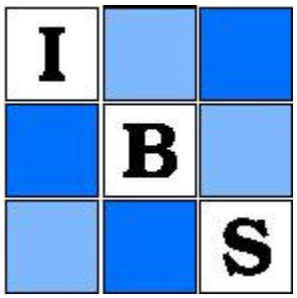

BIOMETRICS 52, 673–684
June 1996

Continual Reassessment Method: A Likelihood Approach

John O'Quigley

Unité 436 INSERM, Paris, France and Department of Mathematics, University of California at San Diego, La Jolla, California 92093, U.S.A.

and

Larry Z. Shen

Biometrics and Statistical Sciences, Procter and Gamble Pharmaceuticals, 11262 Cornell Park Drive, Cincinnati, Ohio 45242, U.S.A.

SUMMARY

The continual reassessment method as described by O'Quigley, Pepe, and Fisher (1990, *Biometrics* **46,** 33–48) leans to a large extent upon a Bayesian methodology. Initial experimentation and sequential updating are carried out in a natural way within the context of a Bayesian framework. In this paper we argue that such a framework is easily changed to a more classic one leaning upon likelihood theory. The essential features of the continual reassessment method remain unchanged. In particular, large sample properties are the same unless the prior is degenerate. For small samples, and as far as the final recommended dose level is concerned, simulations indicate that there is not much to choose between a likelihood approach and a Bayesian one. However, for in-trial allocation of dose levels to patients, there are some differences and these are discussed. In contrast to the Bayesian approach, a likelihood one requires some extra effort to get off the ground. This is because the likelihood equation has no solution until we observe a toxicity. Initially then we suggest working with either a standard Up-and-Down scheme or standard continual reassessment method until toxicity is observed and then switching to the new scheme.

1. Introduction

In Phase I and II dose finding studies, the continual reassessment method (CRM) introduced by O'Quigley, Pepe, and Fisher (1990) was proposed as an alternative to the Up-and-Down schemes described by Storer (1989). Our conclusion was (and is) that CRM is superior to the Up-and-Down schemes in terms of more accurate recommendation of the target level, more efficient use of accrued information, less allocation at suboptimal levels in the course of the trial, and, in particular, as the trial continues, the property of convergence to the target level (O'Quigley et al., 1990; O'Quigley and Chevret, 1991; Shen and O'Quigley, 1996). The Up-and-Down schemes, being "memoryless," sample more or less widely around some percentile but do not converge. Even so, our conviction is not shared by everyone and, in particular, Korn et al. (1994) argue in favor of continuing use of the standard Up-and-Down schemes. Naturally we take issue with Korn et al. but we do not do so here. Nonetheless we do recognize that some of the difficulties associated with the use of CRM make it less appealing than it otherwise might be.

Among these perceived difficulties are the following:

(1) Prescribing initial experimentation at our best prior guess of the target level, rather than the lowest available. This feature of CRM is sometimes viewed with suspicion by practitioners. We have previously argued that the starting level ought to correspond to the practitioner's best guess of dose or dose combination producing the targeted toxicity, and that we should include levels below

Key words: Clinical trial; Continual reassessment method; Dose escalation; Dose finding studies; Maximum likelihood; Phase 1 trial; Toxicity.

and above such a level to enable either escalation or de-escalation. For a design with six levels, it was suggested that we start out at level three. While not wishing to totally abandon this reasoning, we now acknowledge that it has largely failed to convince. The Up-and-Down schemes start out at the lowest available level and this seems to give practitioners some reassurance, probably in part because "undertreating" is viewed less seriously than "overtreating." Also, there are situations, in particular when dealing with a single therapy, where the starting dose is precisely defined on the basis of animal studies, for example 1/10 the lethal dose in mice. Returning to CRM, there is nothing that prevents us, when using the method, allocating to a level other than that indicated so that, for instance, we could start out at the lowest available level. Some efficiency losses may ensue but that is an entirely different issue.

Related to the question of the starting level is the fact that a decision to escalate or de-escalate based on grouped inclusions at the beginning of the study is often felt to be more reliable than one based on a single observation. This also can be made a feature of CRM. It is then quite possible to make CRM closely resemble standard Up-and-Down procedures in the early stages of experimentation.

(2) Incorporating prior information, even when vague, into the decision making is often perceived as implying operational characteristics that are, at least in part, arbitrary. Such criticism is commonly raised against approaches formulated in a Bayesian way and the usual answers to such criticism would apply here. We would also argue that, as sequential inclusion continues, the impact of such prior information diminishes rapidly. Nonetheless such dependence on prior information has been considered to be a serious handicap to the method (Mick and Ratain, 1993) and this concern may be a hindrance to its wider implementation. In part this may be due to the fact that sample size will rarely be much more than 20 and often less, the effect of the prior then being perceived as something we cannot neglect.

(3) The numerical integrals necessary in the implementation of the technique require special programming. We do not anticipate sample sizes becoming very large for such studies, but for samples as small as 12 the calculations begin to become somewhat intensive.

The purpose of this paper is to present a new version of CRM, called CRML, which deals immediately with the above three difficulties. Operating characteristics of CRM and CRML, in the usual case where we do not have strong prior information, are similar overall. Large sample properties are identical. The purpose behind the development of CRML is to give users a choice. If the user is reluctant to use CRM for any of the above reasons, then CRML is an alternative. In practice it will have little impact on the final recommendation and, in most cases, only make a relatively minor difference on in-trial allocation of dose levels to patients.

In the following section we briefly recall the main ideas behind CRM. This is done via 4 short subsections, 2.1 to 2.4. Sections 3.1 to 3.4 parallel this development for CRML, making it easy for the reader to identify the exact nature of the differences between CRM and CRML. In Section 4 we undertake a limited study of the comparative operating characteristics of CRM and CRML.

2. Continual Reassessment Method

2.1 *Basic Setup and Models*

We will need to assume that the true dose toxicity relationship increases monotonically. A fixed number of levels are chosen for study. In our earlier work we concentrated our attention on six levels and generalizing this to more levels presents no conceptual difficulty. There are usually no natural units for the dose levels, each level itself corresponding to a combination of chemotherapeutic and/or radiotherapeutic regimens. We imagine instead some conceptual dose, increasing when one of the constituent ingredients increases, and, under our monotonicity assumption, translating itself as an increase in the probability of a toxic reaction. Choosing the dose levels amounts to selecting levels (treatment combinations) with some guessed-at prescribed probability of a toxic reaction.

