Biometrics 59, 430–440
June 2003

Continual Reassessment Method for Ordered Groups

John O'Quigley* **and Xavier Paoletti**

Department of Biostatistics, Institut Curie, 75005 Paris, France
**email:* oquigley@galois.ucsd.edu

Summary. We investigate the two-group continual reassessment method for a dose-finding study in which we anticipate some ordering between the groups. This is a situation in which, for either group, we have little or almost no knowledge about which of the available dose levels will correspond to the maximum tolerated dose (MTD), but we may have quite strong knowledge concerning which of the two groups will have the higher level of MTD, if indeed they do not have the same MTD. The motivation for studying this problem came from an investigation into a new therapy for acute leukemia in children. The background to this study is discussed. There were two groups of patients: one group already received heavy prior therapy while the second group had received relatively much lighter prior therapy. It was therefore anticipated that the second group would have an MTD higher or at least as high as the first. Generally, likelihood methods or, equivalently, the use of noninformative Bayes priors, can be used to model the main aspects of the study, i.e., the MTD for one of the groups, reserving more informative Bayes modeling to be applied to the secondary features of the study. These secondary features may simply be the direction of the difference between the MTD levels for the two groups or, possibly, information on the potential gap between the two MTDs.

Key words: Bayesian inference; Clinical trials; Continual reassessment method; Dose-finding studies; Maximum tolerated dose; Ordered groups; Pharmacology; Phase I designs; Toxicity; Two-sample problem.

1. Introduction

The purpose of a phase I clinical trial is the identification of a maximum tolerated dose (MTD), usually interpreted as the percentile of some unknown dose-toxicity curve. In such studies we can reasonably well anticipate running into substantial patient heterogeneity, the simplest case occurring where patients can be categorized into two distinct prognostic groups. A common example, arising in cancer studies, concerns the degree of previous treatment which is often related to the patient's ability to tolerate a new treatment regimen (Rosenblum et al., 1999). Another example concerns patients with acute leukemia where it has been observed that, even after having carried out the usual weight adjustments, younger patients are better able to tolerate aggressive treatments. In many situations, groups can be ranked according to the value of a grouping covariate and it is quite plausible that, although we know little about the underlying dose-toxicity relation and, consequently the right dose to recommend for either of the groups, we may be able to say something about the difference between the groups. For instance it may be known in advance that one group will be able to tolerate a higher dose than the other. For very small samples, which are common in this setting, we may be able to do no better than aim to identify some kind of average over the different prognostic groups. However, there will be cases where we have enough information to carry out a more refined analysis, the goal being to identify the most appropriate level for both (or more) groups.

An extension of the continual reassessment method (CRM), specifically focused on this problem, has been developed by O'Quigley, Shen, and Gamst (1999). The development allows for two groups of patients to be treated in the same study simultaneously. The objective in the work of O'Quigley et al. (1999) is to target the MTD for both groups of patients. No assumptions were made about any possible ordering between the groups or about just how far apart the MTDs might be. Although each group may have a different MTD, the idea was to share any common information between the two groups. The design was likelihood based, specifically a two-stage design (Møller, 1995; O'Quigley and Shen, 1996; O'Quigley, 2001) in which initial escalation did not lean upon any model, followed by a model-guided stage employing sequential probability estimates based on likelihood inference. We recall this method in Sections 2 and 3.

It is possible, and potentially useful, to put yet further structure on the two-sample CRM. In this article, we study an extension of the two-sample CRM that incorporates external information about the group differences, in particular the direction of any possible difference as well as limits on the magnitude of a possible difference. We will see that it can be useful to use such information in two-stage designs, for the initial escalation stage as well as the sequential model–based stage of the trial. For two-stage designs, the dose escalation stage allows for considerable flexibility that can be exploited to advantage (Section 3.1). A very simple example helps illustrate this point: Suppose we know that group 2 is more robust to toxicity than group 1. Should the first entered patient in group 2 take place after several patients have already been entered into group 1, the current recommendation for group

1 being level i, then it makes sense to treat this patient at level i, rather than at lower levels.

Current practice will mostly disallow such designs, the guiding principles tending to initiate experimentation for both groups at low, if not the very lowest, levels. If we are in doubt about the ordering, then it may be difficult to do better than this, but whenever we have strong prior information on the ordering then, clearly, we would like to use such information. The introduction of such additional information into the picture, the direction of the difference between the two groups or, possibly, assigning different degrees of plausibility to the final recommendations being more than one level apart provides potentially useful added structure to the problem (Sections 3.2–3.4).

A Bayesian framework provides the natural setting for dealing with such problems, in particular in enabling the experimenter to moderate the strength of additional assumptions. The added structure translates itself as greater confidence in final recommendations. As always there is a price to pay for potential gains, the price being the possibility that precision will be outweighed by bias should the additional assumptions be incorrect. Part of our study here is to look at this balance and, in addition to simulations showing the benefits that can be achieved, we include a study of the size of the penalty we pay if we turn out to be quite wrong. The motivation for this investigation arose from a two-group dose-finding study in acute leukemia in children. We present our findings and dose recommendations for that study and discuss the kinds of options and considerations that might form the basis for constructing appropriate designs for similar studies.

2. Designs for Two Groups

The basic structure is described in O'Quigley et al. (1999) where the details can be found. We recall the main ideas here, rather briefly, before proceeding with those aspects specific to the ordering problem. For the jth included patient we use the indicator variable $Z_j = 0, 1$ to identify to which one of the two groups the patient belongs. Alongside the group indicator Z_j for the jth patient ($j = 1, \ldots, n$), we denote the administered dose, taking values in a discrete set of doses $d_1, \ldots, d_m$, by X_j. The two groups share the m available dose levels. The toxic response to the treatment is coded by a binary variable Y_j, $j = 1, \ldots, n$: 1 for a dose limiting toxicity (DLT) and 0 otherwise. Each dose level is then associated with a probability of DLT that increases with the value of the dose. We can write $P(Y_j = 1 \mid X_j = d_k : Z_j) = R(d_k, Z_j)$ where $R(d_1, 0) < R_i(d_2, 0) < \cdots < R_i(d_m, 0)$ and $R(d_1, 1) < R_i(d_2, 1) < \cdots < R_i(d_m, 1)$. The set of the m doses and their probabilities of toxicity in each group constitute the dose-toxicity relations. The main goal of a phase I clinical trial with group heterogeneity is to identify, for each of the available groups, the MTD. This is defined as the dose where the probability of toxicity is closest to some given value θ. Thus the MTD for the group identified by $Z_j = 0$ is $d_{0i} \in \{d_1, \ldots, d_m\}$ where d_k is the value that minimizes $|R(d_k, 0) - \theta|$. A parallel definition holds for the group identified by $Z_j = 1$. Suppose that an investigator plans a sequential trial with n subjects on m dose levels. The basic idea of the method is to sequentially update our $2m$ estimates of the probabilities of toxicity in both groups. This is done through a working two-parameter model after each inclusion of cohorts of 1 or more patients. The level recommended to the patients within a cohort minimizes the distance between the estimated probability of toxicity at this level and θ, for each member of that particular cohort, given the group status of the members of the cohort. Each patient is then included at the best current estimate of the MTD for that particular patient given his or her group status.

