Journal of Statistical Software

August 2013, Volume 54, Issue 13. *http://www.jstatsoft.org/*

bcrm: Bayesian Continual Reassessment Method Designs for Phase I Dose-Finding Trials

Michael Sweeting
MRC Biostatistics Unit

Adrian Mander
MRC Biostatistics Unit

Tony Sabin
Amgen

Abstract

This paper presents the R package **bcrm** for conducting and assessing Bayesian continual reassessment method (CRM) designs in Phase I dose-escalation trials. CRM designs are a class of adaptive design that select the dose to be given to the next recruited patient based on accumulating toxicity data from patients already recruited into the trial, often using Bayesian methodology. Despite the original CRM design being proposed in 1990, the methodology is still not widely implemented within oncology Phase I trials. The aim of this paper is to demonstrate, through example of the **bcrm** package, how a variety of possible designs can be easily implemented within the R statistical software, and how properties of the designs can be communicated to trial investigators using simple textual and graphical output obtained from the package. This in turn should facilitate an iterative process to allow a design to be chosen that is suitable to the needs of the investigator. Our **bcrm** package is the first to offer a large comprehensive choice of CRM designs, priors and escalation procedures, which can be easily compared and contrasted within the package through the assessment of operating characteristics.

Keywords: Bayesian adaptive designs, continual reassessment method, phase I, dose finding, dose escalation, R.

1. Background

The continual reassessment method (CRM, O'Quigley, Pepe, and Fisher 1990) is an adaptive design used in Phase I clinical trials to find the maximum tolerated dose (MTD). The aim of oncology Phase I trials is to escalate the dose of a new therapy by sequentially entering patients into the trial and selecting the next dose based on the accumulating toxicity data. In oncology, cytotoxic drugs are likely to have severe toxicity at high doses, yet at low doses little efficacy is expected from the drug. Hence, a dose that achieves a tolerable (non-zero) toxicity level is sought that has efficacy potential. Commonly toxicity is the primary outcome on which escalation decisions are based and safety events are classified into dose-limiting toxicities (DLTs) and non-DLTs. The MTD is then defined as the dose that has risk of DLT equal to a chosen target toxicity level (TTL), usually between 20–33% in oncology trials (Tourneau, Lee, and Siu 2009).

Adaptive designs play an important role in dose selection for Phase I trials, with the simple algorithmic '3+3' approach (Storer 1989) still widely used. The design has been widely criticized for being inefficient and inflexible (O'Quigley *et al.* 1990; Heyd and Carlin 1999; Tourneau *et al.* 2009; O'Quigley and Zohar 2006) due partly to escalation decisions being based on outcomes from only the most recent subset of recruited patients. This has led to a number of model-based designs being developed, the most prominent of these being the CRM, where escalation decisions are based on point estimates of the posterior distributions or likelihoods formed from all the accumulated data. A parametric form is often assumed for the dose-toxicity relationship, based either on a one-parameter 'working model' (O'Quigley *et al.* 1990) or a two-parameter logistic function (Whitehead 1997). The one-parameter approach performs well if focusing on local estimation of the dose-response curve, such as treating patients at the estimated MTD, despite probable lack-of-fit in other areas of the dose-toxicity curve (O'Quigley *et al.* 1990). This has been shown to be true in simulations and theoretical work even under misspecification of the one-parameter model (Shen and O'Quigley 1996). A two-parameter model, meanwhile, has been shown to be unstable if dosing tends to be concentrated at only one level (Paoletti and Kramar 2009), since the model is then over-specified for the data being collected. Despite this, the two-parameter model is commonly used due to its ease of interpretation and the ability to make appropriate inferences about the probabilities of DLTs at each dose level (Neuenschwander, Branson, and Gsponer 2008). The latter statement is compatible with a fully-Bayesian philosophy, where historical data or elicitation are used to describe prior beliefs about the shape and uncertainty in the dose toxicity curve and if the model is not misspecified. Escalation decisions can then be based on the whole posterior distribution of toxicity at each dose (Neuenschwander *et al.* 2008), although it should be noted that knowledge about toxicity at untested doses will be dominated by the prior. Given the divergent views in the literature around this topic, it is important for researchers to have easy to use and readily accessible software, that will enable them to make an informed decision around the choice of CRM model for their particular situation, and the **bcrm** software can be used to this effect.

Despite the demonstrated superiority of model-based over traditional algorithmic designs their uptake has been slow, with many oncology trials still implementing '3+3' designs (Tourneau *et al.* 2009). There is a danger that model-based designs may be perceived to be something of a 'black box' in which the process underlying the recommendations is not transparent. To improve the understanding of the methodology requires closer co-operation between the investigators of the trial and the statistician at the design stage.

Another issue is the conceived lack of user-friendly software. Nevertheless, software to implement CRM designs does exist in various guises. The R package **dfcrm** (Cheung 2011) has been developed to implement one-parameter CRM working models and time-to-event CRM (TITE-CRM) designs (Cheung and Chappell 2000). Simulation studies can be performed to assess operating characteristics, and various functions are available to evaluate model sensitivity and to calibrate the design (Cheung 2005; Cheung and Chappell 2002). Other stand-alone software has been developed to implement CRM models, notably **CRMSimulator** (Norris and Cook 2006) and **CRM** (Venier 1999).

In this paper we present a new R (R Core Team 2013) package **bcrm**, which provides a more comprehensive implementation of various Bayesian CRM designs. The package is available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=bcrm, offers extra flexibility in the choice of design aspects, and is intended to allow easy assessment of the designs through summary and graphical outputs. This is often lacking in other software packages, and we feel that an easy to use and interpret package is needed in order to promote the use of CRM designs more widely. The **bcrm** package has other advantages in that it implements both one and two-parameter models, allowing the user to contrast operating characteristics of the different models, and can make use of MCMC computation via either **WinBUGS** (Spiegelhalter, Thomas, Best, and Lunn 2003) or **OpenBUGS** (Lunn, Spiegelhalter, Thomas, and Best 2009). The main features of **bcrm** are as follows:

1. The freedom to specify the functional form of the model, with a variety of one or two-parameter models available.

2. A wide choice of prior distributions and options to calculate standardized doses.

3. Decision rules for escalation can be based on risks of toxicity calculated from plug-in estimates of the mean of the parameter(s), the posterior mean risk of DLT, quantiles of the MTD distribution as used in an escalation with overdose control (EWOC) design (Babb, Rogatko, and Zacks 1998), or a loss function that weights intervals of toxicity risk. A TITE-CRM design is not currently implemented but maybe available in a future release.

4. A variety of possible stopping rules including a minimum and maximum sample size, stopping once a maximum number are treated at the final MTD estimate, and stopping once a required level of precision is obtained.

5. The flexibility to change other design inputs such as the cohort size, the TTL, as well as algorithmic constraints placed on escalation.

6. Easy to assess operating characteristics using built-in simulation code, or an interactive approach is possible with data entered and decisions made after each cohort is recruited.

7. Operating characteristics from the CRM design can be compared side-by-side with the standard 3+3 design.

8. Informative graphical and textual output from both a single CRM trial and those conducted under a simulation study.

9. Estimation using either ‘exact’ methods (adaptive quadrature via the R `integrate` command) or Markov chain Monte Carlo (MCMC) via either **WinBUGS** or **OpenBUGS** using the R packages **R2WinBUGS** (Sturtz, Ligges, and Gelman 2005) and **BRugs** (Thomas, O’Hara, Ligges, and Sturtz 2006), respectively.

The aim of this paper is to therefore demonstrate the options available within the software and to guide users through an example trial with particular emphasis on approaches to critically evaluate a design and its operating characteristics.

2. Methods

Assume that k doses have been chosen for possible experimentation in the trial with the MTD assumed to be within the range of chosen doses. The lowest dose is usually chosen based on animal toxicology studies and is expected to be very safe in humans. The choice of the k doses may be constrained by practical considerations, such as tablet manufacturing. Although this choice is an important consideration for a Phase I trial, and guidelines are available for this purpose (U.S. Department of Health and Human Services 2010), it is not the focus of this paper. We shall, however, assume that the doses $d_1, \ldots, d_k$ are pre-chosen by the investigator and correspond to some a-priori guesses of the risk of DLT $\theta_1, \ldots, \theta_k$ (probabilities) at each dose. The aim is to identify the MTD x (not necessarily one of the doses selected) whose risk of DLT is equal to the TTL θ, where $\theta_1 \leq \theta \leq \theta_k$.

