.[2-4] Briefly, the method is as follows:

1. A mathematical model for dose–response (or dose–toxicity) is proposed. We have used the following one parameter logistic function for p_i, the probability of toxicity at the ith dose, with the family of dose–toxicity curves indexed by α shown in Figure 1:

$$p_i = \varphi(x(i); \alpha) = \frac{\exp(3 + \alpha x(i))}{1 + \exp(3 + \alpha x(i))} \tag{1}$$

   The doses, $x(i)$, that correspond to each dose level, i, are expressed in artificial units, derived from an initial range of toxicity probabilities that are chosen *a priori*. In our simulations, we chose $p_i = 0{\cdot}05, 0{\cdot}10, 0{\cdot}20, 0{\cdot}35, 0{\cdot}5$ and $0{\cdot}70$, corresponding to $x(i)$'s$(= \varphi^{-1}(p_i; \alpha = 1))$ of $-5{\cdot}9$, $-5{\cdot}2$, $-4{\cdot}3$, $-3{\cdot}6$, $-3{\cdot}0$ and $-2{\cdot}15$ (Figure 1). (The correspondence between these doses and the estimated probabilities are changed by the data in the course of the trial, and this initial curve does not necessarily reflect the true dose response relationship.)
2. Assume a prior distribution for α. We explored the use of two priors: the exponential distribution $g(\alpha) = \exp(-\alpha)$, which has been used in most previous work, and a uniform prior, $g(\alpha) = 1/3$, $0 \leqslant \alpha \leqslant 3$.
3. Define an upper limit to the probability (p_{mtd}) of dose limiting toxicity. The dose with this probability is designated as the maximum tolerated dose (MTD).

4. Assign the first patient to the dose whose probability of toxicity is judged closest to p_{mtd}. After observing the dichotomous outcome of toxicity/no toxicity, the posterior distribution of α is calculated via Bayes theorem, and from that, the expected value of α, $E(\alpha)$ (equation (2)). We calculate $\varphi(x_i; E(\alpha))$ to give a new estimate of the dose–toxicity curve, and assign the next subject to the dose level whose associated probability is closes to p_{mtd}. This process continues either until we reach a predetermined fixed sample size, or some other trial terminating condition is satisfied.

$$\text{Expectation of } \alpha = \frac{\int_0^\infty \alpha L(\alpha; \tilde{x}_j, \tilde{t}_j) g(\alpha)\, d\alpha}{\int_0^\infty L(\alpha; \tilde{x}_j, \tilde{t}_j) g(\alpha)\, d\alpha} \tag{2}$$

L is likelihood function for the data, defined as

$$L(\alpha; \tilde{x}_j, \tilde{t}_j) = \prod_1^j [\varphi(x_j; \alpha)]^{t_j} [1 - \varphi(x_j; \alpha)]^{1 - t_j} \tag{3}$$

where x_j is the dose level of the jth subject that is $x_j(\text{i})$, and t_j is the observed toxicity (0, 1) of the jth subject.

The calculation of the expected α produces essentially the same results as calculating the expected values of the probabilities themselves, with substantially less computation. Chevret explored the impact of the standard CRM of using different prior distributions and different constants in the logistic dose–response model and found that in the absence of specific knowledge about the distribution, there were no other choices demonstrably superior to those used here (that is constant = 3 and exponential or uniform prior), and these gave consistently reasonable results.[5]

2.1. Problems with the unmodified CRM

There are several problems with the practical implementation of the unmodified CRM. Foremost is that one needs to know the result from the previous patient before assigning the next patient. This produces designs that require considerable time to complete. Since standard designs typically assign patients in groups of three, and terminate after 12–15 patients, or 5 time cycles, they have a substantial time advantage over the CRM. If more than one person were assigned per dose level in the CRM, one might expect that for any given sample size it would be less accurate, even though it would be completed in a shorter time. There has been no study of the effect of assigning more than one subject per dose level in the CRM.

A second problem with the CRM concerns the assignment of early dose levels. Most clinicians do not feel sufficiently confident enough in their initial toxicity probability estimates to start above the lowest dose (often chosen to be 10 per cent of rodents' LD_{10}). Similarly, they have discomfort with algorithms that dictate an increase of more than one dose level at a time. Closely related to this issue is the fact that the standard CRM seems to produce overall levels of toxicity slightly greater than the probability at the target MTD, and somewhat higher than that produced by the standard method.[4] We addressed these issues by making the modifications of the CRM outlined in the following section.

3. SIMULATION STUDY

We simulated the CRM, modified as follows:

Initiation of experimentation: Experimentation always starts at the lowest dose level.
Number of subjects per level: 1, 2 and 3 subjects at a time are assigned to a dose level.