3. Two-Group Designs Incorporating Ordering

The usual CRM uses an underparametrized working model, specifically, for the case of a homogeneous sample, a one-parameter working model. Reasons for not being able to work with a richer two-parameter model are outlined in Shen and O'Quigley (1996) and O'Quigley (2001). In the current setting, where we have an indicator variable identifying a potential source of heterogeneity, we again use underparameterized models. One approach would be to use two one-parameter models separately, although this would not make use of any shared information. Instead we express the two one-parameter models as a single two-parameter model. Thus, for $k = 1, \ldots, m$ we can express a general formulation as $P(Y_j = 1 \mid X_j = d_k : Z_j) = \psi(d_k, a, bZ_j)$ where $(a, b) \in \mathcal{A} \times \mathcal{B}$ is the pair of unknown one-dimensional parameters. The parameter b models a shift between the two groups. We restrict our attention to the power model

$$P(Y_j = 1 \mid X_j = d_k : Z_j) = \psi(d_k, a, bZ_j) = \alpha_k^{\exp(a+bZ_j)} \quad (1)$$

in which $\alpha_k \in (0, 1)$ and $\alpha_k < \alpha_\ell, k < \ell$. As usual in the CRM context, the models are underparameterized and not assumed to provide an accurate global fit to the true dose-toxicity relations. The only requirement is that they have sufficient flexibility to provide a local fit (Shen and O'Quigley, 1996; O'Quigley et al., 1999). The two-group power model meets the necessary requirements although there are numerous other potential candidates, limited only by some simple technical conditions that are described in Shen and O'Quigley (1996) and O'Quigley et al. (1999). The specific value $b = 0$ corresponds to absence of patient heterogeneity, i.e., that we have a single homogeneous group. Moreover, for any fixed values of x and a, for all b, the difference $\psi_2(x, a, b) - \psi_1(x, a)$ has the same sign. After j subjects have been included in the study, we observe (x_ℓ, y_ℓ, z_ℓ), where x_ℓ is the dose allocated to patient ℓ belonging to group z_ℓ, and y_ℓ his/her response, for $\ell = 1, \ldots, j$. In the first instance inference is based upon the likelihood. Solutions to the likelihood equations are assumed to exist. This will be the case when, for both groups, we have at least one DLT and one non-DLT (O'Quigley et al., 1999). Our calculations are then based on the likelihood $L_j(a, b)$ and its derivatives following our observations on the first j inclusions. We denote the derivatives $U_j^a(a, b), U_j^b(a, b), I_j^{aa}(a, b) I_j^{bb}(a, b)$, and $I_j^{ab}(a, b)$ where, for example,

$$\begin{aligned} U_j^a(a, b) &= \partial \log L_j(a, b)/\partial a \\ I_j^{ab}(a, b) &= \partial^2 \log L_j(a, b)/\partial a \partial b. \end{aligned} \quad (2)$$

Note that before any toxicity has been observed in both groups, likelihood-based estimates are not available to us, or, at least, arising on the boundary of the parameter space, are

not very helpful. Any CRM model will not actually be used until the first toxicity is encountered. We therefore, as for the usual likelihood-based CRM (O'Quigley and Shen, 1996), split the design into two stages and exploit the possibility of using different initial escalation schemes. Two-stage designs and maximum likelihood estimation must necessarily accompany one another in these designs.

3.1 *Two-Stage Designs*

As long as the first two moments exist for any prior, we have toxicity estimates, based only partially or not at all on the observations. Thus from the first included patient the Bayesian approach is operational. A maximum likelihood approach, on the other hand, only becomes operational when the estimates exist, i.e., when the responses are not all non-DLTs. Maximum likelihood approaches then typically require a two-stage design: a model-guided stage preceded by an early initial escalation stage. Rather than view this estimating difficulty as a weakness, we can exploit it to our advantage. Preclinical investigations and other considerations may come into play in establishing the initial escalation scheme and starting dose level. Any CRM model and iterative estimation/allocation will not actually be used until the first toxicity is encountered.

Escalation is subject to ethical requirements and has two, operationally conflicting, guidelines. The first is that we wish to proceed cautiously and not overshoot the target too quickly, thereby putting an unacceptably large number of patients at risk for toxic side effects. The second is that we would like to avoid treating too many patients at levels so far below the target that the probability of seeing any treatment benefit is almost negligible. Addressing both guidelines simultaneously achieves a desirable result, that of reducing the number of patients required to complete such studies. This is not at all obvious, but follows from the fact that our estimating power is focused on a point rather than a whole curve. Nonetheless, additional improvements may be achieved by integrating into the likelihood equation any external information we may have about the group differences. Although this takes place mostly at the second, model-guided, stage it can also be a feature in the first stage, the most obvious example being the inclusion of the first entered patients from one group at levels at least equal to those being recommended for the other group. There is a broad range of possibilities for taking into account group differences. For instance, one can consider completely independent escalation stages in each arm. This is the design used by O'Quigley et al. (1999) in their simulations. Although of wide applicability, this design may not be the most efficient. An alternative approach that would lead to improved efficiency would be, for the escalation stage, to combine both arms into a single group. This makes sense as long as there is no real indication of any difference between the groups. Neither of these two designs incorporates any ordering assumption. The ethical stance here, which will partly translate itself in efficiency terms, is that, until the data tend to indicate the presence of group heterogeneity, we treat the study as a single group. However, we could also adopt an intermediary position. Consider for example the following design. Suppose that patients from group G_2 are known to better tolerate aggressive treatments than subjects from G_1. Escalation designs take place separately in both groups according to any reasonable scheme (an example of the one-group case is given in Table 1). Added to any such design is the incorporation of group ordering.

Table 1
Simple example of first stage in two-stage design

Allocated dose	d_1	d_2	d_3	d_3	d_4	d_4	d_5	d_5	d_5	etc.
Patient no.	1	2	3	4	5	6	7	8	9	etc.

In practice this means we impose a design restriction such that dose allocation in G_2 must be at least equal to the dose currently allocated in G_1. Such a design has the consequence that skipping doses is allowed in G_2, but not in G_1. For instance, suppose that the 3 first patients belonged to G_1 and that they respectively well tolerated the doses d_1, d_2, d_3. If the 4th patient is from G_2, he/she will be directly included at d_3 (Figure 1). As soon as the first toxicity is observed in a group, let us say G_1, we have heterogeneity among the responses in this group. We can proceed in G_1 using the usual CRM design. We fit the working model, $\psi_1(d_k, a)$, to data from G_1,