2.1. One-parameter CRM designs

The original implementation of the CRM assumes a one-parameter monotonically increasing functional form between a standardized or re-scaled version of the doses, $d_1^*, \ldots, d_k^*$, and the risk of a DLT (see Section 2.1.1 for the calculation of the standardized doses). The underlying assumption is that the risk of DLT increases with dose. Let Y_i denote the binary toxicity outcome observed in the ith patient recruited to the trial, where $Y_i = 1$ denotes a DLT, and let the standardized dose chosen for the ith patient be denoted $d^*(i)$. O'Quigley *et al.* (1990) proposed a one-parameter hyperbolic-tangent dose-response model as follows:

$$P(Y_i = 1) = \pi(d^*(i); \alpha) = \left[\left(\tanh\left(d^*(i)\right) + 1\right)/2\right]^{\alpha}, \tag{1}$$

where $\alpha \in \mathbb{R}^+$. This model, despite only having one parameter, fulfils the requirement that there exists an $\alpha_0 \in \mathbb{R}^+$ for which $\pi(x^*; \alpha_0) = \theta$, where x^* is the standardized version of the MTD x. That is we are able to use the model to target doses that have the required toxicity level. Two further 1-parameter models are a logistic model with fixed intercept, c, where c is often taken to be three (O'Quigley and Chevret 1991)

$$\pi(d^*(i); \alpha) = \frac{\exp\left(c + \alpha d^*(i)\right)}{1 + \exp\left(c + \alpha d^*(i)\right)} \tag{2}$$

and a power model of the form

$$\pi(d^*(i); \alpha) = d^*(i)^{\alpha}. \tag{3}$$

The power model is seen to be equivalent to the hyperbolic-tangent model but with a different transformation of the doses (O'Quigley and Shen 1996; Paoletti and Kramar 2009).

Choosing the standardized doses

The doses are standardized to calibrate the chosen functional form of the dose-response curve to the investigator's prior estimates of DLT. This can be achieved by setting $d_i^* = \pi^{-1}(\theta_i; \hat{\alpha})$, where π^{-1} is the inverse of the risk model, and $\hat{\alpha}$ is a prior estimate (mean or median) of the parameter. For example, if the hyperbolic-tangent functional form is used with prior mean for α equal to 1 then the standardized doses can be calculated as

$$d_i^* = \tanh^{-1}(2\theta_i - 1).$$

Meanwhile, if the power model is used with prior mean equal to 1 then the standardized doses are

$$d_i^* = \theta_i,$$

i.e., the standardized doses are equal to the prior estimates of DLT. The standardized doses are treated as fixed quantities over the study duration. However, since they are calculated based on the a-priori estimates of toxicity, they should be thought of as part of the design, which themselves require evaluation.

Bayesian updating

The adaptive nature of the CRM arizes from choosing the dose for the next patient based on the posterior distribution from the currently recruited patients. A Bayesian approach is implemented by placing a prior distribution, $f(\alpha)$, on the model parameter. The posterior distribution for α after n outcomes have been observed is then

$$f(\alpha|y_1,\ldots,y_n) = \frac{f(\alpha)L(\alpha;y_1,\ldots,y_n)}{\int_0^\infty f(u)L(u;y_1,\ldots,y_n)\,du}, \tag{4}$$

where the likelihood is $L(\alpha;y_1,\ldots,y_n) = \prod_{i=1}^n \pi\left(d^*(i);\alpha\right)^{y_i}\left(1-\pi\left(d^*(i);\alpha\right)\right)^{1-y_i}$.
The prior for α could be one of a number of possible, positive valued, distributions such as Gamma, Uniform or Lognormal. Often, a default prior is assumed, such as a Gamma$(1,1)$ (O'Quigley *et al.* 1990). For what follows we shall refer to the prior mean of α as m_0.

Escalation based on point estimates

Decisions regarding the dose to give to the next patient are often based on point estimates of the risk of DLT. This is especially true for the one-parameter CRM designs where the philosophy is to dose at the current 'best' estimate of the dose closest to the MTD (i.e., a "patient gain" philosophy, where the next recruited patient is given the optimal treatment for them, see Whitehead (1997) for example). Two possible point estimates to base dose-escalation decisions on are:

1. Plug-in mean

$$\mu_1(d_i^*;y_1,\ldots,y_n) = \pi\left(d_i^*;\int_0^\infty \alpha f(\alpha|y_1,\ldots,y_n)\,d\alpha\right).$$

2. Posterior mean

$$\mu_2(d_i^*;y_1,\ldots,y_n) = \int_0^\infty \pi(d_i^*;\alpha)\; f(\alpha|y_1,\ldots,y_n)\,d\alpha.$$

One attribute of using the plug-in mean as an estimate of the risk of DLT at each dose is that the standardized doses for one-parameter models are often calculated based on a plug-in prior mean. Hence at the beginning of the trial the point estimates used for escalation are consistent with the a-priori estimates of DLT, i.e.,

$$\mu_1(d_i^*) = \pi\left(d_i^*;\int_0^\infty \alpha f(\alpha)\,d\alpha\right) = \pi\left(\pi^{-1}(\theta_i;m_0);m_0\right) = \theta_i.$$

Whichever point estimate is chosen, the dose for the $(n+1)$-st patient is the one whose point estimate is closest to the TTL, that is

$$d^*(n+1) = \underset{\xi \in \{1,\ldots,k\}}{\arg\min} \left| \; \mu_P(d^*_\xi; y_1, \ldots, y_n) - \theta \; \right|,$$

for $P = 1$ or 2.

2.2. A two-parameter CRM design

A two-parameter model (Whitehead 1997; Neuenschwander *et al.* 2008) is based on the following logistic regression equation

$$\text{logit}(\pi(d^*_i; \beta_1, \beta_2)) = \log(\beta_1) + \beta_2 d^*_i, \tag{5}$$

where $d^*_i = \log(d_i/d_R)$ is a dose standardized to a reference dose d_R, so that $\log(\beta_1)$ becomes the log-odds of toxicity when $d_i = d_R$. The parameters β_1 and β_2 are positive-valued, thus ensuring a monotonically increasing dose-toxicity relationship. Neuenschwander *et al.* (2008) place a bivariate lognormal prior on the two parameters and propose eliciting information on the risk of DLT at two or more doses to obtain prior means and variances. The posterior distribution is then

$$f(\beta_1, \beta_2 | y_1, \ldots, y_n) = \frac{f(\beta_1, \beta_2) L(\beta_1, \beta_2; y_1, \ldots, y_n)}{\int_0^\infty \int_0^\infty f(u, v) L(u, v; y_1, \ldots, y_n) \, du \, dv}, \tag{6}$$

where $f(\beta_1, \beta_2)$ is the joint prior distribution and $L(\beta_1, \beta_2; y_1, \ldots, y_n)$ is the likelihood function.

Escalation with overdose control (EWOC)

Escalation can be based on point estimates of the posterior distribution, as in the one-parameter case. An alternative approach proposed by Babb *et al.* (1998) targets a quantile of the actual MTD distribution. First, the (standardized) MTD is defined based on a transformation of the two-parameter logistic model as follows:

$$MTD^* = \frac{\text{logit}^{-1}(\theta) - \log(\beta_1)}{\beta_2}$$

To avoid overdosing, the next dose can then be chosen based on the quantile of the cumulative posterior distribution function (CDF) for the MTD, $F(\gamma; y_1, \ldots, y_n) = P(MTD^* \leq \gamma; y_1, \ldots, y_n)$. For example, to target the $100q$-percentile, the dose selected for the $(n+1)$-st patient is such that the predicted probability that it exceeds the MTD is equal to q

$$F(d^*(n+1); y_1, \ldots, y_n) = q.$$

The value q is chosen to be less than 0.5, thus ensuring that the next dose is below the posterior median estimate of the MTD. In practice, the $100q$-percentile of the MTD distribution is unlikely to be amongst the set of pre-specified doses and hence the dose for the $(n+1)$-st patient is chosen as the dose closest to the specified quantile of the CDF:

$$d^*(n+1) = \underset{\xi \in \{1,\ldots,k\}}{\arg\min} \left| \; F(d^*_\xi; y_1, \ldots, y_n) - q \; \right|.$$

Escalation based on toxicity intervals

Neuenschwander *et al.* (2008) take a fully-Bayesian approach and allow dose-escalation decisions to be based on the whole posterior distribution. First, desirable and less-desirable regions of toxicity are defined, such as 'Underdosing', 'Target dosing', 'Overdosing' and 'Severe overdosing'. Such regions are defined by the risk of DLT and based on these intervals a loss-function is set-up. Suppose that M regions of toxicity have been defined, $[p_0 = 0, p_1], (p_1, p_2], \ldots, (p_{M-1}, p_M = 1]$, and an associated loss function for dose d_i^* defined as

$$Loss(\beta_1, \beta_2; d_i^*) = \ell_m \text{ if } \pi(d_i^*; \beta_1, \beta_2) \in (p_{m-1}, p_m], \tag{7}$$

which is a function of the parameters in the two-parameter model. The losses $\ell_1, \ldots, \ell_M$ are pre-specified and correspond to each region of toxicity. The Bayes risk, or posterior expected loss, for dose d_i^* is then defined as

$$B(d_i^*) = \int_0^\infty \int_0^\infty Loss(\beta_1, \beta_2; d_i^*)\ f(\beta_1, \beta_2 | y_1, \ldots, y_n)\ d\beta_1 d\beta_2 \tag{8}$$

The dose which minimizes the Bayes risk is selected as the next dose. Under a Bayesian philosophy with well-defined priors this approach allows the current uncertainty in the risk of DLT at each dose level to be accounted for in the dose-escalation procedure. For example, two doses may have very similar point estimates (posterior mean or median) of the risk of DLT but the posterior probability that the risk exceeds an unacceptable toxicity may be quite different. This can be accounted for using the toxicity interval approach.