Table I. Family of true dose toxicity curves used in the simulations, with the prior estimate, which was the same in all calculations. All numbers represent percentage toxicity at that dose level

Dose level	Curve 1 (prior estimate)	Curve 2	Curve 3	Curve 4	Curve 5	Curve 6
1	5	5	10	1	30	5
2	10	10	10	1	40	5
3	20	15	10	5	52	5
4	35	20	10	10	61	5
5	50	25	25	25	76	10
6	70	35	80	80	87	15

Dose escalation: Dosage could not increase by more than one level at a time, although there were no restrictions on dosage decreases. (Starting at the lowest dose and restriction to one dose level escalation has been previously used in simulations of Faries[6] and Korn[4]).

Trial termination: We explored two criteria for terminating a CRM. One was to stop at one of three fixed sample sizes- 12, 18 and 24. We chose these because most clinicians expect to enrol no more than about 20 subjects in phase I trials, and Korn *et al.*[4] showed that the average number of subjects in a standard trial was usually under 20. The second criterion was to use those sample sizes as minimums, continuing until the recommended MTD had at six subjects assigned to it.

We assessed all combinations of the previous design parameters with a variety of underlying dose–toxicity curves, some of which we chose specificially to produce a poor performance of the CRM. These curves are reported in Table I. We used a toxicity probability (p_{mtd}) of 0·20 as the target probability, both to compare these results with prior published work, and because 0·20 is roughly the probability that appears to be implicitly defined by the standard method. We used other p_{mtd}'s in our research, but they had no qualitative effect on any conclusions.

We assessed the performance of each combination of true dose–toxicity curve, patient assignment number, and stopping rule with three indices: the percentage of recommendations at each dose level; overall toxicity; and the percentage of experimentation at each dose level.

We conducted 10,000 simulations of each scenario on an IBM 486 PC and a Macintosh Quadra 700 with programs written in Pascal and C + . The random number generator was based on a linear congruential algorithm with Bays–Durham shuffling, as implemented by Press.[7]

4. SIMULATION RESULTS

The simulations we report are exclusively those of the design that had a minimum sample size of 18 subjects, with 6 subjects required to be tested at the MTD. The sample size of 24 had results that were negligibly different from this stopping rule, and a sample size of 12 is not practical when assigning 3 subjects/group, since this allows escalation only to the 4th level. We found that when the stopping rule was used with the minimum sample size of 12, the average sample size was almost 18 when the MTD was in the upper two dose levels, so this rule in practice results quite close to the rule with a minimum sample size of 18. We used the stopping rule in all of the simulations reported here since some form of data-dependent stopping must be used in practice; if minimal toxicity is observed at the first 6 dose levels, the trial clearly cannot stop. As will be shown, the sample size of 18 was rarely exceeded, and those cases in which it was exceeded were

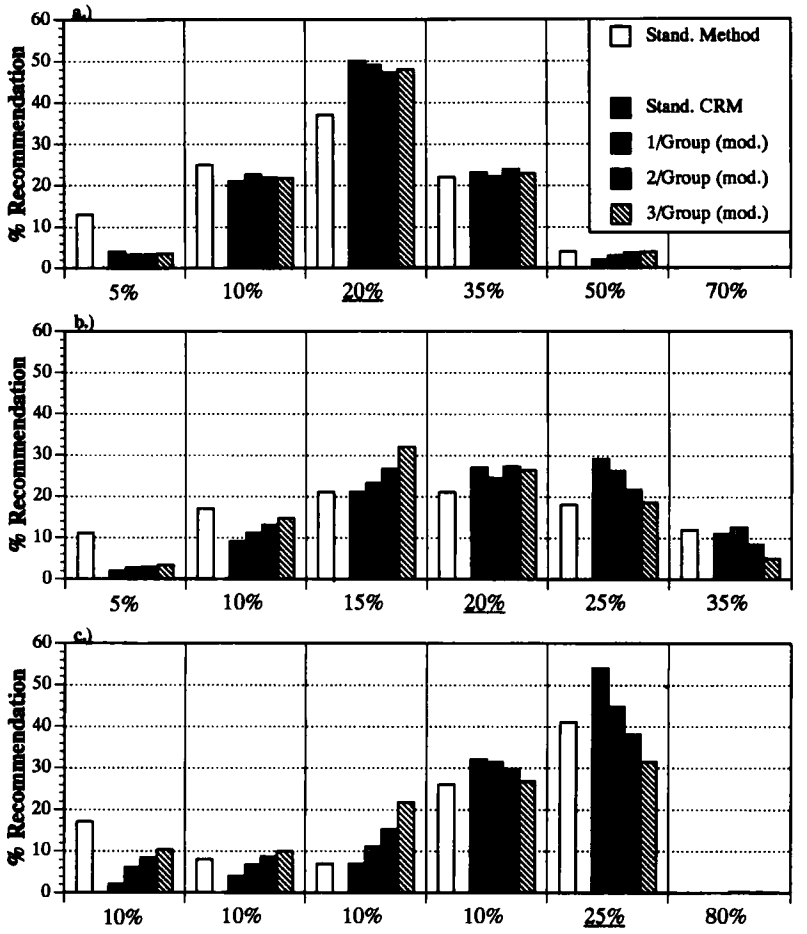

Figure 2. Recommendation percentages under three different dose–toxicity models. The first curve corresponds to when the true dose–toxicity curve is equal to the prior estimate. All CRM designs use a minimum of 18 subjects, but stop only when at least six subjects have been assigned current MTD. Toxicity probabilities are indicated under each dose level. The target probability is 20 per cent. The dose level with the probability of toxicity closest to 20 per cent is underlined. (*a*) curve 1 (*b*) curve 2 (*c*) curve 3