One of the levels will have a guessed-at probability of a toxic reaction close, if not equal to, the aimed for target "acceptable" toxicity level. In our earlier work with six levels we took this to be level 3, the initial starting level, enabling both escalation and de-escalation whereas standard schemes only usually allow for escalation. Values for the target toxicity level, θ, might typically be 0.2, 0.25, 0.3, or 0.35.

We can consider the dose level for the jth entered patient, X_j, to be random, taking discrete values from a fixed range of k levels, $d_1, \ldots, d_k$. Let Y_j be a binary random variable $(0, 1)$ where 1 denotes severe toxic response for the jth entered patient $(j = 1, \ldots, n)$. Suppose that the probability of toxic response when $X_j = x_j$ is modeled via

$$\Pr(Y_j = 1 \mid X_j = x_j) = E(Y_j \mid x_j) = \psi(x_j, a_0)$$

for some one-parameter model $\psi(x_j, a_0)$. The rationale behind the use of a one-, rather than two-, parameter model is given in Sections 2.1, 2.2, and 6.3 of O'Quigley et al. (1990). Very briefly, the reason for using a single-parameter model has to do with identifiability. It is not, as suggested by some workers, because of a need to simplify two-dimensional integration in a Bayesian context. When working with likelihood, we will still be required to restrict attention to one-parameter models.

There is a very wide choice of potential working models and prior distributions. For the working model we require a function of conceptual dose going from zero to one. In our original paper (O'Quigley et al., 1990) we used the following model:

$$\psi(d_i, a) = \{(\tanh d_i + 1)/2\}^a, \qquad i = 1, \ldots, k. \tag{1}$$

As the value of d_i increases, the probability of toxicity increases, going from zero at large negative values to one at large positive values. At each given value of d_i, as the value of the unknown a increases then the probability of toxicity decreases, taking on values between 0 and 1. A requirement of CRM is that at each value d_i the model must be rich enough to model the true probability of toxicity at that same level. This is easily seen to be the case with model (1). In equation (1) the doses d_i, $i = 1, \ldots, k$ are only conceptual values, not necessarily corresponding to so many milligrams of some product. We could equally well overlook them and simply work with the model, $\psi(d_i, a) = \alpha_i^a, i = 1, \ldots, k$, where α_i represents the prior estimated probabilities of toxicity at level d_i. In a study on the accuracy of associated confidence intervals (O'Quigley, 1992), we also worked with a reparameterization of the preceding model whereby a was replaced by $\exp(a)$. This parameterization permitted a prior defined over the whole real line and presented some advantages in terms of coverage accuracy. There are infinitely many possibilities and, in unpublished work carried out by Margaret Pepe, a number of other distribution functions, the Weibull in particular, were evaluated. There does not appear to be a simple clear recommendation for general situations, the chosen working model being to some extent a matter of taste. A word of caution is nonetheless in order because, despite the wide and general conditions for a model to belong to the CRM class, some of the most obvious choices can fail. The particular parameterization chosen for the logistic model by Chevret (1993) and Korn et al. (1994) produces a model outside the CRM class having potentially very poor properties.

2.2 *Getting the Trial Underway*

The original idea of CRM was that the first entered patient would be treated at some level, believed by the experimenter, in the light of all current available knowledge, to be the target level. Available knowledge, possibly together with his or her own subjective conviction, led the experimenter to a 'point estimate' of the probability of toxicity at the starting dose to be the same as the targeted toxic level. In our first paper on CRM (O'Quigley et al., 1990) the targeted toxicity level was 0.2. In equation (1) the 'dose' that satisfied this was -0.69 so that we had $d_3 = -0.69$. In addition, we had $d_1 = -1.47$, $d_2 = -1.10$, $d_3 = -0.69$, $d_4 = -0.42$, $d_5 = 0.0$, and $d_6 = 0.42$ so that for $a_0 = 1$, corresponding to the mean of some prior distribution, the prior point estimates of toxic probabilities were 0.05, 0.1, 0.2, 0.3, 0.5, and 0.7, respectively. We considered that our point estimate 0.2, corresponding to $a = 1$, was very likely to be in error. This notion of error was expressed via a prior density on a, called $g(a)$. In equation (1) the domain of definition of a is the real positive line and so we gave consideration to distributions having support on $\mathcal{R}^+$, the family of gamma distributions in particular. The simplest member of the gamma family, the standard exponential distribution with $g(a) = \exp(-a)$, showed itself to be a prior sufficiently vague for a large number of situations. For this prior, 95% Bayesian confidence intervals for the probability of toxicity at the starting dose lay between 0.003 and 0.96. For the lowest level, the corresponding interval is $(10^{-5}, 0.93)$ whereas for the highest level, having a point prior estimate of 0.7, the interval becomes (0.26, 0.99). Such a prior is therefore not vague at all levels and suggests that the highest level is likely to be too high. A yet more vague prior would help acceleration from the starting level to the highest level when we greatly overestimate the new treatment's toxic potential. Even so, it does not take long for the accumulating information to "override" the prior and this simple exponential formulation appeared to be fairly satisfactory overall (O'Quigley et al., 1990). For the reparameterized model in which a is replaced by $\exp(a)$, the domain of definition of a is now the whole real line. A normal distribution with mean zero and variance sufficiently large to represent vague knowledge gives a suitable prior in this situation.

The first entered patient is then observed for the presence or absence of a toxic reaction. Once we have this information we can proceed to reassessment as described in the following section.