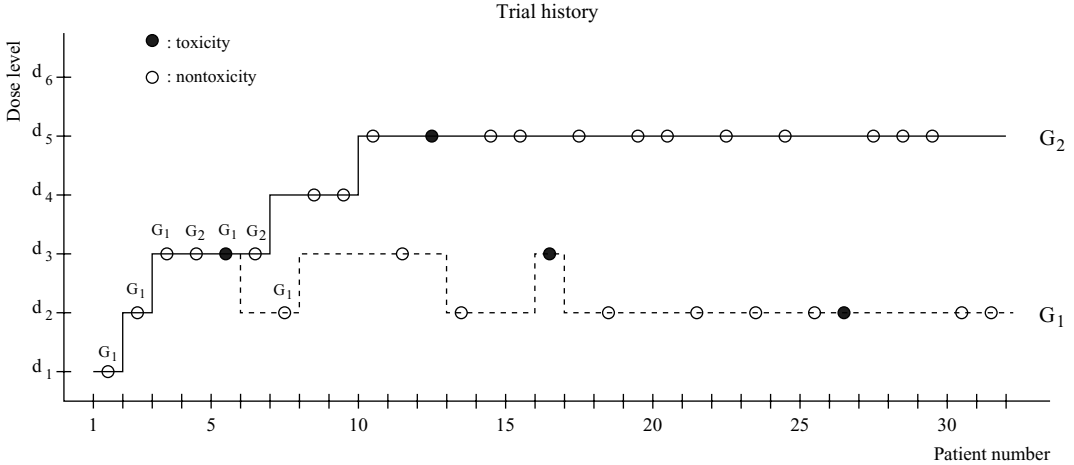

Figure 1. Simulated example of the two-stage design when G_2 is known to better tolerate treatments.

which provides us with estimates of the m probabilities of toxicity. For G_2, escalation continues according to the ordering requirement. Once we encounter a toxicity in G_2, heterogeneity is obtained in both groups (complete heterogeneity). We can fit a two-parameter model to the recorded data. A simulated example of such a design is shown in Figure 1. The x-axis displays the patient number whereas the y-axis gives the level being used. The overall sample size was fixed at 32 patients. Observe that although the first three patients belong to G_1, known to be the weaker group, the first patient from G_2 is included at dose level d_3. Once we have obtained heterogeneity among the responses from both groups, then recommendations are guided by probability estimates, based upon the two-sample CRM iterative scheme (O'Quigley et al., 1999). The next section presents an analysis making extra use of a known ordering.

3.2 *Prior Information on Group Differences*

Prior information concerning the group difference or indeed any other aspects of the experiment is most naturally integrated into the inferential procedures of our designs via a Bayesian framework (Whitehead, 1997; Babb, Rogatko, and Zacks, 1998). We set independent prior distributions for a and b, denoted by $f(a)$ and $g(b)$, and the joint prior density $h_0(a,b) = f(a) \times g(b)$. If $h_{j-1}(a,b)$ is the joint density for (a,b) after $j-1$ patients, then we write simply;

$$h_j(a,b) = K_j^{-1} h_0(a,b) L_{j-1}(a,b) \tag{3}$$

where $K_j = \int_{\mathcal{A}} \int_{\mathcal{B}} h_0(a,b) L_{j-1}(a,b)\, da\, db$. We can then use the mode as a Bayesian estimate; $(\hat{a}_j, \hat{b}_j) = \max_{a,b}\{h_j(a,b)\}$ which, in turn, provide us with estimates of the probabilities of toxicity at the m levels in both groups, i.e., $\psi(d_1, \hat{a}_j, 0)$, $\ldots, \psi(d_m, \hat{a}_j, 0)$ and $\psi(d_1, \hat{a}_j, \hat{b}_j), \ldots, \psi(d_m, \hat{a}_j, \hat{b}_j)$. The recommended dose for the $(j+1)$th patient minimizes either $|\psi(d_k, \hat{a}_j, 0) - \theta|$, if he/she belongs to the first group, or $|\psi(d_k, \hat{a}_j, \hat{b}_j) - \theta|$, if he/she belongs to the second one. After each inclusion or cohort inclusion, we reassess our knowledge of both dose-toxicity relations. If no stopping rule is used, the trial halts once all patients have been included.

In most phase I trials, treatments are generally administered to human patients for the first time and very little is known about the true shape of dose-toxicity relations. It may then be preferable to work with a noninformative prior on a. Nevertheless, as for the examples of the introduction, it is quite plausible that, although we know little about the underlying dose toxicity, expressed via a, we may be prepared to say something much stronger about any difference between the groups. For the situation of heavy previous treatment we would anticipate the less-treated group to have a lower dose-toxicity curve. This information is then less diffuse than that concerning the overall rate and can be expressed via the prior distribution on b. Expressed in more Bayesian terms, we use an improper prior $f(a)$ for a, such as a uniform distribution over the real line and most often a normal or gamma distribution for b of mean μ_b and variance σ_b^2. The mean of this law would reflect the anticipated difference, and the variance, the amount of precision we wish to associate to any anticipated difference. The experimenter can choose both μ_b and σ_b^2 in light of available knowledge from earlier studies. When the distribution for b is improper, then the joint prior distribution itself, $h_0(a,b) = f(a) \times g(b)$, is also improper. Another way to view this is to consider that part of the experiment is being treated with informative Bayes priors and the other part in, arguably, a more objective way, based on the likelihood. Note that the case where both $f(a)$ and $g(b)$ are noninformative on the whole parameters space is equivalent to the two-sample CRM and would produce exactly the same operating characteristics and the same results if the same initial stage is employed. We now focus on two specific cases in which these ideas can be of practical interest.

3.3 *Modeling a Known Ordering of Unknown Magnitude*

In many trials where a covariate is felt to be of prognostic value, the group direction may be known from previous studies even though its magnitude remains uncertain. This is a very common situation. We can express the information that one group has lower toxicity probabilities than the other (i.e., stronger than the other) by constraining b to only take positive values. This can be immediately accomplished by reparameterizing b to $\exp(b)$ so that, for any value of $b \in \mathcal{B}$, the direction of the group difference never changes. Should the accumulating data be in contradiction with the ordering assumption, then the parameter estimate, $\hat{b}$, will tend toward the boundary of the parameter space, which is zero. We would end up treating patients from both groups at the same level. A rigid parameterization, together with maximum likelihood estimation, implies the irreversibility of our ordering assumption. This may be okay in some cases, but can be relaxed by appealing to a Bayesian formulation. We can assume a given ordering with high probability, but nonetheless allow the data to have its say, overriding our initial assumptions if these assumptions turn out to be in contradiction with the the observations themselves.

As an example, let us suppose that G_1 is expected to be more at risk than G_2 for the population under study. In this case, using equation (1), it is unlikely that b be negative. The prior distribution $g(b)$ can be chosen to put small, but nonnull, weight on negative values of b. One can take $g(b)$ to be normal and fix the mean and the variance such that $\int_{\mathbb{R}^-} g(b)db < \alpha$ where α is small. If this is felt not to be flexible enough, the prior can be constructed as a mixture of truncated distributions over the whole definition set $\mathcal{B}$. The unknown magnitude of the group difference can be quantified by choosing a large variance for $g(b)$.