predictable. Figures 2 and 3 show the percentages of recommendation of the various dose levels as the MTD using five designs: the standard method (requiring only 3 subjects at the MTD); the standard CRM (starting at dose level 3, single subject assignment, no limit on dose level escalation); and the modified CRM, assigning 1, 2 and 3 subjects/dose level. Figure 4 shows the same results using a uniform prior for three dose–toxicity curves for which the results were substantively different than under the exponential prior. The target toxicity was 20 per cent in all simulations.

Figure 1(a) shows the relative performance of the designs when the initial curve equals the true dose–response curve. There are minimal differences in probability of MTD selection between the four CRM designs, which all had average sample sizes close to 18 (Table II). The standard design selected the correct third dose level about 10 per cent less frequently, with frequent stoppage at the first or second level. Use of a uniform prior had a negligible effect on selection probabilities in the modified CRM designs (results not shown). The average sample size of the standard method

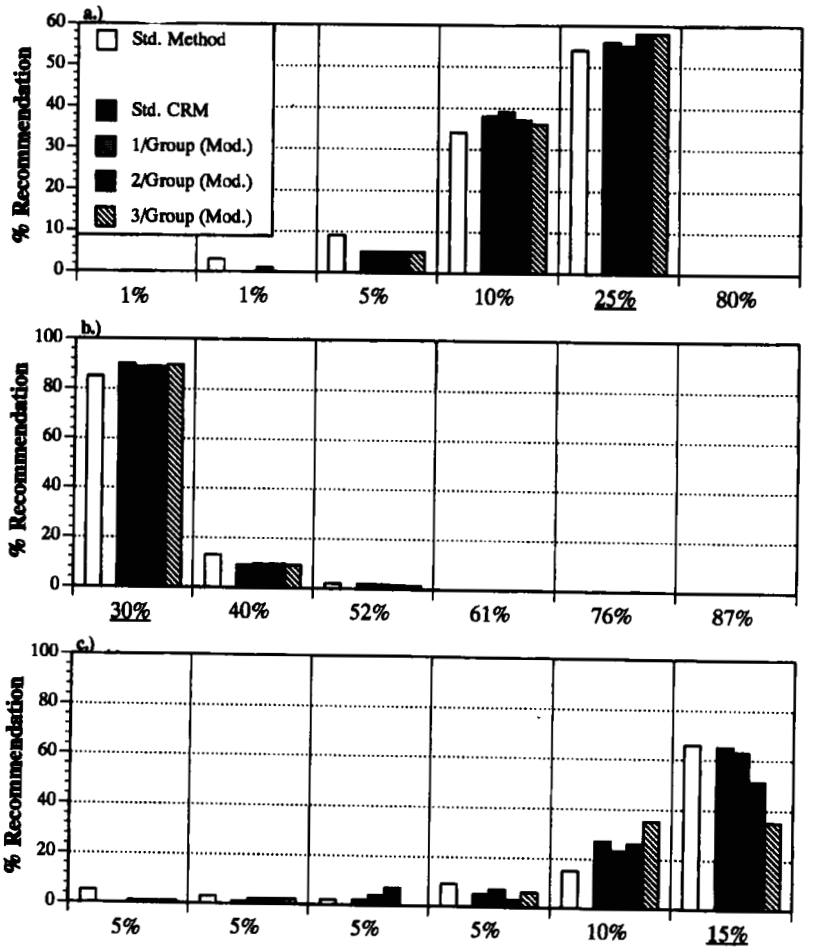

Figure 3. Recommendation percentages for curves 4–6 for different phase I designs. All CRM designs have a minimum sample size of 18, and a minimum testing of six subjects at the recommended MTD. Toxicity probabilities are indicated under each dose level. Note the different vertical scales on the bottom two graphs. The target probability is 20 per cent. The dose level with the probability of toxicity closest to 20 per cent is underlined. (*a*) curve 4 (*b*) curve 5 (*c*) curve 6

was 14·7, or 4·9 cycles of three patients. The 3/group modified CRM had 1·2 additional cycles of three patients (total of 6·1). The toxicity of the unmodified CRM is slightly higher than the standard method (23·3 versus 19·8, Table II), consistent with the findings of Korn *et al.*,[4] but the toxicity of the 2/group and 3/group designs is slightly less, lower than the target toxicity. Compared to the standard design, all of the CRM designs experiment more at levels likely to be therapeutic. There is also less experimentation at the 50 per cent toxicity level in the modified CRM designs than in the standard method. Simulations with a minimum sample size of 12 showed similar qualitative advantages, although the magnitude of the selection advantage decreased by 3–4 percentage points in most situations, with an average addition of 0·2 patient cycles over the standard method.