2.3 *Reassessment Using Bayes Formula*

Let $\Omega_j = \{y_1, x_1, \ldots, y_{j-1}, x_{j-1}\}$ and let $f(a, \Omega_j)$ be a nonnegative function summarizing all available information about the parameter a_0. This is our current prior before experimentation on the jth subject and we will require

$$\int_{\mathcal{A}} f(a, \Omega_j)\, da = 1, \qquad j = 1, \ldots, n,$$

where $\mathcal{A}$ is the domain of definition of a. The response of the jth patient is observed and then we use a formulation of Bayes theorem to obtain $f(a, \Omega_{j+1})$ from $f(a, \Omega_j)$, thereby updating our information about the parameter a_0 as observations become available. Suppose that the treatment level allocated to the jth patient is x_j. Then, having observed whether the jth patient experiences a toxic response, we can evaluate the function $f(a, \Omega_{j+1})$. To be precise, if we let $\phi(x_j, y_j, a) = \psi^{y_j}(x_j, a)\{1 - \psi(x_j, a)\}^{(1-y_j)}$, then we have

$$f(a, \Omega_{j+1}) \int_0^\infty g(u) \prod_{\ell=1}^{j} \phi\{x_\ell, y_\ell, u\}\, du = g(a) \prod_{\ell=1}^{j} \phi\{x_\ell, y_\ell, a\}. \tag{2}$$

Since $f(a, \Omega_1) = g(a)$ we can, in light of x_1 and the observed y_1, evaluate $f(a, \Omega_2)$ by using the above formula. We continue in this way proceeding sequentially. Note that the right-hand side of equation 2, upon removing the prior $g(a)$, is simply the likelihood function. Taking logarithms and differentiating with respect to the parameter a lead to equation 4 of Section 3.3.

2.4 *Dose Allocation*

In order to decide, on the basis of available information, the appropriate level at which to treat the jth patient, we will need some estimate of the probability of toxic response at level $d_i, i = 1, \ldots, k$, given Ω_j. This is denoted as $\theta_j(d_i)$, where

$$\theta_j(d_i) = \int_{\mathcal{A}} \psi(d_i, a) f(a, \Omega_j)\, da, \qquad i = 1, \ldots, k. \tag{3}$$

An alternative estimate might be $\tilde{\theta}_j(d_i)$, where

$$\tilde{\theta}_j(d_i) = \psi(d_i, \mu_j), \qquad i = 1, \ldots, k, \quad \mu_j = \int_{\mathcal{A}} a f(a, \Omega_j)\, da.$$

Finally, let $\Delta(v, w)$ denote some measure of distance between v and w, for example, $\Delta(v, w) = (v - w)^2$. Then for the jth entered patient in the trial, choose dose level d_i such that $\Delta(\theta_j(d_i), \theta)$ or $\Delta(\tilde{\theta}_j(d_i), \theta)$, $i = 1, \ldots, k$, is a minimum. Continuing in this way, the recommended dose will be the dose $d_i, i = 1, \ldots, k$, such that $\Delta(\theta_{n+1}(d_i), \theta)$ or $\Delta(\tilde{\theta}_{n+1}(d_i), \theta)$ is minimized. We can quite easily weight these distances if we so wish. For instance, we may prefer, all else being equal, to allocate below the estimated target rather than above it. It would be quite straightforward even to assign an infinite penalty to estimates above the target so that we would always tend to approach the target level from below.

3. Maximum Likelihood Approach

3.1 *Basic Setup and Models*

The basic set up and models of Section 2.1 apply equally well in this section. We have not carried out an extensive study on the various possible models and instead limit attention to the model of equation (1), expressed as in equation (1) or via $\psi(d_i, a) = \alpha_i^a$. In practice we work with a reparameterization of this model so that $\psi(d_i, a) = \alpha_i^{\exp(a)}$. Of course the maximum likelihood estimate of ψ is the corresponding simple function of the maximum likelihood estimate of a_0.

3.2 *Getting the Trial Underway*

We do not work with a prior and therefore have no summary value (mean or median, for example) as an estimate for a_0. Such information was needed to get CRM off the ground (Section 2.2).

Furthermore, the likelihood expression of the following section will only have solutions at $a = 0$ or $a = \infty$ when all the responses are either toxicities or nontoxicities, i.e., there is no heterogeneity among the responses. A requirement then for CRML is that, before applying the above sequential approach, we must have heterogeneity among the responses. This will ensure the existence of a maximum likelihood estimate for the particular models to which we have limited our attention. For

more general situations and richer models, conditions for existence can be found in Silvapulle (1981).

Thus we do not consider the experiment fully underway until we have a small set of heterogeneous responses. These could be obtained in any way, in particular using ordinary CRM or on the basis of any standard Up-and-Down scheme. Our preference would be to use ordinary CRM although, should the first point of Section 1 be of particular concern, we might wish to closely emulate standard procedures. As we have previously shown, standard procedures are most definitely not the most efficient but they have the advantage of being widely accepted and in practice, for most situations, it will make little difference because they will only be used, on average, for a small part of the experiment. In this case the rule is even simpler than in standard experimentation and consists of escalating after every group of three until the point at which some heterogeneity is observed. Once we have achieved heterogeneity we continue as prescribed in Sections 3.3 and 3.4. Should we strongly suspect that initial experimentation is way below the target level, then we may wish to consider more rapid escalation. For instance, should level 6 be the correct level and we start out at level 1, then we will use at least 18 patients in order to begin experimentation in the area of interest. Increasing the level after observing nontoxicities on two consecutive patients or even after a single patient would be a way to overcome such an inefficient use of resources. We do not explore here the relative merits of different initial designs, but clearly a design that escalates after a single observation of a nontoxicity will have an advantage in terms of efficiency. Some will feel that any such advantage is paid for too dearly since escalation would seem to proceed less cautiously. Against this can be argued the opposite; that, at least in the presence of a sharp dose–toxicity curve, single escalation will only expose a single individual to a toxicity whereas grouped escalation could result in all three experiencing toxicity. There are many other considerations when deciding on an initial scheme and we will limit ourselves here to the observation that the choice of initial scheme will impact the overall operating characteristics of CRML. Some further insight is given in Section 4, although this is by no means exhaustive.