2.3. Stopping rules

Patients continue to be recruited to the trial until either a fixed sample size is achieved or one of a number of possible stopping rules is triggered. Dynamic stopping rules based on the information collected thus far are appealing and may allow a trial to stop earlier than anticipated. O'Quigley and Reiner (1998) propose a stopping rule based on the probability that the patients that remain to be recruited, together with the final recommended MTD, are all assigned the same dose level. If, at any time, this predicted probability is high, then the trial is stopped. Whilst this is an appealing stopping rule, it can be difficult to implement in practice especially if the number of patients yet to be recruited is large, as all combinations of future responses need to be calculated. A simpler 'settling' rule based on whether a dose has been allocated k times (Korn, Midthune, Chen, Rubinstein, Christian, and Simon 1994), for a pre-specified k, compares favorably with the rule based on combinatorial trees (O'Quigley 2002). Alternatively, a stopping rule based on the precision of the final MTD estimate is another advocated approach (Heyd and Carlin 1999), and we demonstrate this approach in the example application in the next section.

3. The bcrm package: an example application

All of the approaches described in Section 2 can be implemented using the function `bcrm` in the **bcrm** package. To install and load the package within R type

```
R> install.packages("bcrm")
R> library("bcrm")
```

Dose	5mg	10mg	15mg	25mg	40mg	50mg	60mg
Prior guess of risk	0.05	0.10	0.20	0.30	0.35	0.40	0.45

Table 1: Prior guesses of risk of DLT given by investigator before trial commences.

The main features of this package will now be highlighted through the design of a fictitious Phase I oncology trial.

Suppose a drug is to be tested in an oncology first-in-man Phase I trial where the objective is to find a maximum tolerated dose corresponding to a risk of DLT occurring in 30% of patients. Seven dose levels have been identified for possible experimentation (5mg, 10mg, 15mg, 25mg, 40mg, 50mg, 60mg), and due to safety concerns the first cohort of patients will receive the lowest dose, identified as a likely safe dose from animal toxicology studies. The investigator wishes to recruit patients in cohorts of size three, and not to allow dose-skipping during escalation. The investigator's best guess of the risk of DLT at the lowest dose is 5% whilst their best guess of the MTD is at dose level 4 (i.e., a 30% risk of DLT at this dose level). After further discussions the initial guesses of risk of DLT are established for each dose level as shown in Table 1.

Using this information, the statistician is to present a report of possible CRM designs and their operating characteristics for a sample size of 42 patients. The investigator requires the design to be 'well-calibrated' in that it should escalate rapidly to the MTD whilst not subjecting too many patients to doses that are too toxic. Under scenarios where the drug is more toxic than expected, the design should be flexible enough to react quickly to accumulating toxicities. The investigator is also worried about how their prior guesses regarding the risk of DLT may affect the escalation procedure.

3.1. Interactively conducting a CRM trial

Suppose our choice of CRM model is a one-parameter power model, where a Gamma(1,1) prior is used, i.e., $f(\alpha) = \exp(-\alpha)$. The one-parameter power model is specified in the function `bcrm` using the argument `ff = "power"`. By default, one-parameter models are fit using numerical methods (adaptive quadrature). Meanwhile the prior is specified using the `prior.alpha = list(a, b, c)` argument, where element `a` specifies the distributional form, and elements `b` and `c` are the parameters of the chosen distribution. The choice of distributions are (i) `a = 1`: Gamma(b, c), where $b =$ shape, $c =$ scale, (ii) `a = 2`: Uniform(b, c), where $b =$ min, $c =$ max, (iii) `a = 3`: Lognormal(b, c), where b = mean on the log scale, c = standard deviation on the log scale, (iv) `a = 4`: Bivariate Lognormal(b, c), where b =mean vector on the log scale, c =Variance-covariance matrix on the log scale. Hence, a Gamma(1,1) prior is specified by `prior.alpha = list(1, 1, 1)`.

Standardized doses will be calculated using a plug-in mean estimate and for consistency escalation decisions will also be based on the plug-in mean. This is the default in **bcrm**, although they can be changed using the `sdose.calculate` and `pointest` arguments if required. Hence escalations are based on a "patient gain" philosophy (Whitehead 1997). A constrained form of the CRM, whereby no dose-skipping is allowed, is also the default, and hence the starting dose *level* must be specified using the argument `start`. An unconstrained design can be conducted by setting `constrain = FALSE`. The `bcrm` function allows a number of stopping rules to be specified using the `stop` argument, which is a list with one or more of the following: (i) `nmax`:

The maximum sample size of the trial, (ii) `nmtd`: The maximum number to be treated at the final MTD estimate (Korn *et al.* 1994), i.e., if the next recommended dose has already been administered to `nmtd` patients then the trial will stop, (iii) `precision`: A vector of the lower and upper percentage points that the MTD 95% credible intervals for the risk of toxicity should lie within (Heyd and Carlin 1999), (iv) `nmin`: The minimum sample size of the trial to be used in conjunction with `nmtd` or `precision`.

We start by considering a fixed sample size of 42, but later consider other options based on the precision of the toxicity estimate at the final recommended dose. To interactively conduct such a trial we run the following code in R, by first specifying the prior guesses of the risk of DLT at each dose:

```
R> p.tox0 <- c(0.05, 0.10, 0.20, 0.30, 0.35, 0.40, 0.45)
R> dose.label <- c(5, 10, 15, 25, 40, 50, 60)
R> bcrm(stop = list(nmax = 42), p.tox0 = p.tox0, dose = dose.label,
+    ff = "power", prior.alpha = c(1, 1, 1), target.tox = 0.30, start = 1)
```

The recommended dose for the first cohort of patients is returned by the program, which in this case is equal to our starting dose (dose level 1, 5mg). The user is allowed to choose whether to accept this recommendation or override the recommendation with another dose. The user then records the DLTs observed in the first cohort. Suppose that we accept the recommendation that the first cohort be dosed at 5mg, and we observe no toxicities in any of the first three patients. After inputting this information, a recommendation for the second cohort is given. The output from the program thus far is as follows:

```
 RECOMMENDED DOSE FOR PATIENTS  1  to  3 IS: 5

 ENTER DOSE FOR PATIENTS  1  to  3
 POSSIBLE CHOICES ARE  5 10 15 25 40 50 60
 (`RETURN' TO ACCEPT RECOMMENDATION, 0 TO EXIT AND RETURN CURRENT RESULTS)

ENTER TOXICITY DATA FOR PATIENT 1 (1=TOX, 0=NO TOX): 0
ENTER TOXICITY DATA FOR PATIENT 2 (1=TOX, 0=NO TOX): 0
ENTER TOXICITY DATA FOR PATIENT 3 (1=TOX, 0=NO TOX): 0


                ENTERED VALUES:
 DOSE ... 5
 TOXICITIES .... 0 0 0
 PRESS `RETURN' IF OK, OR ANY OTHER KEY TO ENTER NEW VALUES


 RECOMMENDED DOSE FOR PATIENTS  4  to  6 IS: 10

 ENTER DOSE FOR PATIENTS  4  to  6
 POSSIBLE CHOICES ARE  5 10 15 25 40 50 60
 (`RETURN' TO ACCEPT RECOMMENDATION, 0 TO EXIT AND RETURN CURRENT RESULTS)
```

The interactive nature of the program continues until all 42 patients have been recruited,

Dose	5mg	10mg	15mg	25mg	40mg	50mg	60mg
No. DLTs	1	0	2	4	5	0	0
No. non-DLTs	2	3	13	5	7	0	0

Table 2: Number of DLTs and non-DLTs observed during a trial with sample size 42.

where the program then terminates and returns an object of class `bcrm`. Printing such an object gives output of posterior estimates (mean, median, standard deviation, and quantiles) of the risk of DLT at each dose together with the recommended MTD.

Data accumulated from a trial can also be entered using the `data` argument, where `data` is a data frame specifying patient numbers, dose levels and toxicities of previously recruited patients. For example, suppose the data shown in Table 2 in summary form were collected during the trial.