This example exhibits some features that appear in all of the simulations. Group sizes of three produce consistently less toxicity than the other two because they force more experimentation at lower dose levels. When the true dose–response curve is a member of the parametric family of

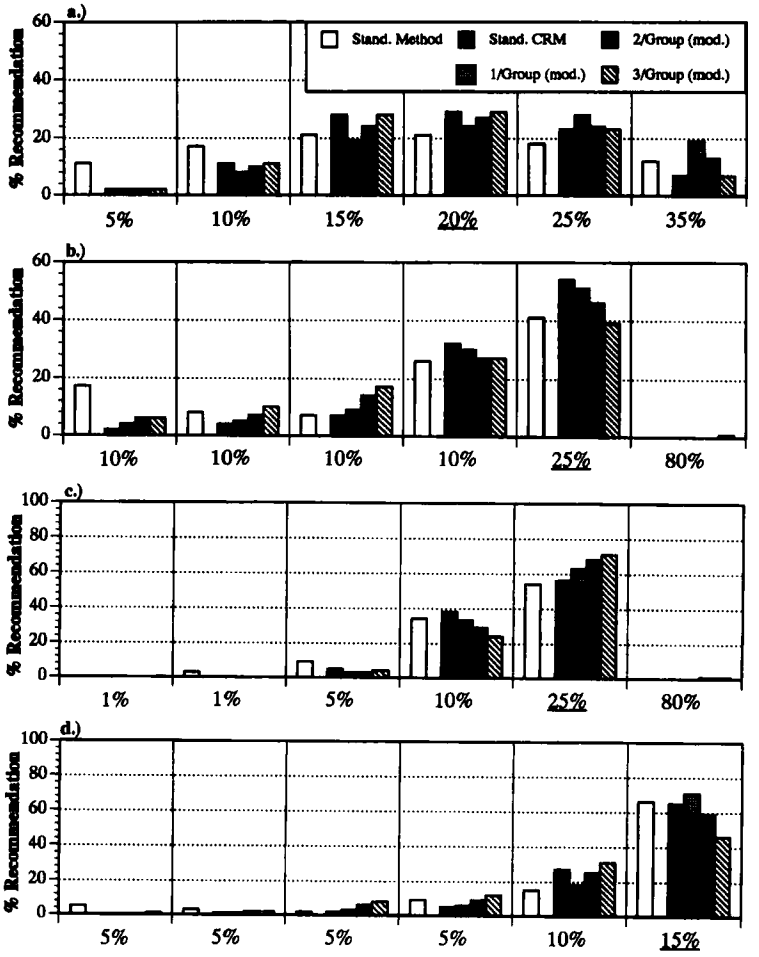

Figure 4. Recommendation percentages using a constant prior for curves that had appreciably different results than under an exponential prior (Figures 2 and 3). All CRM designs have a minimum sample size of 18, and a minimum testing of six subjects at the recommended MTD. Toxicity probabilities are indicated under each dose level. Note the different vertical scales on the bottom two graphs. The target probability is 20 per cent. (*a*) curve 2 (*b*) curve 3 (*c*) curve 4 (*d*) curve 6

curves used in the model, experimentation below the MTD is quite informative about the location of the MTD, and there is no accuracy penalty for more experimentation below the MTD. As we will see, there is some decrease in accuracy of MTD estimation with larger group designs when the true curve is not part of that family, which can be partially compensated for by use of the constant prior.

Figure 2(*b*) shows the results of these methods under a true dose–toxicity curve that is flatter and plateaus earlier than the prior estimate. All the designs have difficulty discriminating between the three middle dose levels, although all of the CRMs are superior to the standard method. All the modified CRM designs have difficulty discerning the 5 per cent differences between levels 3, 4 and 5, with 1 per group slightly overestimating the MTD, 3 per group slightly underestimating it, and 2 per group being intermediate. The reason for this pattern is due to the restricted shape of the fitted dose–toxicity curve. Figure 1 shows that the only members of the logistic family (with constant = 3) that model the higher doses with probabilities as low as 35 per cent predict that the

Table II. Experimentation percentages, total toxicity, and sample sizes for all designs described in the text. All CRM's have a minimum sample size of 18, stopping when at least six subjects have been observed at the MTD