3.4 *Modeling a Known Ordering of Given Magnitude*

We now consider the practical situation in which, in addition to the group ordering, investigators can also say something about its expected magnitude. $f(a)$ is uniform on the real line and $g(b)$ $(b \in \mathcal{B})$ is chosen to be normal $\mathcal{N}(\mu_b, \sigma_b^2)$. To tie down the design, we have to give consideration to the choice of μ_b and σ_b^2. Denoting the difference between the two MTDs, d_{01} and d_{02}, by ℓ and assuming that the associated probabilities of toxicity at these levels are not too distant from θ, then the aim is to identify a^* and μ_b such that

$$\psi_1\big(d_k, a^*\big) \simeq \psi_2\big(d_{k+\ell}, a^*, \mu_b\big) \simeq \theta. \tag{4}$$

Of course the parameter a^*, which enables us to identify d_{01}, is unknown in any real experiment. Thus μ_b will generally

Table 2
Examples of values of μ_b for an expected difference of one level (right) or two levels (left)

$d^*_{(1)}$	$d^*_{(2)}$	μ_b	$d^*_{(1)}$	$d^*_{(2)}$	μ_b
1	3	0.69	1	2	0.19
2	4	0.93	2	3	0.50
3	5	0.99	3	4	0.43
4	6	1.08	4	5	0.56
-	-	-	5	6	0.52
	$\bar{\mu}_b = 0.92$			$\bar{\mu}_b = 0.42$	

depend on both the coding $(\alpha_k, k = 1, \ldots, m)$ and the location of d_{01}. However, for the particular case of the exponential power model, equation (4) becomes:

$$\mu_b \simeq \ln \ln \alpha_k - \ln \ln \alpha_{k+\ell}, \tag{5}$$

in which a^* no longer appears. This is helpful, but we are not yet quite out of the woods since the solution for μ_b depends upon the value of α_k and therefore varies from one level to another. For instance, for the coding (0.04, 0.07, 0.2, 0.35, 0.55, 0.7), an expected difference of either one or two levels leads to the values of μ_b given in Table 2.

At each level we have a different value for μ_b. This is still awkward but note that, in the right-side part of the table, for all levels, apart from the first, a good approximation for any level is given by the average value close to 0.5. This value, $\bar{\mu}_b$, provides a logical choice for the mean of the prior normal law. The only problem that still causes us some concern is the code at level 1. However, the choice of the codes, $\alpha_k, k = 1, \ldots, m$, is largely arbitrary and, mostly, does not affect the asymptotic convergence property of the method (see Shen and O'Quigley, 1996). The choice does have an influence on the behavior of the method for finite samples and deeper study of this question (the best choice of coding) still needs to be carried out. Nevertheless, a criterion that would guide the coding would be a choice such that we obtain the same ratio $\ln\{(\ln \alpha_{k+\ell})/(\ln \alpha_k)\}$ at all levels $k = 1, \ldots, m - \ell$.

Table 3
Dose coding giving a constant ratio: $(\ln \alpha_{k+\ell})/(\ln \alpha_k) = 0.4$

d_k	d_1	d_2	d_3	d_4	d_5	d_6
α_k	0.04	0.11	0.28	0.41	0.60	0.70

Table 3 gives an example of a coding for $\ell = 2$ whose associated operating characteristics, based on a simulation study on various dose-toxicity relations, appears promising. Similarly, one could decide to recode α_1 so that a good approximation is the mean through all the levels.

Figure 2 shows the history of a simulated trial where the MTD in group 2 was anticipated to be higher than that for group 1. The chosen prior here leaned on an initial estimate that the two MTDs would be two levels apart. It is very interesting to compare this to Figure 1, where knowledge of potential group differences is only exploited during the initial escalation stage. The effect does not appear at all strong, but there is a clear impression that the prior information results in pulling upward, more than would happen in the absence of such information, the level for group 1. Also, no doubt, the level for group 2 is pulled downward, the parameterization effecting a compromise on the recommended levels via the shared information. In both cases we conclude that the MTDs are three levels apart, but this conclusion is arrived at a little earlier in the second case, despite, or most probably, because of, the noisier path that group 1 follows in this situation compared with that shown in Figure 1.

The parameter σ_b^2 quantifies the precision we wish to associate with our prior assumptions. The more confident in the assumption we are, the smaller the variance and the more informative the prior. We could be more formal here, fixing the variance in such a way that a precision interval for b corresponds to a precision interval for d_k. This would necessarily be approximate since the dose levels themselves are discrete. However, such a formal approach implies a stronger belief in

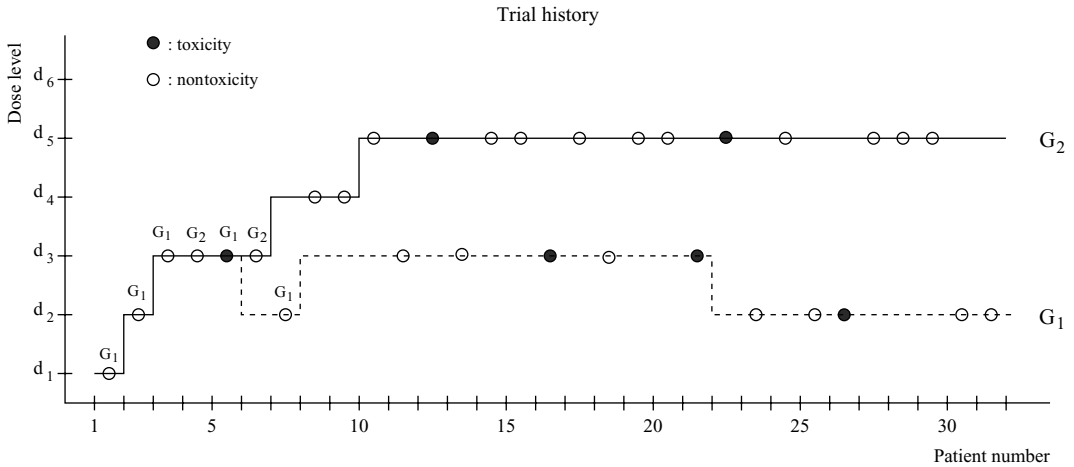

Figure 2. Simulated example of the two-stage two-sample CRM with informative prior.

our models that can be warranted in this context. It seems more reasonable, and certainly more reassuring to the user, to simply investigate a number of candidate values for σ_b^2 via simulations over classes of potential situations that we can expect to arise. On the basis of these, we can choose a value of σ_b^2 that provides a satisfactory compromise between an inflexible assertion of how far apart the levels are ($\sigma_b^2 = 0$) and knowledge that is so vague (σ_b^2 very large) that the operating characteristics coincide with those for the design in which the ordering information is only used in the initial escalation stage.

4. Simulations

4.1 *Simulations Setup*

This section is devoted to studying the behavior of the two-stage two-sample design, compared to the two-sample CRM, exclusively based on likelihood inference, in various situations. In order to generate responses of "simulated patients," we specified dose-toxicity relations. It is desirable that our main conclusions do not depend strongly upon the working model being true, whatever the chosen model is (logistic, probit, power, etc.). We therefore chose four arbitrary dose-toxicity relations (cf. Table 4), none of which were generated from the working model, to illustrate the main aspects of the method.

The two proposed extensions to the two-sample CRM—the escalation stage and the modeling of a known intergroup difference—enable us to incorporate external information in both the design and the statistical analysis of the trial. The basic setup of the simulations is similar to previous articles on the CRM. We consider $m = 6$ dose levels with the codes 0.04, 0.07, 0.2, 0.35, 0.55, 0.7. The $n = 32$ subjects are randomly assigned to either G_1 or G_2 with probability one-half. The target is $\theta = 0.2$. Each trial is replicated 5000 times. In the two-sample CRM and the two-stage two-sample CRM, the trial is split into two arms from the very start of the experiment, which involves two separate escalation stages. Both groups start at the first dose d_1 and dose levels are cleared after having observed the sequences of nontoxicities shown in Table 5.

In addition, in the two-stage two-sample CRM, group G_1 is taken to be more at risk than patients from G_2 and dose allocation during the first stage in this latter group is constrained to never be inferior to that for G_1's current MTD.