3.3 *Reassessment Using Likelihood*

We limit our attention to the power model outlined in Section 2.1. This has a simple translational form whereby

$$\log\log\psi(d_i,a) = \log\log\psi(d_i,1) + \log a.$$

For this model, heterogeneity is equivalent to saying that the equation $U(a) = 0$, where

$$U(a) = \sum_{\ell=1}^{j-1} y_\ell \log\psi(x_\ell,1) - \sum_{\ell=1}^{j-1}(1-y_\ell)\{1-\psi(x_\ell,a)\}^{-1}\psi(x_\ell,a)\log\psi(x_\ell,1) \tag{4}$$

has a solution. The solution is given by $a = \hat{a}_j$. Of course, once established, heterogeneity persists so that if a solution exists at some value of j then one exists at all j' where $j' \geq j$. It is worth mentioning that the above expression, based on the log-likelihood for a Bernouilli law, assumed one by one inclusion of patients. Were we to work with grouped inclusions, then the appropriate log-likelihood at each level is Binomial. However, this distinction only has the effect of introducing combinatorial constants into the equation and these all disappear upon taking derivatives. Equation 4 is therefore valid in the presence of grouped inclusions.

On the basis of the observed information matrix, we can calculate an approximate variance, $v(\hat{a}_j)$, for $\hat{a}_j$, which becomes more accurate with increasing sample size:

$$v^{-1}(\hat{a}_j) = \sum_{\ell=1}^{j-1}(1-y_\ell)\psi(x_\ell,\hat{a}_j)\log\psi(x_\ell,1)^2/\{1-\psi(x_\ell,\hat{a}_j)\}^2. \tag{5}$$

After each observation, we can calculate approximate $100(1-\alpha)\%$ confidence intervals for $\psi(d_i,\hat{a}_j)$ as $(\psi_j^-(d_i),\psi_j^+(d_i))$, where

$$\psi_j^-(d_i) = \psi\{d_i,(\hat{a}_j + z_{1-\alpha/2}v(\hat{a}_j)^{1/2})\}, \qquad \psi_j^+(d_i) = \psi\{d_i,(\hat{a}_j - z_{1-\alpha/2}v(\hat{a}_j)^{1/2})\},$$

and z_α is the αth percentile of a standard normal distribution. A recursion can be set up so that the mth iteration for $\hat{a}_j$, denoted $\hat{a}_j^{(m)}$ say, is given by

$$\hat{a}_j^{(m)} = \hat{a}_j^{(m-1)} - U(\hat{a}_j^{(m-1)})v(\hat{a}_j^{(m-1)}).$$

A starting value to the iterations, in the vicinity of the solution, is obtained by fitting the model to the mean response together with the mean dose level used. This equation solves immediately to give

$$\hat{a}_j^{(0)} = \frac{\log \sum_{\ell=1}^{j-1} y_\ell - \log(j-1)}{\log \sum_{\ell=1}^{j-1} \alpha_\ell - \log(j-1)}, \qquad j \geq 2.$$

The reparameterized power model (i.e., the model as described above except that a is replaced by $\exp(a)$) can be expressed as

$$\log\log\psi(d_i, a) = \log\log\psi(d_i, 0) + a\,,$$

still maintaining a translational form. For this model, equation (4) still holds in $\exp(\hat{a}_j)$ so that it is only necessary to take the logarithm of the solution to (4). The large sample variance will then tend to $\hat{a}_j^{-2} v(\hat{a}_j)$ and can be used to obtain approximate $100(1-\alpha)\%$ confidence intervals as above. A starting value for the iterations for the reparameterized power model will be the logarithm of the value provided above.

We are now in a position to assess the precision of the probability of toxicity at the currently recommended level. For sample sizes of 16 and 20, intervals calculated in the above way were shown to be reasonably reliable and meaningful (O'Quigley, 1992), especially after having carried out a Cornish–Fisher correction. We need bear in mind that samples much larger than these are unlikely to be attained so that simulations are probably more useful than theoretical asymptotics. For small sample sizes we need to be circumspect about appealing to large sample results. Even so, a table for sample size 12, not presented in O'Quigley (1992), showed overall performance to be as satisfactory as for the larger samples. For very small samples, below 10 say, the reliability of the confidence intervals becomes to some extent less of a practical concern since the intervals are mostly too wide to be of real help. This is an area that may warrant deeper investigation.

3.4 *Dose Allocation*

It is not necessary to carry out any integration as in Section 2.4 and we work directly with the maximum likelihood estimate of a_0. Using our model, this enables us to estimate the probability of toxicity at level d_i given the accumulated information on the first $j-1$ patient responses. Indeed the maximum likelihood estimator of the probability of toxicity at dose d_i for patient j is $\psi(d_i, \hat{a}_j)$, where $\hat{a}_j$ is assumed to exist. Next we allocate patient j to treatment level d_i such that $\Delta(\psi(d_i, \hat{a}_j), \theta)\,, i = 1, \ldots, k$, is minimized. The recommended dose will be the dose d_i such that $\Delta(\psi(d_i, \hat{a}_{n+1}), \theta)\,, i = 1, \ldots, k$, is minimized. We are at liberty to choose from a wide range of distance functions as well as weighted distances such that, for example, we always approach the target from below. Such a choice would depend upon the context. In this paper we only work with a simple quadratic distance.

4. Simulations

4.1 *Illustration*

It is instructive to look at a single, simple simulation to get a feel for how things might work in a practical setting. Operating characteristics for small and moderate samples are most usefully examined via large numbers of such simulations and this is considered in the next two sections. Before considering comparative operating characteristics, we simulated a single set of 16 responses from six levels having true toxic probabilities of 0.03, 0.22, 0.45, 0.60, 0.80, and 0.95, respectively. Here we limit our description to the case where we work initially with the traditional method. Having no natural units for the "doses" $x_1, \ldots, x_6$, we equate them to our baseline working model so that $x_i = \alpha_i, i = 1, \ldots, 6$, where

$$\psi(x_i, a) = \alpha_i^{\exp(a)}, \qquad i = 1, \ldots, 6,$$

and where $\alpha_1 = 0.04$, $\alpha_2 = 0.07$, $\alpha_3 = 0.2$, $\alpha_4 = 0.35$, $\alpha_5 = 0.55$, and $\alpha_6 = 0.7$.