	Experimentation percentage						Per cent toxicity observed	Average number of subjects	Average number of cycles
	1	2	3	4	5	6			
Curve 1 (% toxicity)	*5%*	*10%*	*20%*	*35%*	*50%*	*70%*			
Standard method	23	25	25	19	8	1	19·8	14·7	4·9
Unmodified CRM	11	19	36	23	9	1	23·3	18·5	18·5
Exponential prior									
Mod CRM (1/group)	14	22	33	21	8	2	22·2	18·6	18·6
Mod CRM (2/group)	19	23	33	19	6	1	19·8	18·8	9·4
Mod CRM (3/group)	22	28	31	16	4	0	17·3	18·9	6·1
Uniform prior									
Mod CRM (1/group)	11	20	33	23	10	3	24·4	18·7	18·7
Mod CRM (2/group)	16	22	31	21	8	1	21·2	18·8	9·4
Mod CRM (3/group)	22	24	30	18	6	0	18·9	19·1	6·4
Curve 2	*5%*	*10%*	*15%*	*20%*	*25%*	*35%*			
Standard method	20	21	21	17	13	8	15·7	15·7	5·2
Unmodified CRM	7	9	21	23	26	13	20·3	18·4	18·4
Exponential prior									
Mod CRM (1/group)	13	15	20	20	19	13	18·5	18·5	18·4
Mod CRM (2/group)	19	20	25	19	13	5	15·6	18·9	9·4
Mod CRM (3/group)	22	25	26	16	9	2	13·8	19·5	6·5
Uniform prior									
Mod CRM (1/group)	10	13	18	19	20	18	20·0	18·6	6·2
Mod CRM (2/group)	16	18	23	20	15	8	16·6	18·8	9·4
Mod CRM (3/group)	21	21	25	19	11	3	14·7	19·5	6·5
Curve 3	*10%*	*10%*	*10%*	*10%*	*25%*	*80%*			
Standard method	21	19	17	16	16	11	20·2	17·9	6·0
Unmodified CRM	5	4	14	30	42	5	19·6	18·4	18·4
Exponential prior									
Mod CRM (1/group)	16	12	14	24	30	5	17·8	18·6	18·6
Mod CRM (2/group)	23	17	19	19	19	3	15·2	19·1	9·6
Mod CRM (3/group)	27	21	21	16	12	2	13·3	20·1	6·7
Uniform prior									
Mod CRM (1/group)	13	11	13	23	34	7	19·9	18·9	18·6
Mod CRM (2/group)	20	16	18	19	21	5	16·8	19·0	9·5
Mod CRM (3/group)	25	19	21	17	15	3	14·4	20·3	6·7
Curve 4	*1%*	*1%*	*5%*	*10%*	*25%*	*80%*			
Standard method	16	16	17	19	19	13	18·4	19·6	6·5
Unmodified CRM	2	2	14	33	45	5	19·3	18·3	18·3
Exponential prior									
Mod CRM (1/group)	6	7	12	30	39	6	18·1	18·3	18·3
Mod CRM (2/group)	11	11	16	26	30	6	15·6	18·8	9·4
Mod CRM (3/group)	15	16	19	23	23	4	12·4	20·1	6·7
Uniform prior									
Mod CRM (1/group)	6	6	10	26	42	9	21·0	18·4	18·3
Mod CRM (2/group)	11	11	14	23	32	8	17·8	18·8	9·4
Mod CRM (3/group)	15	15	17	22	25	6	14·5	20·2	6·7

Table II. (continued)

	Experimentation percentage						Per cent toxicity observed	Average number of subjects	Average number of cycles
	1	2	3	4	5	6			
Curve 5	*30%*	*40%*	*52%*	*61%*	*76%*	*87%*			
Standard method	60	30	9	2	0	0	35·7	7·4	2·5
Unmodified CRM	71	12	11	4	1	0	35·5	18·2	18·2
Exponential prior									
Mod CRM (1/group)	76	14	7	2	1	0	33·9	18·2	18·2
Mod CRM (3/group)	79	15	6	1	0	0	33·1	18·2	9·1
Mod CRM (3/group)	80	16	4	0	0	0	32·4	18·4	6·1
Uniform prior									
Mod CRM (1/group)	72	17	8	3	1	0	35·1	18·2	18·2
Mod CRM (2/group)	75	17	7	1	0	0	33·6	18·2	9·1
Mod CRM (3/group)	78	15	6	1	0	0	33·1	18·2	6·3
Curve 6	*5%*	*5%*	*5%*	*5%*	*10%*	*15%*			
Standard method	17	17	16	16	17	17	7·5	19·7	6·6
Unmodified CRM	2	2	9	13	28	47	11·4	18·3	18·3
Exponential prior									
Mod CRM (1/group)	9	8	10	11	20	41	10·2	18·5	18·5
Mod CRM (3/group)	15	13	16	16	19	21	8·0	19·3	9·7
Mod CRM (3/group)	18	18	20	18	16	11	6·9	20·8	6·9
Uniform prior									
Mod CRM (1/group)	8	8	9	10	16	48	10·7	18·4	18·4
Mod CRM (2/group)	14	13	15	15	18	25	8·4	19·0	6·3
Mod CRM (3/group)	18	16	18	17	17	14	7·1	21·3	7·1