The data can be used directly as input to obtain the estimated MTD at the end of the trial or as starting data for a larger trial. The code and the full sequential data is therefore as follows:

```
R> data <- data.frame(patient = 1:42, dose = rep(c(1, 2, 3, 4, 5, 5, 5, 5,
+    4, 4, 3, 3, 3, 3), each = 3), tox = c(1, rep(0, 13), 1, 1, rep(0, 4),
+    rep(1, 3), rep(0, 3), rep(1, 5), rep(0, 7), 1, rep(0, 3)))
R> trial.output <- bcrm(stop = list(nmax = 42), data = data,
+    p.tox0 = p.tox0, dose = dose.label, ff = "power",
+    prior.alpha = c(1, 1, 1), target.tox = 0.30)

 Stopping: Reached maximum sample size

R> print(trial.output)

 Estimation method:  exact

 Model:  1-parameter power

 Prior:  Gamma( Shape:1, Scale:1)

 Standardised doses (skeleton):
   5   10   15   25   40   50   60
0.05 0.10 0.20 0.30 0.35 0.40 0.45

 Modified (constrained) CRM used, starting dose:  5

 Plug-in estimate of probability of toxicity used to select next dose

 Toxicities observed:
            Doses
             5 10 15 25 40 50 60
  n          3  3 15  9 12  0  0
```

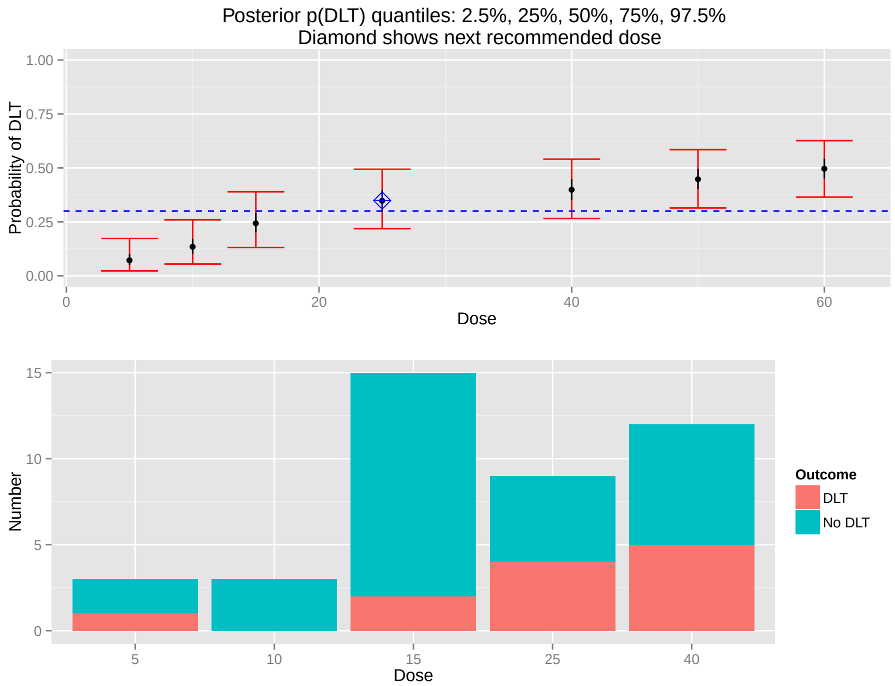


Figure 1: Plot of estimated risk of DLT and outcomes observed at the end of the trial (after 42 patients have been recruited).

```
  Toxicities 1  0  2  4  5  0  0

 Posterior estimates of toxicity:
        Doses
              5    10     15     25     40     50    60
  Mean   0.0793 0.140 0.2490 0.3510 0.4000 0.4490 0.497
  SD     0.0391 0.053 0.0665 0.0707 0.0705 0.0693 0.067
  Median 0.0727 0.133 0.2440 0.3490 0.3990 0.4480 0.497
         Doses
Quantiles      5     10    15    25    40    50    60
    2.5%  0.0227 0.0545 0.131 0.219 0.265 0.314 0.365
    25%   0.0505 0.1010 0.201 0.301 0.351 0.401 0.451
    50%   0.0727 0.1330 0.244 0.349 0.399 0.448 0.497
    75%   0.1010 0.1720 0.292 0.398 0.448 0.496 0.543
    97.5% 0.1730 0.2600 0.390 0.494 0.541 0.585 0.626

 Plug-in estimates of toxicity:
          5    10    15    25    40    50    60
[1,] 0.0699 0.129 0.239 0.343 0.394 0.443 0.492

 Next recommended dose:  25
```

A plot of the posterior estimates can be obtained after each cohort has been recruited by

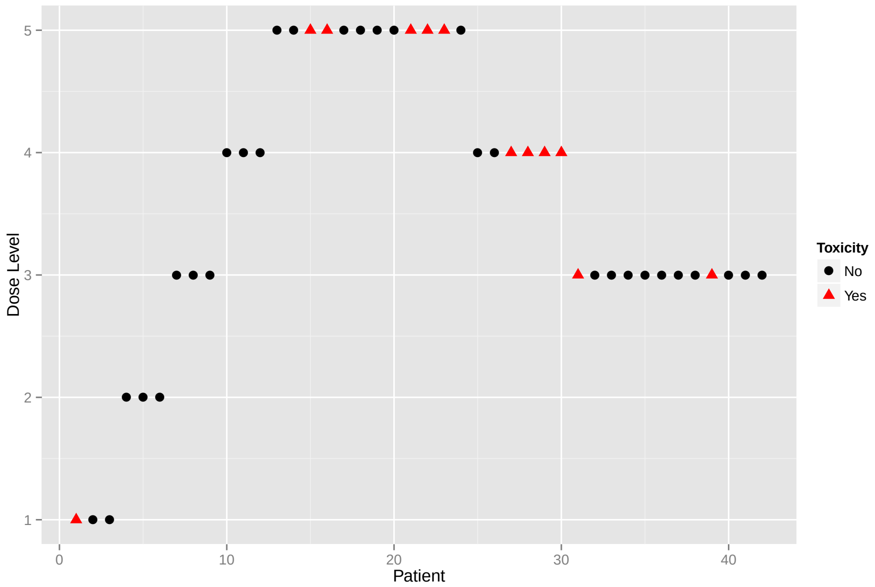


Figure 2: Plot of allocated doses and DLTs observed throughout the trial.

specifying `plot = TRUE` in the initial call to `bcrm`, or by using the `plot` function to plot a `bcrm` object. For example `plot(trial.output)` displays the results in Figure 1. In addition, to plot the dose trajectory and DLT outcomes from the trial, use the command `plot(trial.output, trajectory = TRUE)`. This gives the plot shown in Figure 2.

3.2. Assessing operating characteristics through simulations

Simulations are a key tool used to assess design operating characteristics. Of particular interest is the (average) percentage of times each dose is experimented on during the Phase I trial and the percentage of times each dose is recommended as the MTD at the end of the trial. Different simulation scenarios should be investigated assuming the a-priori guesses (e.g., in Table 1) are both correct and incorrect. For the latter a number of scenarios could be envisaged, such as assuming toxicity is 20% greater than expected, 20% less than expected, or that the risk of DLT increases at a greater/lesser rate than expected. Table 3 shows five simulation scenarios that will be investigated in this application, although it is common for a more exhaustive list to be considered.

To conduct a simulation study, set the argument `simulate = TRUE`, specify the true risks of DLT using `truep`, and choose the number of simulations using `nsims`. An object of class `bcrm.sim` is returned, which can be printed or plotted to obtain operating characteristics. The code used to carry out 1000 simulations from scenario 1 is

```
R> truep <- c(0.05, 0.10, 0.20, 0.30, 0.35, 0.40, 0.45)
R> sim.scen1 <- bcrm(stop = list(nmax = 42), p.tox0 = p.tox0,
+    dose = dose.label, ff = "power", prior.alpha = c(1, 1, 1),
+    target.tox = 0.30, start = 1, simulate = TRUE, nsims = 1000,
+    truep = truep)
R> print(sim.scen1)
```

Dose	5mg	10mg	15mg	25mg	40mg	50mg	60mg
Prior guess of risk	0.05	0.10	0.20	0.30	0.35	0.40	0.45
Scenario 1 (Toxicity as expected)	0.05	0.10	0.20	0.30	0.35	0.40	0.45
Scenario 2 (20% greater than expected)	0.06	0.12	0.24	0.36	0.42	0.48	0.54
Scenario 3 (20% less than expected)	0.04	0.08	0.16	0.24	0.28	0.32	0.36
Scenario 4 (greater rate of DLT increase)	0.05	0.11	0.24	0.39	0.49	0.60	0.72
Scenario 5 (lesser rate of DLT increase)	0.05	0.09	0.16	0.21	0.23	0.24	0.25

Table 3: Scenarios to be investigated in simulation studies.

```
Operating characteristics based on  1000  simulations:


Sample size 42

                           Doses
                                 5     10    15    25    40    50     60
  Experimentation proportion 0.0734 0.0942 0.230 0.264 0.168 0.106 0.0635
  Recommendation proportion  0.0000 0.0050 0.199 0.380 0.228 0.129 0.0590

                           Probability of DLT
                            [0,0.2] (0.2,0.4] (0.4,0.6] (0.6,0.8] (0.8,1]
  Experimentation proportion   0.398     0.539    0.0635         0       0
  Recommendation proportion    0.204     0.737    0.0590         0       0
```

The simulation shows that the correct dose (25mg) will be recommended as the MTD approximately 38% of the time. At first this may appear to be quite low. However, it should be noted that neighboring doses (15mg and 40mg) also have risk of DLT not far from the TTL of 30%. In fact, the design selects a dose with true risk of DLT between 20% and 40% approximately 74% of the time. Given the gradually increasing toxicity profile, this may be an acceptable operating characteristic for the clinician. In terms of both experimentation and recommendation, only around 6% of participants receive a dose and 6% of trials recommend a dose whose risk of DLT is greater than 0.4.