After heterogeneity among the toxic responses is obtained, the power model of equation (1) comes into play. The prior distribution $f(a)$ is chosen to be uniform on the real line and

Table 4
Dose-toxicity relations

Case	Gr. i	Probability of toxicity $R_i(d_k)$					
A	1	.07	.23	.31	.35	.45	.57
	2	.01	.03	.05	.09	.20	.40
B	1	.08	.20	.35	.50	.70	.80
	2	.01	.05	.18	.40	.55	.7
C	1	.02	.03	.05	.10	.25	.60
	2	.01	.02	.06	.07	.23	.40
D	1	.17	.33	.40	.60	.80	.90
	2	.02	.03	.05	.07	.10	.25

Table 5
Possible escalation scheme

Dose	d_1	d_2	d_3	d_4	d_5
No. of nontox.	1	1	2	2	3

$g(b)$ is taken to be normal $\mathcal{N}(\mu_b, \sigma_b)$. Simulation parameters (μ_b, σ_b and the accrual ratio), as well as results are presented in Tables 6–10. The main conclusions are highlighted below. We describe as designs

1. The first design is simply the unstructured two-sample design described in O'Quigley et al. (1999). This is helpful for our comparisons.
2. Design 2 has $\mu = 1.5$ and $\sigma^2 = 0.15$. This variance is quite small, indicating a strong prior in favor of ordering.
3. Design 3 has $\mu = 1.5$ and $\sigma^2 = 0.35$. Although much weaker, this variance is still small enough to be informative. Prior knowledge concerning an ordering is moderate. Groups were unbalanced 35%/65%.
4. Design 4 has $\mu = 1.5$ and $\sigma^2 = 1.5$. This variance is large enough to be noninformative.
5. Design 5 has $\mu = 0.5$ and $\sigma^2 = 0.15$. Again an informative prior but, given the relatively small mean shift, this corresponds to the groups being not very different.
6. Design 6 has $\mu = 1.5$ and $\sigma^2 = 1.5$, as for design 4, but with an unbalanced group composition in the ratio 30%/70%.
7. Design 7 has $\mu = 0.93$ and $\sigma^2 = 0.2$, i.e., a small variance and a moderate shift between the groups. Prior knowledge is then quite strong but for a mild difference, in practice no more than one or two levels apart.
8. Design 8 has $\mu = -0.19$ and $\sigma^2 = 0.25$. Again this is quite an informative variance and, given that μ is in the wrong direction (for the case studied), we may anticipate problems. Here our prior information is plain wrong and is strong, typically the hardest situation for the data to overcome.
9. Design 9 has $\mu = 0.0$ and $\sigma^2 = 1.0$. No difference is assumed with a moderate size variance. The data would be expected to dominate the study.
10. Design 10 has $\mu = -0.19$ and $\sigma^2 = 1.0$. Another characterization in the wrong direction, for the case studied, this time with a more moderate variance. The data can rectify the erroneous prior more quickly.
11. Design 11 has $\mu = 0.93$ and $\sigma^2 = 0.35$ This is highly informative in the context of the case studied (groups as far apart as possible) and the chosen μ is not large enough. The data then have to work to convince us that group differences are greater than we anticipated.

In each group, we give the frequency that a level is allocated within the trial (% alloc.), the distribution of final recommendations (% rec.), and the frequency with which toxicities occur. The first couple of rows in Tables 6, 8, 9, and 10 (rows A, B, C and D) refer to performances obtained with the two-sample CRM and are useful for comparisons. The bolded parameter highlights the specific feature of the simulation.

Table 6
Final recommendations and in-trial allocations with the two-stage two-sample CRM

		Group 1							Group 2						
		d_1	d_2	d_3	d_4	d_5	d_6		d_1	d_2	d_3	d_4	d_5	d_6	
	$R_i(d_k)$	.07	.23	.31	.35	.45	.57	#T[a]	.01	.03	.05	.09	.20	.40	#T[a]
A	% of rec.	.24	.40	.22	.11	.03	.0	.18	.0	.01	.08	.27	.43	.21	.14
	% of alloc.	.16	.16	.11	.05	.02	.0		.04	.04	.10	.13	.14	.06	
		Ratio[b] = 0.5/0.5,					$b = 1.58$,		$\mu_b = 1.49$,				σ_b= **0.15**		
A.1	% of rec.	.29	.52	.17	.02	.0	.0	.20	.0	.0	.0	.12	.70	.18	.16
	% of alloc.	.19	.18	.10	.03	.0	.0		.03	.03	.07	.11	.20	.06	
		Ratio[b] = 0.35/0.65,					$b = 1.58$,		$\mu_b = 1.49$,				$\sigma_b = 0.35$		
A.2	% of rec.	.30	.47	.19	.03	.01	.0	.20	.0	.0	.01	.20	.64	.15	.15
	% of alloc.	.14	.11	.07	.02	.01	.0		.03	.04	.09	.16	.25	.08	
		Ratio[b] = 0.05/0.05,					$b = 1.58$,		$\mu_b = 1.49$,				σ_b= **1.5**		
A.3	% of rec.	.24	.42	.22	.10	.02	.0	.21	.0	.0	.05	.27	.48	.20	.15
	% of alloc.	.16	.16	.12	.05	.01	.0		.03	.04	.09	.13	.15	.06	

[a]#T: Average frequency of observed toxicities.
[b]Ratio: Accrual ratio.

Table 7
Recommendation and in-trial allocation for the four schemes

	d_1	d_2	d_3	d_4	d_5	d_6	
$R(d_k, 0)$	0.07	0.23	0.31	0.35	0.45	0.57	
$R(d_k, 1)$	0.01	0.03	0.05	0.09	0.20	0.40	
% recommendation in group 0	0.24	0.40	0.22	0.11	0.03	0.00	design 1
% recommendation in group 1	0.00	0.01	0.08	0.27	0.43	0.21	
% recommendation in group 0	0.29	0.52	0.17	0.02	0.00	0.00	design 2
% recommendation in group 1	0.00	0.00	0.00	0.12	0.70	0.18	
% recommendation in group 0	0.30	0.47	0.19	0.03	0.01	0.00	design 3
% recommendation in group 1	0.00	0.00	0.01	0.20	0.64	0.15	
% recommendation in group 0	0.24	0.42	0.22	0.10	0.02	0.00	design 4
% recommendation in group 1	0.00	0.00	0.05	0.27	0.48	0.20	
$R(d_k, 0)$	0.08	0.20	0.35	0.50	0.70	0.80	
$R(d_k, 1)$	0.01	0.05	0.18	0.40	0.55	0.70	
% recommendation in group 0	0.23	0.48	0.25	0.01	0.00	0.00	design 1
% recommendation in group 1	0.01	0.15	0.58	0.23	0.03	0.00	
% recommendation in group 0	0.19	0.65	0.16	0.00	0.00	0.00	design 5
% recommendation in group 1	0.00	0.05	0.75	0.20	0.00	0.00	
% recommendation in group 0	0.20	0.65	0.14	0.01	0.00	0.00	design 6
% recommendation in group 1	0.00	0.06	0.76	0.18	0.00	0.00	
% recommendation in group 0	0.56	0.42	0.02	0.00	0.00	0.00	design 7
% recommendation in group 1	0.00	0.01	0.48	0.49	0.02	0.00	
% recommendation in group 0	0.02	0.37	0.55	0.06	0.00	0.00	design 8
% recommendation in group 1	0.02	0.45	0.50	0.03	0.00	0.00	
$R(d_k, 0)$	0.02	0.03	0.05	0.10	0.25	0.57	
$R(d_k, 1)$	0.01	0.02	0.06	0.07	0.23	0.40	
% recommendation in group 0	0.00	0.01	0.09	0.35	0.47	0.07	design 1
% recommendation in group 1	0.00	0.01	0.07	0.29	0.45	0.18	
% recommendation in group 0	0.00	0.00	0.07	0.36	0.50	0.07	design 9
% recommendation in group 1	0.00	0.00	0.06	0.29	0.50	0.15	
% recommendation in group 0	0.00	0.01	0.07	0.35	0.49	0.08	design 10
% recommendation in group 1	0.00	0.01	0.06	0.30	0.48	0.15	
$R(d_k, 0)$	0.17	0.33	0.40	0.60	0.80	0.90	
$R(d_k, 1)$	0.02	0.03	0.05	0.07	0.10	0.25	
% recommendation in group 0	0.61	0.29	0.09	0.01	0.00	0.00	design 1
% recommendation in group 1	0.00	0.02	0.07	0.15	0.29	0.47	
% recommendation in group 0	0.33	0.42	0.22	0.03	0.00	0.00	design 11
% recommendation in group 1	0.00	0.00	0.06	0.29	0.52	0.13	