lower dose levels will have very low toxicity probability. Thus, when more experimentation is done at the higher doses, the tendency is to underestimate the probabilities at lower dose levels, and therefore slightly overestimate the MTD. Conversely, the 3/group designs gather most information at the lower levels, leading them to overestimate the toxicity at higher levels, and slightly underestimate the MTD. Use of a uniform prior (Figure 4(*a*)) reduces some of the disparity between the three designs by giving more weight to higher values of alpha (corresponding to lower toxicity at high doses). The most significant disadvantage of the standard method relative to the CRM with this dose–toxicity curve is the frequent selection of the lowest dose by the standard method. Korn *et al.*,[4] showed that this feature of the standard method is not appreciably affected by requiring six subjects to be tested at the MTD.

Table II shows that the experimentation, toxicity and sample size patterns mirror those from the first curve. The standard method experiments slightly more at the highest levels than the 2 and 3 subject/group modified CRM designs, but at roughly similar frequency below that level. The toxicity of the standard method (15·6 per cent) is a bit lower than the 1/group designs (18·5 per cent), and higher than the 2 and 3 group designs (15·6 per cent and 13·8 per cent). Under a uniform prior, experimentation at higher levels rises slightly, as does total toxicity, although the 3/group design still retains an advantage over the standard method (15·7 per cent and 14·7 per cent). On average, the standard method takes about three subjects less than the CRM designs, and 1·3 cycles less than this 3/group CRM design.

Figure 2(*c*) shows the performance of these designs in a situation that clinicians hope not to encounter; a dose at which there is a dramatic increase in toxicity. We chose this to occur at the

last dose level, which should produce the most severe problems for the CRM, particularly with three subjects/group, since that design cannot have much experimentation at the maximum dose levels with a minimum sample size of 18. A curve like this, with non-negligible and non-increasing toxicity for the first four doses, is not a part of the parametric family we modelled. The standard method had a very high rate of premature stopping, choosing the lowest dose, or one below it, 17 per cent of the time. Although all CRM designs correctly peak at the 5th dose level, we see that with more experimentation at the lower doses (with more subjects per group), the error in extrapolating to higher doses increases, and the MTD is more frequently underestimated. As before, the use of a uniform prior, by increasing the prior weight on low toxicity curves, greatly ameliorates that effect (Figure 4(*b*)).

Toxicity figures (Table II) show an advantage for all of the CRM designs over the standard designs, and less experimentation at the highest dose level. The 3/group designs have the greatest toxicity advantage, which was maintained with the use of a uniform prior.

The results of these designs under a dose–toxicity curve similar to the previous one, except with negligible toxicity at low doses (curve 4), is shown in Figure 3(*a*). In this setting, we would not want to pick too low or too high a dose level, assuming toxicity is correlated with efficacy. The simulations show good and similar performance of all modified CRM designs because curve 4 is closer in shape to the logistic family than was curve 3. This curve did not include a dose level with exactly the target probability of 20 per cent. The accuracy of all the methods was comparable, although the standard method again chose the 5 per cent dose level slightly more often. The use of the uniform prior increased the accuracy of the larger group size designs enough to actually exceed that of the other designs (Figure 4(*c*)). The standard method again had more experimentation at the highest dose level and less at therapeutic levels than any of the CRMs (Table II). The 3/group designs had at least a 25 per cent proportional reduction in toxicity with greater accuracy and the same number of subject cycles as the standard design.

Figure 3(*b*) shows the results of simulations when all dose levels are too high (curve 5). The lowest dose level (and, had we allowed it, perhaps a dose below) was chosen about 90 per cent of the time by all CRM designs. The very high accuracy was due to the fact that almost all experimentation occurred at the two lowest dose levels, producing very precise estimates of their toxicity probabilities. Toxicity was still lowest by a small amount in the three per group designs (Table II, curve 5). The standard method achieved almost the same levels of accuracy with an average of only 7·2 subjects/trial. Simulations of modified CRM with no minimum sample size, but retaining the requirement of six subjects being tested at the chosen MTD, lowered the average sample size to about nine subjects, with a drop of 3–5 percentage points in accuracy. Situations like this point to the need for flexible CRM stopping rules, which allow for termination when some predetermined degree of precision in p_{mtd} estimation has been reached. In practice, it is likely that investigators would add doses below the planned levels if excessive toxicity were seen at the lowest dose level.

The final curve used in the simulations had low toxicity at all levels (Figure 3(*c*)). The closeness of the toxicity probabilities, combined with the inability of the 3/group designs to experiment substantially at the highest dose levels, resulted in the largest differences between the modified CRM designs. The 1/group design chose the correct 6th level 62 per cent of the time, versus 37 per cent of the 3/group design. The top two levels were chosen 83 per cent of the time with 1/group, and 68 per cent with 3/group. This effect was reduced moderately by using a uniform prior (Figure 4(*d*)). There was a very large difference in experimentation at the highest dose levels for example, 42 per cent versus 21 per cent versus 7 per cent for 1, 2, 3 per group designs at level 6). Because of the closeness and low probability of the toxicity at all levels, the errors in this situation are not as serious as some of those in the previous designs.