The target toxic probability is 0.2 so that level 2 is the correct level. The first three patients are included at level 1 and none of the three experience toxicity. The next three are included at level 2 and again no toxicity is observed. Patients 7, 8, and 9 are treated at level 3 and two out of the three experience a toxicity. We now have heterogeneity among the responses so that the maximum

likelihood estimate for a_0 exists and is, in fact, equal to -0.335. The starting value to the iterations equals -0.412. The maximum likelihood estimates of toxicity at all six levels are 0.1, 0.149, 0.316, 0.472, 0.652, and 0.775, respectively. Level 2 therefore is the level with an estimated toxicity closest to the target level. Patient 10 is treated at this level and has no toxicity. The maximum likelihood estimate for a_0 moves to -0.275 and level 2 remains the nearest level to the target. Continuing in this way, we see that the last seven patients in the trial are all treated at level 2 which, after including 16 patients, is the recommended level for future use. Figure 1 provides a graphical representation of this trial.

4.2 *Setup to Simulations*

In this section we present some limited simulations and comment upon the results in the following section. It is not possible, of course, to cover all cases and our goal is to present a reasonably wide spectrum of situations in order to get a feel for how things might work in practice. We have limited ourselves to six dose levels for no better reason than this often occurs in practice and that, apart from extreme situations, i.e., two levels or above 20 levels, we believe our main conclusions hold more generally. We believe it is important to underline that our conclusions do not hinge upon the working model being true and so half of the simulations are generated according to the model, the other half generated from a model whose only constraint is that of monotonicity. Tables 1 and 2 concern final recommendation, whereas Tables 3 and 4 concern in-trial allocation. Each trial was repeated 2000 times.

Two versions of the two-stage continual reassessment method are presented in this section. In each version, the maximum likelihood estimates of the parameter are used after heterogeneity has been observed in the toxic responses. The two versions differ at the beginning of the trials.

CRM I: The original CRM is used at the beginning of the trial, i.e., the Bayesian estimates are employed until heterogeneity is observed, after which we use CRML.

CRM II: The traditional method is used at the beginning of the trial. In other words, three patients are included at the lowest level. We assume all three are not toxicities since, for such an eventuality, we would undoubtedly abandon the trial and rethink appropriate levels. If there areno toxicities among the three, then we escalate to the next level and include an additional three pa-

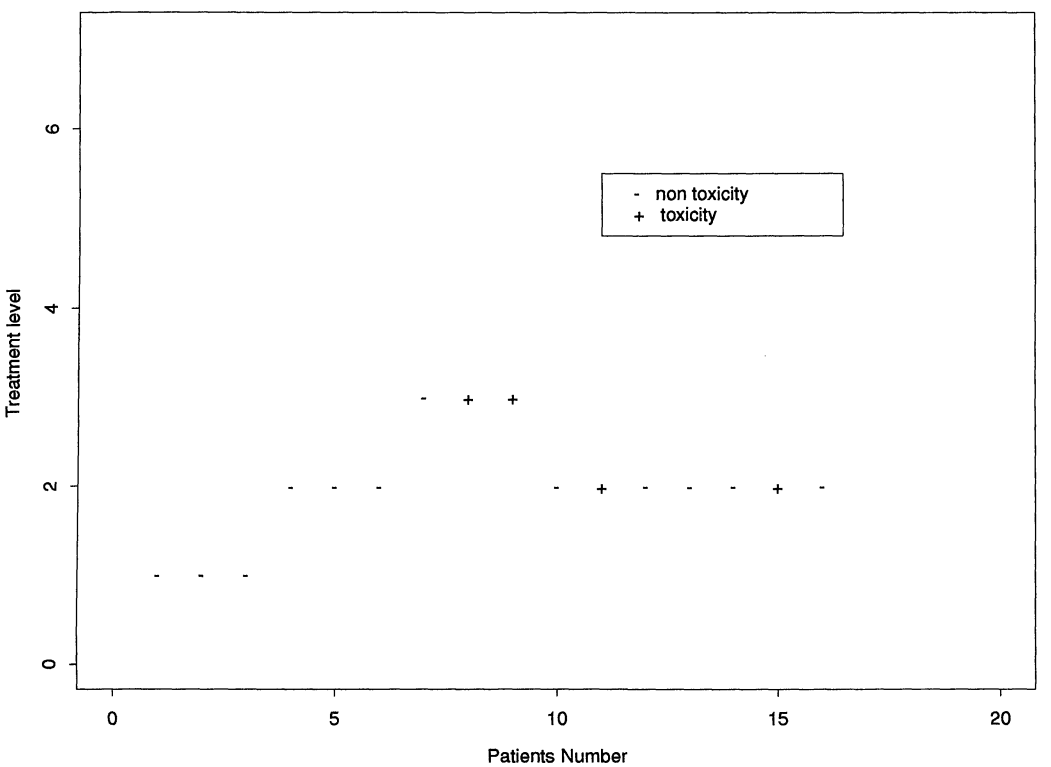

Figure 1. Results of simulated experiment.