Table 8
In-trial allocations and final recommendations with the two-stage two-sample CRM

		Group 1							Group 2						
		d_1	d_2	d_3	d_4	d_5	d_6		d_1	d_2	d_3	d_4	d_5	d_6	
	$R_i(d_k)$	.08	.20	.35	.50	.70	.80	#T[a]	.01	.05	.18	.40	.55	.70	#T[a]
B	% of rec.	.23	.48	.25	.04	.0	.0	.23	.01	.15	.58	.23	.03	.0	.14
	% of alloc.	.16	.17	.13	.03	.01	.0		.05	.10	.22	.11	.02	.0	
		Ratio[b] $= 0.5/0.5$,				$b = 0.56$,			$\mu_b = 0.50$,			$\sigma_b = 0.15$			
B.1	% of rec.	.19	.65	.16	.0	.0	.0	.21	.0	.05	.75	.20	.0	.0	.21
	% of alloc.	.15	.22	.11	.02	.0	.0		.03	.07	.27	.11	.02	.0	
		Ratio[b] $= 0.30/0.70$,				$b = 0.56$,			$\mu_b = 0.50$,			$\sigma_b = 0.15$			
B.2	% of rec.	.20	.65	.14	.01	.0	.0	.22	.0	.06	.76	.18	.0	.0	.20
	% of alloc.	.09	.13	.07	.02	.0	.0		.04	.11	.39	.14	.02	.0	
		Ratio[b] $= 0.50/0.50$,				$b = 0.56$,			$\mu_\mathbf{b} = \mathbf{0.93}$,			$\sigma_b = 0.20$			
B.3	% of rec.	.57	.43	.03	.0	.0	.0	.17	.0	.01	.48	.49	.02	.0	.26
	% of alloc.	.25	.16	.07	.02	.0	.0		.03	.05	.21	.19	.03	.0	
		Ratio[b] $= 0.50/0.50$,				$b = 0.56$,			$\mu_\mathbf{b} = \mathbf{-0.19}$,			$\sigma_b = 0.25$			
B.4	% of rec.	.02	.37	.55	.06	.0	.0	.27	.02	.45	.50	.03	.0	.0	.14
	% of alloc.	.08	.16	.21	.05	.0	.0		.07	.18	.19	.05	.01	.0	

[a]#T: Average frequency of observed toxicities.
[b]Ratio: Accrual ratio.

4.2 *Simulations Results*

We consider some of the main cases of interest that may be anticipated to occur in practice. Our focus is on the impact of the added structure when we have a situation in which we have no knowledge of the right level and some knowledge of group differences. This latter knowledge comes under three broad headings; strong prior knowledge that is reasonably accurate, strong information that is erroneous and, finally, information that is very weak.

Cases A.1, B.1, *and* B.2: These three sets of simulations covered the case where the modeling assumptions on group differences are reasonably accurate: $\psi_2(d_{02}, a^*, \mu_b) \simeq R_2(d_{02})$. The variance σ_b is small, producing an informative prior. Results are encouraging and show the potential gains; for B.1 and B.2 the frequency of accurate final recommendations is increased by 33% $(1/2(+35\% + 29\%))$ compared to the likelihood design, and in A.1 gains are as high as +45% $(1/2(+30\% + 59\%))$! On average the right level in G. is correctly identified 75%, with a sample size between 16 and 22.

Effects of the escalation stage are particularly noticeable in cases A.1 and A.3. For situations in which, for at least one of the treatment arms, the MTD is distant from the starting level, we can achieve quite substantial gains in efficiency. This makes intuitive sense and it is reassuring that the method captures this. As G_2 is less at risk than the other group (there is a three-level difference), the escalation stage is speeded up and many more of the patients are directly included at potentially effective doses. Even so, the frequencies of observed toxicities are very close to those found with a more conservative initial design and remain below or equal to the targeted percentile. When sample sizes are unbalanced (B.2, A.2), the shared information enables significant gains concerning the group most likely to have low recruitment.

Table 9
In-trial allocations and final recommendations with the two-stage two-sample CRM

		Group 1							Group 2						
		d_1	d_2	d_3	d_4	d_5	d_6		d_1	d_2	d_3	d_4	d_5	d_6	
	$R_i(d_k)$	.02	.03	.05	.10	.25	.60	#T[a]	.01	.02	.06	.07	.23	.40	#T[a]
C	% of rec.	.0	.01	.09	.35	.47	.07	.15	.0	.01	.07	.29	.45	.18	.14
	% of alloc.	.04	.04	.10	.13	.14	.05		.04	.04	.10	.13	.14	.05	
		Ratio[b] $= 0.5/0.5$,				$b = 0.06$,			$\mu_b = 0$,			$\sigma_\mathbf{b} = \mathbf{1.0}$			
C.1	% of rec.	.0	.0	.07	.36	.50	.07	.15	.0	.0	.06	.29	.50	.15	.15
	% of alloc.	.04	.04	.10	.14		.03		.03	.03	.08	.14	.16	.05	
		Ratio[b] $= 0.5/0.5$,				$b = 0.06$,			$\mu_\mathbf{b} = \mathbf{-0.19}$,			$\sigma_b = 1.0$			
C.2	% of rec.	.0	.01	.07	.35	.49	.08	.16	.0	.01	.06	.30	.48	.15	.14
	% of alloc.	.04	.04	.10	.14	.15	.03		.03	.04	.09	.14	.16	.05	

[a]#T: Average frequency of observed toxicities.
[b]Ratio: Accrual ratio.