Experimentation and recommendation percentages from all simulation scenarios are shown in Table 4, by regions of true toxicity. For the scenarios investigated, the recommendation proportions are generally high for drugs whose risk is in the range $(0.2, 0.4]$. Indeed the recommendation proportions are above 90% at these levels of risk for Scenarios 3-5. This is partly due to more doses having true risk within this range for these scenarios. Of slight possible concern for the clinician is Scenario 2, where all doses have 20% higher toxicity than anticipated. Here, doses with risk of DLT > 0.40 are recommended as MTDs approximately 14% of the time. Furthermore, around 19% of patients would be dosed at these levels of risk

	True risk DLT			
Scenario	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]
Experimentation proportions				
Scenario 1	0.398	0.539	0.064	0.000
Scenario 2	0.192	0.617	0.191	0.000
Scenario 3	0.301	0.699	0.000	0.000
Scenario 4	0.184	0.675	0.136	0.004
Scenario 5	0.294	0.706	0.000	0.000
Recommendation proportions				
Scenario 1	0.204	0.737	0.059	0.000
Scenario 2	0.022	0.837	0.141	0.000
Scenario 3	0.051	0.949	0.000	0.000
Scenario 4	0.019	0.926	0.054	0.001
Scenario 5	0.033	0.967	0.000	0.000

Table 4: Operating characteristics obtained from simulations with Gamma(1,1) prior.

	True risk DLT			
Scenario	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]
Experimentation proportions				
Scenario 1	0.280	0.719	0.000	0.000
Scenario 2	0.143	0.743	0.113	0.000
Scenario 3	0.236	0.764	0.000	0.000
Scenario 4	0.143	0.775	0.082	0.000
Scenario 5	0.234	0.766	0.000	0.000
Recommendation proportions				
Scenario 1	0.082	0.916	0.002	0.000
Scenario 2	0.000	0.904	0.096	0.000
Scenario 3	0.012	0.988	0.000	0.000
Scenario 4	0.000	0.954	0.046	0.000
Scenario 5	0.005	0.995	0.000	0.000

Table 5: Operating characteristics obtained from simulations with Gamma(20,0.05) prior.

during the trial.

We can improve the behavior of the design if we are willing to be more informative in our prior beliefs. For the one-parameter power model being used here, this relates to a smaller prior variance on the parameter α. Recall, that we gave a Gamma(1,1) prior for α, which has a mean and variance of one. Hence the standardized doses are equal to the prior estimates of DLT (see Section 2.1.1). Suppose the clinician is willing to make the following statement: "I am 95% confident that the true risk of DLT at dose level 4 (our best guess of the MTD) is not higher than 0.45." This translates to the following probability statement,

$$P(\pi(d_4^*;\alpha) < 0.45) = P(\theta_4^\alpha < 0.45) = P\left(\alpha < \frac{\log(0.45)}{\log(0.30)}\right) = 0.95.$$

Evaluating this quantile and keeping the mean of the Gamma distribution as one, a suitable prior is therefore Gamma(20,0.05), i.e., a Gamma distribution with shape 20 and scale 0.05.

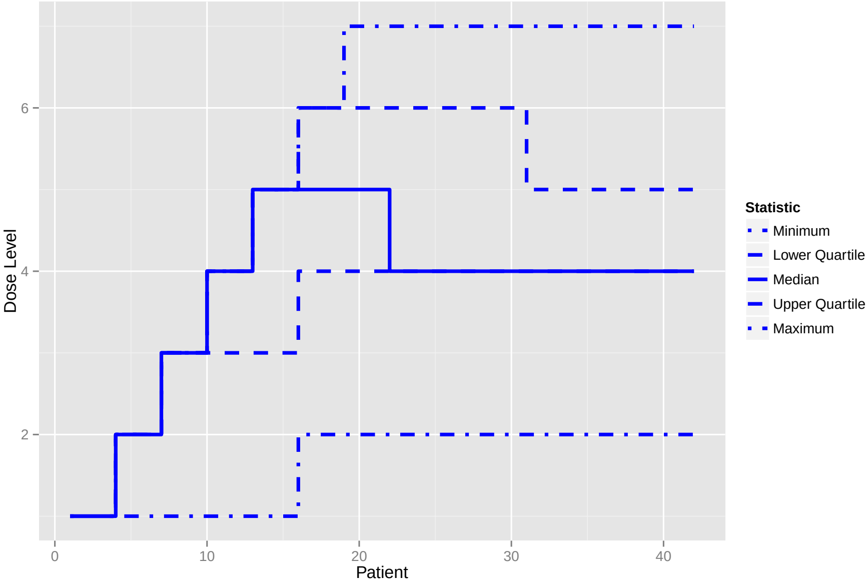


Figure 3: Summary statistics of the simulated trajectories under Scenario 1.

Table 5 shows the operating characteristics of the chosen scenarios using this more informative prior. Under Scenario 2, the experimentation proportion for doses with risk >0.4 has now reduced to 0.113, and the recommendation proportion has reduced to 0.096. In 90% or more of the simulated trials across all scenarios investigated the recommended MTD has true risk of DLT between 20% and 40%. This demonstrates how a more informative prior can improve the operating characteristics of the design amongst a range of plausible scenarios. Naturally, scenarios that deviate markedly from the prior estimates could perform worse if the prior is too precise.

With any simulation study conducted, the individual trial dose levels and outcomes can be extracted by using the `trajectories = TRUE` argument in the `print` command. Alternatively, a summary of the trajectories can be plotted. For example, `plot(sim.scen1, trajectories = TRUE)` produces Figure 3 for Scenario 1.

Comparison with the standard 3+3 design

To compare operating characteristics with the standard 3+3 design, the option `threep3 = TRUE` can be added to the original `bcrm` command, or alternatively it can be added to the `plot` command post-estimation. For example, a comparison of the operating characteristics for Scenario 1 can be made by specifying `plot(sim.scen1, threep3 = TRUE)`, which then produces Figure 4.

Variable sample sizes

Thus far, it has been assumed that the sample size has been fixed at 42 patients, perhaps based on a feasible recruitment rate. However, suppose the investigator wishes to implement a stopping rule based on the precision of the MTD estimate (Heyd and Carlin 1999), and is willing to stop the trial early if the 95% posterior credible interval for the risk of toxicity

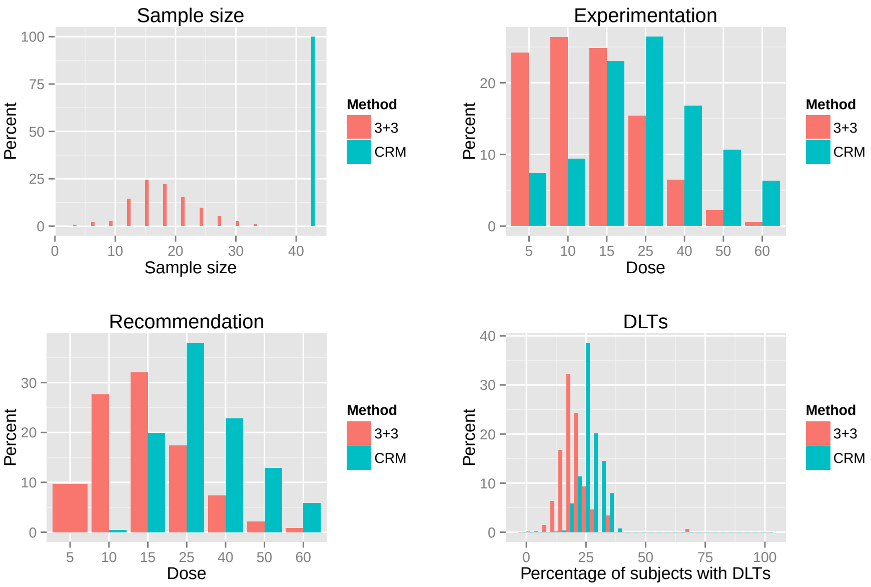


Figure 4: Comparison of operating characteristics for the CRM and 3+3 designs under Scenario 1.

at the final dose lies within 15% to 45%. However, the investigator also wishes to keep the maximum sample size at 42. Such a stopping rule can be invoked using the argument `stop = list(nmax = 42, precision = c(0.15, 0.45))`. The distribution of required sample sizes based on this stopping rule can then be investigated by conducting a simulation study. The code below does this for Scenario 1 (Table 3) and plots the operating characteristics (Figure 5), where the sample size distribution is shown in the top left panel.

```
R> truep <- c(0.05, 0.10, 0.20, 0.30, 0.35, 0.40, 0.45)
R> sim.stop <- bcrm(stop = list(nmax = 42, precision = c(0.15, 0.45)),
+    p.tox0 = p.tox0, dose = dose.label, ff = "power",
+    prior.alpha = c(1, 1, 1), target.tox = 0.30, start = 1,
+    simulate = TRUE, nsims = 1000, truep = truep)
R> plot(sim.stop)
```

The expected sample size in this scenario is 40.7, only slightly lower than the maximum specified sample size. This suggests that the majority of simulations did not reach the required precision for the MTD estimate before reaching the maximum sample size. This supports findings from other studies that the sample sizes commonly seen in Phase I trials are insufficient to achieve good precision around the MTD estimate (Iasonos, Wilton, Riedel, Seshan, and Spriggs 2008).