Table 1
Data are generated according to the model

a_0		i:	% of recommendation for the ith level ($n = 16$)						% of recommendation for the ith level ($n = 25$)					
			1	2	3	4	5	6	1	2	3	4	5	6
−.68	R_i		.20	.26	.44	.58	.74	.83	.20	.26	.44	.58	.74	.83
	CRM		.58	.32	.09	.01	.00	.00	.63	.32	.05	.00	.00	.00
	CRMI		.57	.32	.11	.01	.00	.00	.60	.35	.05	.00	.00	.00
	CRMII		.56	.32	.11	.01	.00	.00	.62	.32	.06	.00	.00	.00
−.50	R_i		.15	.20	.38	.52	.70	.80	.15	.20	.38	.52	.70	.80
	CRM		.38	.40	.20	.02	.00	.00	.36	.46	.17	.01	.00	.00
	CRMI		.37	.39	.21	.02	.00	.00	.33	.50	.16	.01	.00	.00
	CRMII		.35	.41	.22	.03	.00	.00	.33	.50	.17	.01	.00	.00
0.0	R_i		.04	.07	.20	.35	.55	.70	.04	.07	.20	.35	.55	.70
	CRM		.02	.20	.57	.19	.01	.00	.01	.17	.63	.19	.00	.00
	CRMI		.03	.19	.50	.25	.03	.00	.01	.19	.60	.20	.00	.00
	CRMII		.03	.19	.50	.26	.02	.00	.01	.17	.59	.23	.01	.00
.42	R_i		.01	.02	.09	.20	.40	.58	.01	.02	.09	.20	.40	.58
	CRM		.00	.02	.29	.56	.13	.00	.00	.01	.26	.63	.09	.00
	CRMI		.00	.02	.25	.52	.20	.01	.00	.00	.23	.63	.13	.00
	CRMII		.00	.01	.27	.55	.14	.03	.00	.01	.22	.61	.16	.00
.92	R_i		.00	.00	.02	.07	.22	.41	.00	.00	.02	.07	.22	.41
	CRM		.00	.00	.02	.35	.51	.12	.00	.00	.00	.29	.64	.07
	CRMI		.00	.00	.01	.26	.54	.19	.00	.00	.00	.24	.63	.13
	CRMII		.00	.00	.02	.38	.38	.21	.00	.00	.01	.29	.55	.16
1.50	R_i		.00	.00	.00	.01	.07	.20	.00	.00	.00	.01	.07	.20
	CRM		.00	.00	.00	.03	.33	.64	.00	.00	.00	.00	.27	.73
	CRMI		.00	.00	.00	.01	.22	.77	.00	.00	.00	.00	.19	.81
	CRMII		.00	.00	.00	.06	.30	.64	.00	.00	.00	.01	.19	.80

tients. As soon as we observe heterogeneity among the responses, we use CRML and thereafter include patient by patient.

The simulation studies compare the results for the original CRM and the two versions of CRML, CRM I and CRM II.

The basic set up of the simulation is similar to that of O'Quigley et al. (1990). We consider six dose levels, say $x_1, \ldots, x_6$, and, having no natural units for the x_i, we equate them to our baseline working model so that $x_i = \alpha_i, i = 1, \ldots, 6$, where

$$\psi(i, a) = \alpha_i^{\exp(a)}, \qquad i = 1, \ldots, 6,$$

and where $\alpha_1 = 0.04$, $\alpha_2 = 0.07$, $\alpha_3 = 0.2$, $\alpha_4 = 0.35$, $\alpha_5 = 0.55$, and $\alpha_6 = 0.7$. The prior distribution of a is the Normal distribution $N(0, 1.34)$. We investigate two situations: (i) the real dose–toxicity relation coincides with $\psi(\cdot, a_0)$ for some a_0; (ii) the more usual case in which the model is misspecified, i.e, the real dose–toxicity curve is different from $\psi(\cdot, a)$. For each particular model, we shall calculate the posterior mean for CRM; the maximum likelihood estimate for CRM I and CRM II. Sample sizes of $n = 16$ and $n = 25$ were chosen for each set of simulations. Tables 1 and 3 contain simulation results for situation (i). The probabilities of toxicity for generating the data have the form $R_i = \alpha_i^{\exp(a)}$ for $i = 1, \ldots, 6$. The true parameter a takes values in $\{-0.68, -0.50, 0.0, 0.42, 0.92, 1.50\}$.

The main entries in Table 1 are the percentages for the final recommendation for the six levels using the approach of CRM, CRM I, and CRM II, respectively. Note that the percentages in each row may not add up to exactly 100% due to rounding. The main entries in Table 3 are the percentages of in-trial allocation of dose level to patients. For all tables the R_i denotes the true probability of toxicity at dose level i.

The main entries of Tables 2 and 4 parallel the results of Tables 1 and 3 for situation (ii) in which the observations are not generated from the working model. The probabilities of toxicity

Table 2
Data are not generated from the model

	% of recommendation for the ith level ($n = 16$)						% of recommendation for the ith level ($n = 25$)					
i:	1	2	3	4	5	6	1	2	3	4	5	6
R_i	.10	.20	.30	.45	.58	.70	.10	.20	.30	.45	.58	.70
CRM	.24	.40	.30	.05	.00	.00	.23	.45	.29	.03	.00	.00
CRMI	.24	.39	.29	.08	.00	.00	.20	.47	.29	.04	.00	.00
CRMII	.23	.41	.29	.06	.00	.00	.21	.48	.27	.03	.00	.00
R_i	.02	.11	.20	.35	.58	.70	.02	.11	.20	.35	.58	.70
CRM	.04	.22	.53	.20	.02	.00	.01	.22	.61	.16	.00	.00
CRMI	.04	.23	.47	.24	.02	.00	.02	.20	.57	.20	.01	.00
CRMII	.02	.24	.48	.25	.01	.00	.01	.23	.56	.19	.01	.00
R_i	.01	.07	.10	.20	.35	.50	.01	.07	.10	.20	.35	.50
CRM	.00	.05	.27	.49	.17	.02	.00	.03	.28	.54	.15	.00
CRMI	.01	.05	.23	.44	.23	.03	.00	.03	.23	.54	.19	.01
CRMII	.01	.08	.32	.43	.12	.04	.00	.05	.27	.51	.15	.02
R_i	.01	.05	.09	.15	.20	.40	.01	.05	.09	.15	.20	.40
CRM	.00	.03	.16	.33	.37	.12	.00	.01	.11	.34	.45	.08
CRMI	.00	.02	.11	.24	.43	.19	.00	.01	.10	.29	.48	.12
CRMII	.00	.05	.26	.38	.19	.11	.00	.02	.18	.40	.27	.08
R_i	.12	.22	.30	.45	.55	.70	.12	.22	.30	.45	.55	.70
CRM	.30	.33	.31	.06	.00	.00	.29	.40	.29	.02	.00	.00
CRMI	.29	.34	.28	.08	.01	.00	.27	.40	.29	.04	.00	.00
CRMII	.32	.40	.22	.06	.00	.00	.31	.45	.22	.02	.00	.00
R_i	.02	.08	.13	.23	.35	.50	.02	.08	.13	.23	.35	.50
CRM	.01	.09	.34	.40	.14	.02	.00	.06	.37	.45	.12	.01
CRMI	.01	.09	.30	.39	.19	.03	.00	.06	.32	.47	.15	.01
CRMII	.01	.13	.37	.38	.08	.03	.00	.09	.38	.40	.11	.01

that generate the data satisfy only $0 < R_1, \ldots, R_6 < 1$. This is the situation of practical interest in view of the restrictive nature of the one-parameter working model. It is nonetheless interesting to note that the results hinge only very weakly upon whether or not the working model is correct.