The actual comparative performance of the CRM versus the standard design is difficult to assess in this last simulation because 66 per cent of the standard trials did not formally terminate; these trials have to progress beyond the 6th dose level in order to declare the 6th dose level to be the MTD. In the case of the CRM, there is a parallel situation; the fitted dose–response curve at the termination of many of these trials indicated that the true MTD was likely to exist at dose levels higher than the sixth. As with the standard method, in practice this could lead to continuation of dose escalation to higher levels.

A consistent pattern in all of these simulations is that the accuracy of the two and three per group designs suffers when the MTD is at the highest dose levels and the dose–toxicity curve is misspecified. The two and three subject designs have more difficulty than the unmodified CRM compensating for this misspecification by oversampling at the MTD. The use of constant prior distribution increases the CRM's accuracy in those situations, with only a minimal increase in toxicity (never higher than the standard method). When the MTD is at a low dose level (that is, the true curve is a high toxicity curve) putting a higher weight on low toxicity curves produces no disadvantage because the larger number of patients assigned to the MTD overwhelms the effect of the prior.

One early report on the unmodified CRM[3] showed somewhat higher percentages of correct predictions than we obtained with our algorithms (modified to simulate the reported designs). For example, at 62 per cent ($N = 25$) correct prediction frequency was reported when the dose–toxicity curve was equal to the prior, in contrast to the 54 per cent we found, and approximately an absolute 5–10 per cent fewer correct MTD choices with similar shaped underlying curves. Our toxicity and experimentation results were similar, as were all results on extreme curves (like our curves 5 and 6). A small error discovered in the numerical integration routine used for that paper may be responsible for these differences (O'Quigley, personal communication).

5. DISCUSSION

The most striking feature of these results is that assigning three subjects at each dose level usually did not produce a pronounced drop in the accuracy compared to one per group designs. The determinant of its relative accuracy was whether the misspecification of the dose–toxicity curve was such that extrapolation from high to low or low to high dose levels was more accurate, within the restricted family of logistic curves used, and whether adequate experimentation was permitted at the MTD. When there was no misspecification, there was essentially no accuracy difference between the two designs. The consistent features of three subject per group designs for all underlying dose–toxicity curves were that when combined with restricted dose escalation, they forced experimentation at lower levels, produced less toxicity, and chose MTD levels more conservatively than one subject per group designs. When there was a substantive disparity between modified CRM designs, this was lessened or eliminated by use of a constant prior. Priors that put even more weight on low toxicity curves could reduce this difference further, but at the price of more toxicity, which is likely to be unacceptable.

One criticism of the CRM is that the summary toxicity numbers do not take into account the dose level at which the toxicities occur, because higher doses can produce not just more frequent but more serious toxicities.[4] In none of the simulations did the modified CRMs with two or three subjects per group produce more experimentation at high toxicity levels (that is $\geqslant 35$ per cent) than the standard method, and they were usually substantially lower. It must be emphasized that toxicity far below the target level is not always good; if there is a monotonic relationship between dose and efficacy, it can mean that too much experimentation took place at possibly sub-therapeutic levels.

The unmodified CRM, which does not restrict the dose escalation to one step, produces only modest increases in accuracy over the modified CRM, but at the price of greater toxicity, and, most important, clinical acceptability. In our experience, clinical investigators are extremely wary about accepting 'black box', statistically driven trial designs, particularly when the bases for these designs are dose–toxicity curves that are, at best, educated guesses. We have found that keeping as many features as possible similar to conventional designs greatly enhances the acceptance of the CRM. This fact, together with the dramatic decrease in trial duration made possible by using more subjects per group, and the lower toxicity, make a compelling argument that one should always implement CRMs with 2–3 subjects assigned at each recommended dose level, and with only single dose level escalations. It also appears that a constant prior may optimize selection probabilities.

Other aspects of CRM implementation merit comment. Use of the CRM forces clinical investigators to acknowledge the presence of a dose–toxicity curve, and to consider explicitly the appropriate probability of toxicity at the MTD. The automaticity of the conventional method has led to a striking lack of awareness that a dose–toxicity relation exists that one can or should model. In fact, the conventional designs preclude reliably estimating the dose–toxicity function.[1] We have found that among many scientists very familiar with dose–response models, dose–toxicity curves in a phase 1 setting are a new concept.