3.3. ‘What if’ scenario analyses

A further useful check of the performance of the design is to see how it performs under some pre-defined scenarios. In particular, suppose the investigator wishes to know the following information:

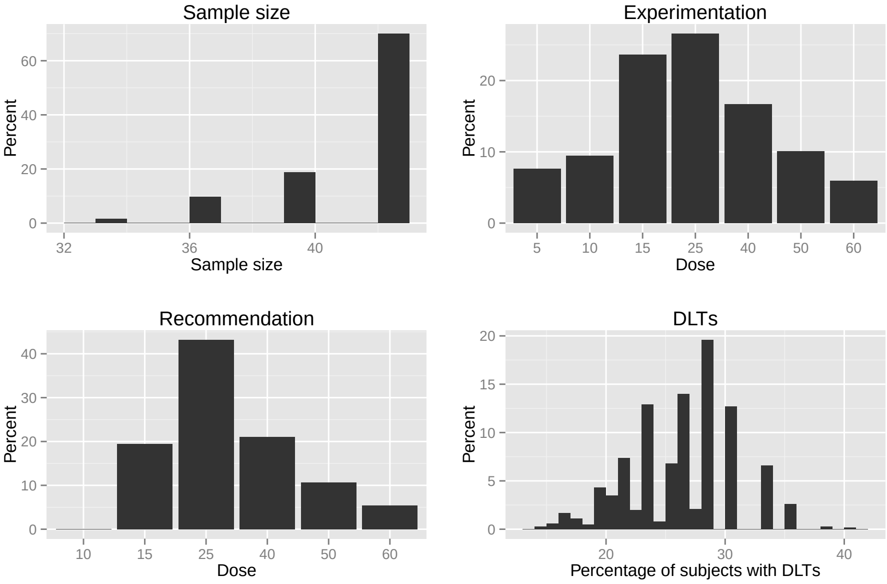


Figure 5: Plot of operating characteristics from Scenario 1 simulation study, with variable sample size based on precision of the final dose.

1. Given no toxicities are observed, how many cohorts need to be recruited before the top dose level is reached?

2. If toxicities are seen early on, does the design still escalate?

Scenario (a) can be investigated via a single simulation, setting the true risks of DLT to be zero. This is done below for the design with Gamma(1,1) prior:

```
R> truep <- c(0, 0, 0, 0, 0, 0, 0)
R> scenarioa <- bcrm(stop = list(nmax = 42), p.tox0 = p.tox0,
+    dose = dose.label, ff = "power", prior.alpha = c(1, 1, 1),
+    target.tox = 0.30, start = 1, simulate = TRUE, nsims = 1,
+    truep = truep)
R> scenarioa
```

```
Operating characteristics based on  1  simulations:

                            Doses
                                 5     10     15     25     40     50    60
  Experimentation proportion 0.0714 0.0714 0.0714 0.0714 0.0714 0.0714 0.571
  Recommendation proportion  0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 1.000

                            Probability of DLT
                             [0,0.2] (0.2,0.4] (0.4,0.6] (0.6,0.8] (0.8,1]
  Experimentation proportion       1         0         0         0       0
  Recommendation proportion        1         0         0         0       0
```

We see that 3 patients (7.14%) are dosed at each level below the uppermost (60mg), indicating that the design escalates continuously up the dose levels if no toxicities are observed (since the design uses cohorts of size 3). If a Gamma(20,0.05) prior is used, however, then 9 patients are dosed at the second highest level (50mg) before escalation to the highest dose, with all doses below 50mg receiving 3 patients each.

The code presented below addresses scenario (b) for the case where one out of three toxicities are seen in the first cohort. This is investigated for both the Gamma(1,1) prior and the Gamma(20,0.05) prior.

```
R> data <- data.frame(patient = 1:3, dose = rep(1, 3), tox = c(0, 0, 1))
R> scenariob.gamma.1.1 <- bcrm(stop = list(nmax = 3), data = data,
+    p.tox0 = p.tox0, dose = dose.label, ff = "power",
+    prior.alpha = c(1, 1, 1), target.tox = 0.30, plot = TRUE)
R> scenariob.gamma.20.0.05 <- bcrm(stop = list(nmax = 3),
+    data = data, p.tox0 = p.tox0, dose = dose.label,
+    ff = "power", prior.alpha = c(1, 20, 0.05), target.tox = 0.30,
+    plot = TRUE)
```

Both designs recommend escalation if one out of three toxicities are observed in the first cohort. However, if two out of three toxicities are observed, the design with Gamma(1,1) prior recommends remaining at the first dose level, whilst the design with Gamma(20,0.05) prior continues to recommend escalation to the second level. Indeed, although the performance of the Gamma(20,0.05) prior design was superior in the simulation scenarios investigated (Section 3.2), for this more extreme scenario where toxicities are observed early on, the design may perform not as required due to the prior now being too informative. This behavior would also have been identified if an extremely toxic simulation study was undertaken, for example assuming $> 30\%$ risk of toxicity at the lowest dose.

Indeed, the poor performance of the Gamma(20,0.05) prior when early toxicities are observed is even more apparent when considering the scenario where one out of three toxicities are observed in the first cohort and three out of three are seen in the second cohort (at the second dose level). Under this scenario, the design with Gamma(1,1) prior would recommend de-escalating back to dose level 1, whilst a Gamma(20,0.05) prior still recommends escalating to dose level 3. The performance under such 'worst case' scenarios, even if not likely, are important to report to the investigator to allow an informed judgement.

3.4. Two-parameter models

An alternative to the one-parameter models is the two-parameter logistic model. To use this model, a bivariate lognormal prior is required for the parameters (β_1, β_2). Recall that $\log(\beta_1)$ is the log-odds of the risk of toxicity at the reference dose, d_R (Equation 5). Hence, by setting

$$d_i^* = \log(d_i/25)$$

we can set the prior mean for $\log(\beta_1)$ as the investigator's prior log-odds estimate of the risk of toxicity at dose level 4 (25mg), $\log(0.30/0.70) = -0.85$. We wish to allow enough uncertainty in the prior risk of toxicity at this dose to allow escalation to higher doses if it is shown to be safe. To allow for this scenario the investigator is willing to accept the risk of toxicity at level 4 may be lower than 0.30, but they are almost certain (95%) that the risk is not lower than 0.05. Hence a prior variance for $\log(\beta_1)$ can be derived as

$$\mathrm{Var}(\log(\beta_1)) = \left(\frac{\log(0.05/0.95) - \log(0.3/0.7)}{\Phi^{-1}(0.05)}\right)^2 = 1.28^2,$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function.