The simulation is carried out for six special cases. In each of the first four cases there exists a particular level which is exactly the target level, whereas in the last two cases one can only find a level that is the closest to the target level. Again, performance does not depend very much on this. If an existing level is close to the target, the method will tend to converge to that level. If the target lies midway between two available levels, then we tend to allocate to those two levels. The methods behave as we might intuitively expect in this regard.

4.3 *Some Comments on Simulation Findings*

Perhaps the most important take home message, for those cases studied at least and in so far as concerns final recommendation, is that there are rarely very marked differences between the methods. Additionally, the conclusions are much the same whether the working model is correct or not. In the case where the probability of toxicity at the first five levels is very low and only at level 6 do we reach the target, it is clear that grouped inclusions of three will have exhausted 15 patients before even getting there. For this case then the strong differences in operational characteristics are only to be expected. More generally, differences are much slighter. Tables 3 and 4 deal with in-trial allocation and here the differences between CRMII and the other two are more pronounced. Almost always the differences between CRM and CRMI can be ignored. Only in the last entry of Table 3, where the target is the highest level, do we detect any difference of note. As pointed out above, the lack of prior information enables CRMI to attain more easily the highest level. Conversely, had we increased the variance on the prior $g(a)$, then we would see the behavior of CRM approach that of CRMI. In this same situation in which the higher levels turn out to be the correct ones, it is clear, as expected, that both CRM and CRMI do better than CRMII, the latter spending too much time at the lower levels due to the slow escalation caused by grouping.

Table 3
Data are generated according to the model

			% of recommendation for the ith level ($n = 16$)						% of recommendation for the ith level ($n = 25$)					
a_0		i:	1	2	3	4	5	6	1	2	3	4	5	6
−.68	R_i		.20	.26	.44	.58	.74	.83	.20	.26	.44	.58	.74	.83
	CRM		.48	.30	.12	.07	.03	.00	.53	.30	.10	.04	.02	.00
	CRMI		.49	.28	.12	.08	.03	.00	.52	.31	.10	.05	.02	.00
	CRMII		.62	.28	.09	.01	.00	.00	.61	.29	.09	.01	.00	.00
−.50	R_i		.15	.20	.38	.52	.70	.80	.15	.20	.38	.52	.70	.80
	CRM		.40	.28	.18	.10	.04	.00	.38	.33	.19	.07	.02	.00
	CRMI		.40	.26	.18	.11	.04	.00	.38	.33	.19	.08	.03	.00
	CRMII		.50	.32	.16	.03	.00	.00	.44	.37	.16	.02	.00	.00
0.0	R_i		.04	.07	.20	.35	.55	.70	.04	.07	.20	.35	.55	.70
	CRM		.14	.18	.35	.24	.09	.01	.09	.19	.41	.24	.06	.01
	CRMI		.14	.17	.30	.27	.10	.02	.10	.20	.39	.24	.07	.01
	CRMII		.24	.30	.31	.13	.02	.00	.16	.25	.39	.18	.02	.00
.42	R_i		.01	.02	.09	.20	.40	.58	.01	.02	.09	.20	.40	.58
	CRM		.08	.06	.23	.39	.19	.04	.05	.04	.25	.47	.16	.03
	CRMI		.08	.06	.19	.39	.23	.05	.05	.04	.21	.45	.21	.04
	CRMII		.20	.20	.29	.23	.07	.00	.13	.14	.26	.35	.12	.00
.92	R_i		.00	.00	.02	.07	.22	.41	.00	.00	.02	.07	.22	.41
	CRM		.07	.01	.07	.32	.39	.15	.04	.01	.05	.32	.46	.13
	CRMI		.07	.01	.05	.28	.40	.20	.04	.01	.03	.27	.47	.18
	CRMII		.19	.19	.20	.24	.16	.02	.12	.12	.14	.25	.28	.08
1.50	R_i		.00	.00	.00	.01	.07	.20	.00	.00	.00	.01	.07	.20
	CRM		.06	.00	.01	.12	.36	.44	.04	.00	.01	.08	.33	.54
	CRMI		.06	.00	.01	.10	.29	.53	.04	.00	.00	.07	.27	.62
	CRMII		.19	.19	.19	.20	.19	.05	.12	.12	.12	.13	.21	.30

We might expect CRMII to have an advantage in situations in which the target level turns out to be among the lower or the lowest one. The simulation results indicate this to be the case as far as in-trial allocation is concerned, although not in an overwhelming way. As far as final recommendation is concerned, there is very little to choose between the methods, the ability of both CRM and CRMI to quickly react to toxicities meaning that all three methods tend to provide very similar recommendations. The fact that CRMII performs more conservatively as far as allocation goes, yet pays no apparent penalty for this conservatism in terms of final allocation, would strongly argue in favor of its use when we feel the lowest levels are likely to have nonnegligible toxic effects. When we suspect that the probabilities of toxicity are likely to be very small over most of the dose levels, then CRMII would not seem a good choice.

5. Conclusion

As far as the recommended level is concerned, large sample theory indicates there will be little to choose between the three approaches unless priors are strong. In the majority of situations we would only wish to work with weak priors. For a large number of situations examined here and elsewhere, the large sample approximation appears reasonably accurate for sample sizes as small as 16 (O'Quigley, 1992). We can thus conclude that deciding to work with one rather than any other of the three approaches will have little impact on the final recommendation. For in-trial allocation, the differences between the methods again disappear as sample size increases, but for small samples may need to be considered. The simulations of the previous section tend to confirm our intuition. When sample size is as small as 16 and the highest level is indeed the correct level, then CRMII does poorly. For the other extreme, in which the correct level is the lowest one, CRMII performs better than the other two designs in terms of in-trial allocation while maintaining almost identical results in terms of final recommendation. Intermediary designs, having properties lying between CRMI and CRMII, could also be considered: for example escalating every two instead of three observations or, like Barry Storer's suggestion for a hybrid Up-and-Down design, after single observations (Storer, 1989) or a combination of the two.