In our comparisons with the standard method, we chose a target toxicity probability (20 per cent) that we knew, based on simulations, would optimize its performance. Whereas it is quite straightforward to use another target toxicity with the CRM, it is impossible to predict in a straightforward manner from the nominal stopping criteria (for example, 2/3, 2/4, 2/6 etc.) what the effective target probability of the standard method will be. When freed to make an explicit choice of p_{mtd}, investigators at our institution have used targets as low as 10 per cent and as high as 50 per cent, varying according to the severity and reversibility of the toxicity, and the nature of the drug's proposed benefit. More importantly, they find the discussions about the appropriate target level to be quite useful. There were several instances where investigators who considered the CRM but eventually used the standard method (so they would not have 'approval trouble') mentioned that the process of choosing a p_{mtd} was still extremely helpful for them and their colleagues. Unfortunately, explicitness about the choice of p_{mtd} occasionally invites criticism from reviewing bodies. One response to such criticism is that the standard method has an implicit median stopping dose, and this implicit definition is not more likely to be appropriate simply because it is unknown or unstated.

Comparisons between the 'standard method' and the CRM are difficult to interpret, since there is no clear p_{mtd} selected by the standard method, and its stopping behaviour can vary with a change in dose levels even when the underlying dose–toxicity curve is unchanged. The criteria by which we should judge whether the standard method produces a 'correct' result are unclear. Also, in practice we can increase or decrease the CRM sample size to modify accuracy, whereas it is not easy to change the standard method to produce larger trials or results of predictable accuracy.

Korn *et al.*[4] concluded that the one subject/group CRM should not yet supplant the traditional method, based on higher toxicity, longer duration of trials, too much experimentation at high dose levels, and the comparable accuracy of the two methods (with p_{mtd}'s near 0·20). Their implementation of the CRM used the stopping rule used here with no minimum sample size, that is, trials stopped when six subjects had been tested at the current MTD estimate. We think that the modifications presented here address all of these problems and greatly strengthen the case for the CRM.

There are several aspects of the CRM that we did not explore. One was the use of a two-parameter logistic model, which would theoretically lessen the problem with extrapolation using

sometimes ill-fitting one parameter models. Both O'Quigley *et al.* and Chevret investigated this and found that there is only a small impact on the accuracy of MTD selection in the standard CRM.[2,5] Whether there is greater advantage when more than one subject per group is used, which relies more heavily on extrapolation, is a subject for further study.

Another modification we did not examine was using the width of the Bayesian posterior probability interval as a stopping criterion, thereby permitting floating sample sizes. This has a number of attractive features, one being the possibility of terminating the CRM when there has been considerable experimentation at a few dose levels, such as in our curve 4. Even if such a plan was used, however, at least 18 subjects would have to be planned for if one used the three subjects/group design. With all designs, 18 is a number at which one achieves reasonable levels of accuracy for a broad range of underlying dose–toxicity functions. Chevret's work,[5] as well as our simulations, show that as the sample size exceeds 20, the increase in accuracy is fairly slow.

Finally, we did not explore the possibility of a mixture of group sizes. For example, one could start with three subjects assigned at the lowest levels, and then shift to 2/group or 1/group assignments. This might combine the conservative features of the larger groups with the slightly better selection probabilities of the smaller groups. However, using the smaller groups would prolong the trial.

In summary, we re-emphasize the importance to the clinical investigator of the modifications to the CRM described in this paper. While the unmodified CRM may have some small statistical advantages over these designs, clinicians find the duration of those trials, the perceived aggressive push towards higher dose levels in the face of little knowledge, and the possibility of excessive toxicity (quantitative and qualitative), as unacceptable. We have shown that restricting dose escalation and assigning more than one subject per group addresses all those concerns, and makes the CRM as modified here a practical, and we believe preferable alternative to standard designs.

REFERENCES

1. Storer, B. 'Design and analysis of phase I clinical trials', *Biometrics*, **45**, 925–937 (1989).
2. O'Quigley, J., Pepe, M. and Fisher, L. 'Continual reassessment method: a practical design for phase I clinical trials in cancer', *Biometrics*, **46**, 33–48 (1990).
3. O'Quigley, J. and Chevret, S. 'Methods for dose finding studies in cancer clinical trials: a review and results of a Monte Carlo study', *Statistics in Medicine*, **10**, 1647–1664 (1991).
4. Korn, E., Midthune, D., Chen, T., *et al.* 'A comparison of two phase 1 trial designs', *Statistics in Medicine*, 1994; in press.
5. Chevret, S. 'The continual reassessment method in cancer phase 1 clinical trials: a simulation study', *Statistics in Medicine*, **12**, 1093–1108 (1993).
6. Faries, D. 'The modified continual reassessment method for phase I clinical trials', *in* Association A.S. (ed.) ASA Meeting, Biopharmaceutical section, 1991.
7. Press, W., Terkolsky, S., Vetterling, W., *et al. Numerical Recipes in C*, 2nd edn, Cambridge University Press, 1992.

学霸图书馆
www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航：

图书馆首页　　文献云下载　　图书馆入口　　外文数据库大全　　疑难文献辅助工具