We elicit the prior mean and variance for the slope parameter β_2 using the investigator's opinion about the risk of toxicity at the lowest dose. Specifically, the investigator's best guess of the risk of toxicity is 5%. However, to allow flexibility in the design to account for possible early toxicities, the investigator is willing to accept that there is a 10% chance that the risk is greater than 40%. These two statements translate into the following prior estimate and variance for $\log(\beta_2)$:

$$\log(\hat{\beta}_2) = \log\left(\frac{\log(0.05/0.95) - \log(0.30/0.70)}{\log(5/25)}\right) = 0.265$$

$$\mathrm{Var}(\log(\beta_2)) = \left(\frac{\log(0.4/0.6) - \log(\hat{\beta}_1) - \log(\hat{\beta}_2)\log(5/25)}{\Phi^{-1}(0.90)}\right)^2 = 1.98^2.$$

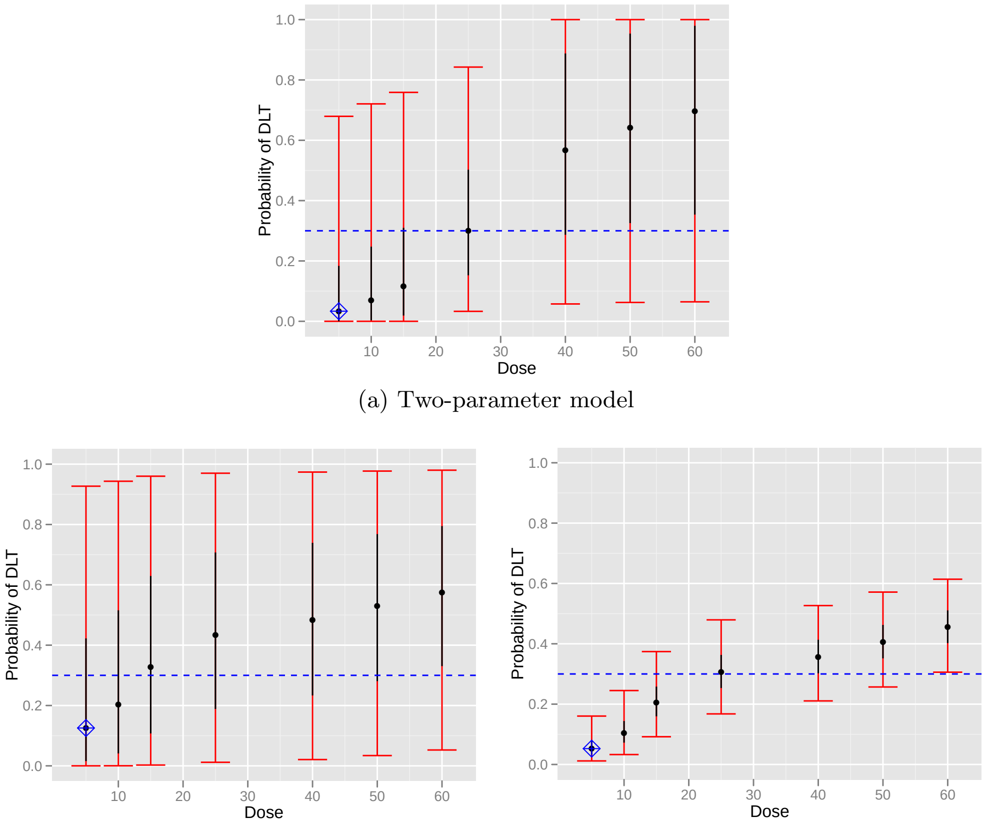
Furthermore, we assume zero a-priori correlation between the parameters.

A two-parameter model can be fit using the argument `ff = "logit2"`. As we wish to calculate the standardized doses from the doses themselves rather than from prior estimates of risk, we replace the `p.tox0` argument with the `sdose` argument, which allows us to manually specify the standardized doses. The bivariate lognormal prior is specified using `prior.alpha = list(4, mu, Sigma)`, where `mu` is the prior mean vector and `Sigma` is the prior variance-covariance matrix. Finally, it is preferable, though not necessary, to calculate the posterior distributions in a two-parameter model using MCMC methods as it is computationally faster than quadrature methods. Two options are available for this; either using **WinBUGS** via the **R2WinBUGS** package or **OpenBUGS** via the **BRugs** package. The latter is the most efficient, computationally. In either case, the user must ensure that the appropriate program has been downloaded and installed on their system before running **bcrm**. The program used to calculate the posterior distributions can then be specified using the `method` argument, e.g., `method = "BRugs"`.

The corresponding prior estimates of toxicity risk for the two-parameter model can be plotted by specifying a trial of size zero as follows:

```
R> mu <- c(-0.847, 0.265)
R> Sigma <- rbind(c(1.28^2, 0), c(0, 1.98^2))
R> bcrm(stop = list(nmax = 0), sdose = log(dose.label/25),
+    dose = dose.label, ff = "logit2", prior.alpha = list(4, mu, Sigma),
+    target.tox = 0.30, start = 1, method = "BRugs", plot = TRUE)
```

Figure 6 shows the resulting plot of the prior dose-toxicity curve, and contrasts it to the one-parameter model with Gamma(1,1) and Gamma(20,0.05) priors. The priors for the two-parameter model are seen to impose less precise risks of toxicity than the Gamma(20,0.05) one-parameter model but more precise risks than the one-parameter Gamma(1,1) model.



(a) Two-parameter model

(b) One-parameter Gamma(1,1) model

(c) One-parameter Gamma(20,0.05) model

Figure 6: Plot of prior dose-toxicity quantiles (2.5%, 25%, 50%, 75%, 97.5%) for a) the two-parameter model, b) the one-parameter model with Gamma(1,1) prior, c) the one-parameter model with Gamma(20,0.05) prior.

	Dose						
Model	5mg	10mg	15mg	25mg	40mg	50mg	60mg
One-parameter Gamma(1,1)	3	3	3	3	3	3	24
One-parameter Gamma(20,0.05)	3	3	3	3	3	3	24
Two-parameter ‘posterior mean’	3	3	3	6	3	3	21
Two-parameter EWOC $q = 0.25$	3	3	3	15	3	3	12
Two-parameter Toxicity intervals	3	3	3	9	3	3	18

Table 6: Number of individuals treated at each dose level if no toxicities are observed.

Escalation methods

The escalation decisions are to be based on the posterior mean rather than a plug-in estimate; this can be set by specifying `pointest = "mean"`. A two-parameter design based on posterior mean escalation was found to reach the highest dose (60mg) after seven cohorts had been recruited (21 patients) if no toxicities were observed (Table 6).

Alternatively, an EWOC design can be fit by specifying the quantile of the MTD distribution to be used for escalation in the `pointest` argument, e.g., `pointest = 0.25`. Since this design

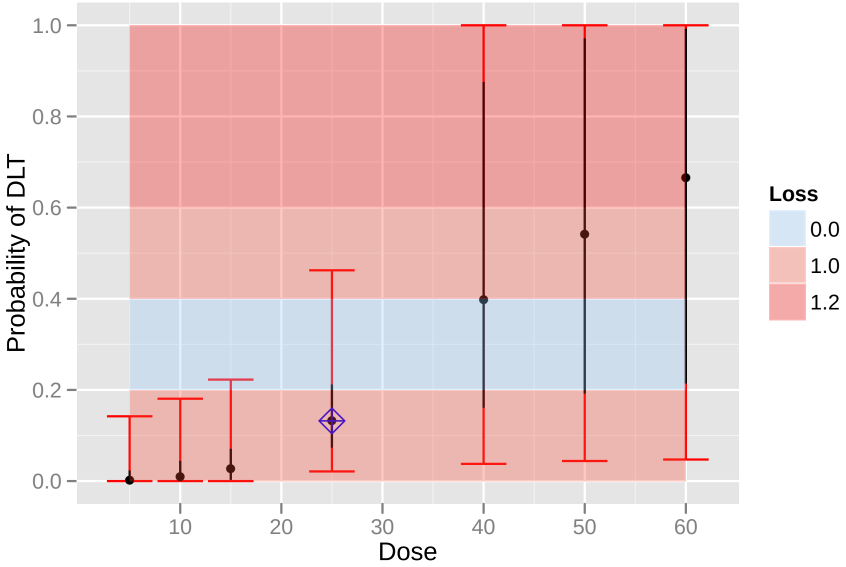


Figure 7: Plot of posterior dose-toxicity quantiles (2.5%, 25%, 50%, 75%, 97.5%) for the two-parameter model after four cohorts have been recruited to dose levels 1–4 with no toxicities observed.

is more conservative than a design that targets the posterior mean, it takes longer to reach the highest dose when no toxicities are observed. In this situation, the EWOC design would reach the highest dose after ten cohorts (30 patients) had been recruited (Table 6).