Table 4
Data is not generated from the model

	% of recommendation for the ith level ($n = 16$)						% of recommendation for the ith level ($n = 25$)					
i:	1	2	3	4	5	6	1	2	3	4	5	6
R_i	.10	.20	.30	.45	.58	.70	.10	.20	.30	.45	.58	.70
CRM	.27	.32	.22	.12	.05	.01	.26	.36	.24	.10	.04	.00
CRMI	.29	.30	.20	.14	.06	.01	.26	.35	.24	.11	.04	.01
CRMII	.41	.35	.19	.04	.00	.00	.36	.38	.22	.04	.00	.00
R_i	.02	.11	.20	.35	.58	.70	.02	.11	.20	.35	.58	.70
CRM	.14	.20	.33	.24	.08	.01	.10	.20	.41	.23	.05	.01
CRMI	.14	.18	.30	.27	.10	.01	.10	.19	.37	.26	.07	.01
CRMII	.24	.32	.30	.12	.02	.00	.17	.30	.36	.15	.02	.00
R_i	.01	.07	.10	.20	.35	.50	.01	.07	.10	.20	.35	.50
CRM	.09	.07	.21	.36	.21	.06	.06	.06	.23	.41	.20	.04
CRMI	.10	.07	.17	.34	.24	.08	.06	.06	.20	.39	.23	.06
CRMII	.21	.26	.28	.19	.06	.00	.14	.18	.28	.29	.10	.02
R_i	.01	.05	.09	.15	.20	.40	.01	.05	.09	.15	.20	.40
CRM	.08	.05	.13	.27	.32	.14	.05	.04	.12	.30	.36	.12
CRMI	.08	.04	.10	.24	.34	.19	.05	.04	.10	.25	.38	.18
CRMII	.21	.24	.27	.20	.08	.01	.13	.16	.25	.25	.15	.05
R_i	.12	.22	.30	.45	.55	.70	.12	.22	.30	.45	.55	.70
CRM	.33	.23	.22	.14	.06	.01	.32	.28	.24	.11	.04	.01
CRMI	.35	.22	.20	.16	.07	.01	.31	.28	.23	.13	.05	.01
CRMII	.49	.32	.15	.04	.00	.00	.41	.36	.19	.04	.00	.00
R_i	.02	.08	.13	.23	.35	.50	.02	.08	.13	.23	.35	.50
CRM	.11	.10	.23	.32	.19	.06	.07	.09	.27	.36	.17	.04
CRMI	.11	.09	.20	.31	.22	.07	.07	.09	.24	.36	.20	.05
CRMII	.23	.28	.28	.17	.04	.00	.15	.22	.31	.24	.07	.01

Overall, the possible choices, as far as final recommendation is concerned, behave similarly. In-trial allocation will be slightly affected and this may be a consideration. The main purpose of this paper then is not to make claims that either of the outlined approaches can improve on what is already available. The purpose is rather to add an angle to the CRM construction which might, hopefully, make the approach more widely acceptable. In addition, since more is generally known about maximum likelihood estimation than its Bayesian counterpart, the current work should facilitate deeper theoretical understanding of CRM and related procedures.

Acknowledgements

This work was supported in part by a Contrat de Recherche 1817 by the French Association Pour la Recherche Sur le Cancer. The referees and associate editor provided a very detailed list of concerns on a previous version of this work, in particular questions concerning the in-trial allocation probabilities. We would like to thank them for this input.

Résumé

La méthode dite de réévaluation séquentielle (CRM), développée par O'Quigley, Pepe et Fisher (1990, *Biometrics* **46,** 33–48), s'appuie en grande partie sur une approche Bayesienne. Le début de l'expérience, ainsi que les remises à jour successives se font tout naturellement dans un contexte Bayesien. Dans ce papier on démontre que ce contexte peut être aussi bien remplacé par une méthodologie s'appuyant sur l'inférence classique à partir de la vraisemblance. Les caractéristiques principales de la CRM demeurent les mêmes. Les propriétés asymptotiques, en particulier, restent inchangées à moins d'utiliser une loi a priori de Dirac. Pour les échantillons aux faibles effectifs, en ce qui concerne le niveau recommandé, les simulations confirment que les deux approches sont très semblables. Quant aux niveaux utilisés pendant l'essai il peut y avoir quelques petites différences et celles-ci font le sujet d'une discussion. Une approche fondée sur la vraisemblance ne peut démarrer avant de recontrer une première réaction toxique. Ceci, puisque la vraisemblance s'avère monotone

alors que toutes les réponses sont des non toxicités. A ce stade de l'expérience nous proposons alors l'utilisation du schéma standard dit Up-and-Down ou du schéma CRM Bayesien jusqu'à ce qu'une réaction toxique se produise. A ce moment là l'approche bascule et désormais c'est la nouvelle approche basée sur la vraisemblance qui est utilisée.

References

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

Korn, E. L., Midthune, D., Chen, T. T., Rubinstein, L. V., Christian, M. C., and Simon, R. (1994). A comparison of two Phase I trial designs. *Statistics in Medicine* **13**(18), 1799–1806.

Mick, R. and Ratain, M. J. (1993). Model-guided determination of maximum tolerated dose in Phase I clinical trials: Evidence for increased precision. *Journal of the National Cancer Institute* **85**(3), 217–223.

O'Quigley, J. (1992). Estimating the probability of toxicity at the recommended dose following a Phase I clinical trial in cancer. *Biometrics* **48,** 853–862.

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.

O'Quigley, J. and Chevret, S. (1991). Methods for dose finding studies in cancer clinical trials: A review and results of a Monte Carlo study. *Statistics in Medicine* **10,** 1647–1664.

Shen, L. Z. and O'Quigley, J. (1996). Consistency of continual reassessment method in dose finding studies. *Biometrika*, in press.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society, Series B* **43**(3), 310–313.

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

Received March 1994; *revised May–October* 1995; *accepted November* 1995.