Table 10
In-trial allocations and final recommendations with the two-stage two-sample CRM

		Group 1							Group 2						
		d_1	d_2	d_3	d_4	d_5	d_6		d_1	d_2	d_3	d_4	d_5	d_6	
	$R_i(d_k)$	.17	.33	.40	.60	.80	.90	#T[a]	.02	.03	.05	.07	.10	.25	#T[a]
D	% of rec.	.61	.29	.09	.01	.0	.0	.15	.0	.02	.07	.15	.29	.47	.14
	% of alloc.	.28	.13	.07	.02	.0	.0		.04	.05	.10	.10	.12	.11	
		Ratio[b] $= 0.5/0.5$,					$b = 2.20$,			$\mu_b = \mathbf{0.93}$,			$\sigma_b = .35$		
D.1	% of rec.	.0	.0	.07	.36	.50	.07	.15	.0	.0	.06	.29	.50	.15	.15
	% of alloc.	.04	.04	.10	.14		.03	.15	.03	.03	.08	.14	.16	.05	

[a]#T: Average frequency of observed toxicities.
[b]Ratio: Accrual ratio.

Situations C.1 *and* A.3: The assumption concerning the difference is correct, but the variance of the prior estimate is large. Prior influence is weak and operating characteristics are similar to those found by O'Quigley et al. (1999), which themselves are very close to what can be obtained with two separate and independent trials. Prior influence being negligible, likelihood and Bayesian inferences will be similar, even for quite small samples. Improvements are nonetheless significant and arise not from the model or parameter precision, but from the sensible use of information in the initial escalation stage. In practical situations, the truth will lie somewhere between these two extremes: a noninformative and a very informative prior. Clinicians may have good intuition about the group difference in light of other studies, but they may not wish to assume high precision about any anticipated differences. The prior distribution will then be mildly informative, resulting in moderate gains. Case A.2 is an example of such moderate prior information with unbalanced sample sizes. A sizeable part of the gain will be obtained from the escalation stage rather than the added prior information.

Situations C.2, D.1, B.3 *and* B.4: As for any analysis based on the use of informative prior assumptions, it is important to investigate the consequences of errors on the added prior information. The procedure appears to be robust in this regard to model misspecification. This is probably due to a property of the CRM to quickly concentrate most inclusions at the targeted dose. In case C.2, the average of the prior is very distant from the true modeled value (actually, the ordering assumption is even inverted and G_2 is expected to be weaker than G_1). Even so, the consequences are not great since the variance is large enough for the observations to drive the inclusions and recommendations. In B.2 and D.1, errors concern the assumption about the distance $(d_{02} - d_{01})$ whereas in B.3 the assumption on the direction is also incorrect. Results are encouraging. If the ordering assumption turns out to be incorrect, then it will only impact the running of the trial and final recommendation in the case where the chosen prior is not only incorrect, but strongly informative. If the prior is strongly informative, and correct, then we will, of course, do very much better than we would have otherwise. This we expect. If the prior is weak, as in case C, then behavior is close to that obtained by O'Quigley et al. (1999) where no assumptions were made. The prior can then impact in a nonnegligible way the final decision recommendation, especially with small sample sizes. This happens of course with any design using Bayesian inference and is the inevitable trade-off we accept for the ability to include information beyond that provided only by the data. Nonetheless, in practice, were data to be in such strong contradiction with prior assumptions, investigators may wish to reevaluate such assumptions even during the course of the trial itself. Otherwise, performance can be seen to be improved over the traditional scheme when information on group difference is not in serious conflict with the data. Of course, the more accurate our knowledge, as always, the greater would be the gain. Other work, not presented here, indicates that model misspecification itself has a much weaker impact and we can consider the approach, in the same way as for single-sample CRM, to be relatively robust to model choice. The procedure outlined by O'Quigley and Reiner (1998) could be applied to each group separately in order to determine a rule for early stopping.

5. Illustration

The example concerned a phase I dose-finding study in children with acute leukemia. Rather more levels than are typical in many phase I studies were considered, specifically a total of ten. The spacing between the dose levels was based on pharmacological notions, mainly the use of modified Fibonacci dosing, leading to approximate doubling of the dose at the lower levels and increases of between 50% and 66% later on. Little was known about the amount of new agent, taken in conjunction with other agents, that may provoke toxic side effects. The MTD itself was taken to be that level in which 25% of patients would encounter dose limiting toxicity. The lowest levels were such that the investigators were close to being certain that the probability of toxicity in either group was effectively zero. The patients fell into two distinct prognostic groups: those patients having already received substantial prior treatment (heavy pretreatment group) and those having received much less prior treatment (light or moderate pretreatment group). It was expected that the heavy pretreatment group would not be able to take as strong a dose as the other group. In addition to this, it is known for this class of treatments that the therapeutic window is relatively narrow, which translates as a relatively steep dose toxicity curve. Thus two levels below the true, unknown, MTD we expect to see almost no toxicity at all, and two levels above we might see substantial toxicity. This information can be used in various ways, one being to consider that whatever the true unknown MTD should turn out to be for, say, group 1, then for

group 2 it is not realistically going to be very far from that level; possibly one or two levels different, but rather unlikely to be three or more removed. This knowledge can be incorporated into our design to our advantage.

This particular study had some other peculiar design features that, although not featuring strongly in our main discussion, ought be mentioned here. There were many fewer patients available in the heavy pretreated group so that, in seeking the right dose for the smaller group, the current recommended level for the larger group is providing a piece of valuable information we need exploit. As it turned out so often in practice, our example becomes slightly messy since a management decision was made not to use CRM, but to carry out a standard up-and-down design, in particular escalating patients in groups of three until the two patients out of six experienced a DLT at some level. This meant separate trials for the two groups. Only after it became clear that there was a considerable waste of resources was it decided to use CRM, the decision being made when, following standard up-and-down algorithms, it was felt the MTD had been established in group 1. These are not ideal conditions for using either one-sample or two-sample CRM, but this is what took place in practice. Our broader methodological development, the subject of this article, is not that well illustrated by the example. Nonetheless, it was the example that motivated the methodological investigation and we believe it should be reported as it turned out, even though decisions, beyond the competence of the statistician, make the actual study less relevant and less clear-cut than it may have been. This said, the example is, all the same, very interesting.

Data at the time the switch was made from up-and-down to CRM are given in Table 11. n_k denotes the number of patients included at dose level d_k.

The MTD in group 1 was close to being established and data from this group then constituted a considerable amount of information that we wished to exploit in searching for a dose for group 2 patients. Our proposal was to integrate this information through an informative prior distribution. Given the number of included patients and the fact that the ordering was known, it was convenient to work with a gamma law for the prior. The average and the variance were taken from the parameter estimate in $G_1, \hat{a}_{53}$. We fit the CRM model to the data gathered in the first group, and $\hat{a}_{53}$ gave us the average of the prior gamma law, our fit here being simply a method of moments estimate. The selected model was

$$\psi_1(d_k, a) = \alpha_k^a, \quad \psi_2(d_k, a, b) = \alpha_k^{(a+b)} \tag{6}$$

with the coding given in Table 12.

Notice that these values attenuate the influence of the first four levels, considered to be most likely quite removed from the MTD. Model ψ_1, fit on data from G_1, gave a parameter estimate $\hat{a}_{53} = 1.078$ with variance $\text{var}(\hat{a}_{53}) = 0.11$.

However, since the accrual rate in the second group was much slower than in the first, if we directly took $\text{var}(\hat{a}_{53})$ as the variance of the gamma law, the prior would have been over-weighted, i.e., the information from group 1 was having too strong an impact on our inference for group 2. We therefore decided to inflate the variance in order to dilute the prior's influence. Trial and error, guided by operating characteristics, resulted in our choosing a factor 3 with which to scale up the variance. Both distributions are displayed in Figure 3. We finally built up a prior distribution with an average of 1.078 and a variance of 0.33, which, for a gamma distribution, translates as the two parameters being 3.88 and 0.28, respectively.

Table 11
Data from phase I study in acute leukemia

Group 1			Group 2		
Patient no.	x_j	$\sum y_j/n_k$	Patient no.	x_j	$\sum y_j/n_k$
1,2,3	d_1	0/3	1,2,3	d_1	0/3
4,5,6,7	d_2	0/4	4,5,6	d_2	0/3
8,9,10	d_3	0/3	7,8,9	d_3	0/3
11,12,13	d_4	0/3	10,11,12	d_4	0/3
14,15,16	d_4	0/3	13,14,15	d_4	0/3
17,18,19,20,21	d_5	0/5	16,17,18	d_5	0/3
22,23,24	d_6	1/3	-	-	-
25,26,27	d_6	0/3			
28,29,30	d_7	1/3			
31,32,33	d_7	0/3			
34,35	d_8	1/2			

Table 12
Model coding

k	1	2	3	4	5	6	7	8	9	10
α_k	.005	.007	.01	.02	.03	.07	.23	.35	.45	.55

6. Discussion

In some sense we are always dealing with patient heterogeneity. Treating a group of patients as though they were a single group, comparable in all ways, as far as any prognostic factors, known or unknown, are concerned is really an abstract statistical construction. The construction is useful in enabling us to identify, with relatively few patients, often no

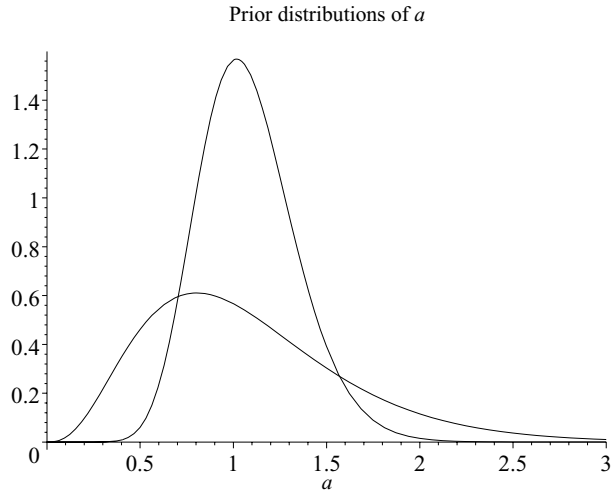

Figure 3. Gamma distributions of variance $\text{var}(\hat{a}_{53}) = 0.11(\Gamma(17.15, 0.06))$ and $\text{var}(\hat{a}_{53}) = 0.33(\Gamma(3.88, 0.28))$.

more than 20, an MTD for the group in question. It may be helpful to view such an MTD as an average for the group, a level at which some given percentage of patients will experience dose-limiting toxicities. If the knowledge of prognostic factors can enable us to more precisely identify the MTD for different prognostic groups, then the more involved modeling described in this work can be rewarding. The simulations give some idea about the potential gains and under which circumstances. Nonetheless some skill and preliminary work is required on the part of the statistician if such gains are to be realized in practice. This is because the extra variability and added difficulty of fitting two or more parameter models may offset any hoped-for increases in precision, in particular when relatively few patients are available for study. Before using the methods described here we advise simulation studies similar to those presented above, customized to the application in mind, i.e., using the sample sizes and relative group sizes that we expect to obtain. Thus we do not recommend any generic design, capable of working well in very broad circumstances. We suggest instead that, for two or more groups, careful thought be given to the nature of the groups and whether or not it is realistic to quantify prior information concerning ordering of the MTDs, how far apart the MTDs may be, or other information that is secondary to the main question of identifying the average MTD. If such an endeavor appears to be realistic, then we can anticipate carrying out our study using fewer patients. Savings are unlikely to be great but, in this context where patients can be rare and recruitment lengthy, even using two or three less patients than we would have otherwise needed may make the difference between successfully proceeding to a phase II study or not.

Acknowledgements

The authors would like to thank the editors and reviewers for a very extensive and detailed report on an earlier version of this article.

Résumé

Nous étudions une extension de la Continual Reassessment Method pour deux groupes dans le cadre d'une étude de recherche de dose lorsque les deux groupes peuvent être ordonnés. Dans une telle situation, nous avons peu ou pas d'information sur la localisation de la Dose Maximale Tolérée (DMT) dans chacun des deux groupes, mais nous pouvons savoir beaucoup plus précisément dans quel groupe, celle-ci sera la plus élevée, si différence il y a. Ce problème nous fut soulevé lors d'un essai sur une nouvelle thérapie chez les enfants atteints de leucémie aiguë. Le contexte de cette étude est discuté. Il y avait deux groupes de patients, distingués par le degré des traitements reçus avant d'entrer dans l'étude. C'est pourquoi, il était attendu que la DMT dans le second groupe, le plus légèrement pré-traité, soit atteinte à un niveau plus élevé. Des méthodes basées sur la vraisemblance, ou de façon équivalente sur une distribution a priori non informative, peuvent être utilisées pour modéliser la localisation de la DMT, tandis qu'une modélisation bayésienne plus informative peut être utilisée pour le second aspect du problème; par exemple, pour la direction de la différence entre les niveaux de MTD des deux groupes ou, même éventuellement pour la différence elle-même.

References

Babb, J., Rogatko, A., and Zacks, S. (1998). Cancer phase 1 clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine* **17,** 1103–1120.

Møller, S. (1995). An extension of the CRM using a preliminary up and down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Statistics in Medicine* **14,** 991–22.

O'Quigley, J. (2001). Continual reassessment method in cancer clinical trials. In *Handbook of Statistics in Clinical Oncology,* J. Crowley (ed), 35–72. New York: Marcel Dekker.

O'Quigley, J. and Shen, Z. L. (1996). Continual reassessment method: A likelihood approach. *Biometrics* **52,** 163–174.

O'Quigley, J. and Reiner, E. (1998). A stopping rule for the continual reassessment method. *Biometrika* **85**(3), 741–48.

O'Quigley, J., Shen, Z. L., and Gamst, A. (1999). Two sample continual reassessment method. *Journal of Biopharmaceutical Statistics* **9**(1), 17–44.

Rosenblum, M. G., Verschraegen, C. F., Murray, J. L., Kudelka, A. P., Gano, J., Cheung, L., and Kavanagh, J. J. (1999). Phase I study of 90Y-labeled B72.3 intraperitoneal administration in patients with ovarian cancer: Effect of dose and EDTA coadministration on pharmacokinetics. *Clinical Cancer Response* **5**(5), 953–61.

Shen, L. Z. and O'Quigley, J. (1996). Consistency of continual reassessment method under model mispecification. *Biometrika* **83,** 395–406.

Whitehead, J. (1997). Bayesian decision procedures with application to dose-finding studies. *International Journal of Pharmaceutical Medicine* **11,** 201–208.

Received January 2001. *Revised November* 2002.
Accepted November 2002.