The final escalation design that can be fit in **bcrm** is an escalation based on toxicity intervals design. To fit such a model, the risk of toxicity cutpoints must be specified as a vector using the `tox.cutpoints` argument, and the associated losses using the `loss` argument. For example the arguments `tox.cutpoints = c(0.2, 0.4, 0.6)` and `loss = c(1, 0, 1, 1.2)` specifies a design that places a loss of 1 to a toxicity risk < 0.2, a loss of 0 if the risk if between 0.2 and 0.4, a loss of 1 for risk between 0.4 and 0.6 and a loss of 1.2 for risk > 0.6. Using these losses, the design reaches the highest dose after eight cohorts (24 patients) (Table 6). Indeed, with the choice of doses and priors used in this example, we found that the design could easily get stuck at dose level 4 for a long period of time, especially if a high loss was specified for risks of toxicity > 0.6. The reason for this is that while dose levels 5-7 remain untested, the posterior distribution gives a non-negligible probability to high risks of toxicity at these doses, primarily due to the logistic dose-toxicity shape being imposed. This can be seen in Figure 7 where the first 4 cohorts are specified to have no toxicities as coded below. The relatively sparse priors used in this example are driving the dose allocation and reflect the fact that there is often a higher cost for escalating to dose level 5 than staying at level 4. If the priors truly reflect the investigators uncertainty in the toxicity then this lack of escalation represents ‘reasonable behavior’ for the investigator. However, the presentation of such operating characteristics may cause the investigator to reassess their chosen priors. Indeed a loss function such as `loss = c(1, 0, 1, 2)` that penalizes high toxicity risk more extremely would not reach the highest dose when no toxicities are observed.

```
R> data <- data.frame(patient = 1:12, dose = rep(1:4, each = 3), tox = 0)
R> loss <- c(1, 0, 1, 1.2)
```

```
R> tox.intervals.bcrm <- bcrm(stop = list(nmax = 12), data = data,
+    sdose = log(dose.label/25), dose = dose.label, ff = "logit2",
+    prior.alpha = list(4, mu, Sigma), target.tox = 0.30,
+    tox.cutpoints = tox.cutpoints, loss = loss, method = "BRugs")
R> plot(tox.intervals.bcrm)
```

Simulation results for two-parameter models

A simulation study is the best way to comprehensively assess the performance of the two-parameter model and the three possible escalation procedures described. The code shown below, utilizes BRugs to conduct a simulation study of the two-parameter model using the posterior mean for escalation decisions, under scenario 1 (Table 3):

```
R> truep <- c(0.05, 0.10, 0.20, 0.30, 0.35, 0.40, 0.45)
R> sim.scen1.twoparam <- bcrm(stop = list(nmax = 42),
+    sdose = log(dose.label/25), dose = dose.label, ff = "logit2",
+    prior.alpha = list(4, mu, Sigma), target.tox = 0.30,
+    pointest = "mean", start = 1, simulate = TRUE, nsims = 1000,
+    truep = truep, method = "BRugs")
R> sim.scen1.twoparam

Operating characteristics based on  1000  simulations:


Sample size 42

                            Doses
                                 5     10    15    25     40     50     60
  Experimentation proportion 0.0753 0.0933 0.226 0.431 0.0967 0.0389 0.0391
  Recommendation proportion  0.0000 0.0060 0.196 0.558 0.1410 0.0480 0.0510

                            Probability of DLT
                             [0,0.2] (0.2,0.4] (0.4,0.6] (0.6,0.8] (0.8,1]
  Experimentation proportion   0.395     0.566    0.0391         0       0
  Recommendation proportion    0.202     0.747    0.0510         0       0
```

More generally, the performance of the three designs under scenarios 1 and 2 are shown in Table 7. The EWOC design doses fewer patients at levels with high risk of DLT but in contrast it doses more patients at suboptimal dose levels. This design also accrues fewer patients to dose levels close to the TTL (20–40% risk of DLT). The toxicity intervals escalation design performs the best in terms of recommending a dose close to the TTL, for both scenario 1 and 2. When comparing these designs to the one-parameter model with Gamma(1,1) prior (Table 4) we see that the designs are comparable under scenario 1, but perform slightly better under scenario 2. These comparisons can be very dependent on the choice of doses, model, and priors, and hence cannot usually be generalized to other situations. Of course in reality, when designing a Phase I trial, a more comprehensive simulation study should be conducted to assess the possible models and designs.

		True risk DLT		
Scenario	Model	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]
		Experimentation proportions		
Scenario 1	Two-parameter	0.395	0.566	0.039
	EWOC $q = 0.25$	0.532	0.462	0.007
	Toxicity intervals	0.415	0.573	0.012
Scenario 2	Two-parameter	0.193	0.714	0.093
	EWOC $q = 0.25$	0.288	0.687	0.025
	Toxicity intervals	0.190	0.766	0.044
		Recommendation proportions		
Scenario 1	Two-parameter	0.202	0.747	0.051
	EWOC $q = 0.25$	0.324	0.662	0.014
	Toxicity intervals	0.200	0.784	0.016
Scenario 2	Two-parameter	0.020	0.894	0.086
	EWOC $q = 0.25$	0.055	0.916	0.029
	Toxicity intervals	0.020	0.925	0.055

Table 7: Operating characteristics obtained from simulations with two-parameter models for Scenarios 1 (toxicity as expected) and 2 (toxicity 20% greater than expected).

4. Summary

The **bcrm** package has been designed to allow a comprehensive assessment and implementation of a variety of Phase I Bayesian adaptive designs of a ‘continual reassessment method’ nature. These include common one and two-parameter models with escalation decisions being based on either a plug-in estimate, the posterior mean, an ‘escalation with overdose control’ quantile, or the use of a loss function to weight intervals of toxicity risk. The syntax has been designed to be relatively simple to use, where all options are specified as arguments within an overarching function, `bcrm`. Trials can be conducted interactively or operating characteristics can be obtained using built-in simulation code.

This package focuses on a Bayesian implementation of CRM models. However, a likelihood based version of the CRM can be applied (O’Quigley and Shen 1996). Here, a two-stage design is often employed, where a rule-based ‘start-up’ is used until heterogeneity is observed in the response. A two-stage design will not be particularly advantageous in terms of efficacy but can be particulary attractive to clinicians used to dealing with the standard 3+3 design (Iasonos, Zohar, and O’Quigley 2011). The ‘start-up’ rules can also be based on lower grade toxicity information, facilitating an accelerated escalation when few low grade toxicities are observed (Iasonos *et al.* 2011). In practice, a ‘start-up’ rule can also easily be applied with the Bayesian designs described in the **bcrm** package; the accumulated information on DLTs in the ‘start-up’ stage are just incorporated using the `data` argument in the normal way. An automatic way to assess the operating characteristics of such two-stage designs is not yet implemented but will be a consideration for future releases.

The emphasis of the package is to allow an easy implementation and clear presentation of CRM designs. To this end, relevant summary and graphical output of the ongoing trial or simulation study can be easily obtained using the `print` and `plot` wrapper functions. This enables clinical trial statisticians to easily present results to clinicians and to facilitate discussion about the appropriateness of the design.

Acknowledgments

We would like to thank Graham Wheeler for providing the initial code to implement the standard 3+3 design. MJS and APM are funded by the UK Medical Research Council (grant codes U1052.00.001 and U1052.00.014)

References

Babb J, Rogatko A, Zacks S (1998). "Cancer Phase I Clinical Trials: Efficient Dose Escalation with Overdose Control." *Statistics in Medicine*, **17**(10), 1103–1120.

Cheung K (2011). ***dfcrm**: Dose-Finding by the Continual Reassessment Method*. R Package Version 0.2-1, URL http://CRAN.R-project.org/package=dfcrm.

Cheung YK (2005). "Coherence Principles in Dose-Finding Studies." *Biometrika*, **92**(4), 863–873.

Cheung YK, Chappell R (2000). "Sequential Designs for Phase I Clinical Trials with Late-Onset Toxicities." *Biometrics*, **56**(4), 1177–1182.

Cheung YK, Chappell R (2002). "A Simple Technique to Evaluate Model Sensitivity in the Continual Reassessment Method." *Biometrics*, **58**(3), 671–674.

Heyd JM, Carlin BP (1999). "Adaptive Design Improvements in the Continual Reassessment Method for Phase I Studies." *Statistics in Medicine*, **18**(11), 1307–1321.

Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR (2008). "A Comprehensive Comparison of the Continual Reassessment Method to the Standard 3+3 Dose Escalation Scheme in Phase I Dose-Finding Studies." *Clinical Trials*, **5**(5), 465–477.

Iasonos A, Zohar S, O'Quigley J (2011). "Incorporating Lower Grade Toxicity Information into Dose Finding Designs." *Clinical Trials*, **8**, 370–379.

Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM (1994). "A Comparison of Two Phase I Trial Designs." *Statistics in Medicine*, **13**, 1799–1806.

Lunn D, Spiegelhalter DJ, Thomas A, Best NG (2009). "The BUGS Project: Evolution, Critique and Future Directions." *Statistics in Medicine*, **28**(25), 3049–3067.

Neuenschwander B, Branson M, Gsponer T (2008). "Critical Aspects of the Bayesian Approach to Phase I Cancer Trials." *Statistics in Medicine*, **27**(13), 2420–2439.

Norris C, Cook J (2006). ***CRMSimulator** Version 1.1*. URL https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=13.

O'Quigley J (2002). "Continual Reassessment Designs with Early Termination." *Biostatistics*, **3**(1), 87–99.

O'Quigley J, 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**(11), 1647–1664.

O'Quigley J, Pepe M, Fisher L (1990). "Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer." *Biometrics*, **46**(1), 33–48.

O'Quigley J, Reiner E (1998). "A Stopping Rule for the Continual Reassessment Method." *Biometrika*, **85**(3), 741–748.

O'Quigley J, Shen LZ (1996). "Continual Reassessment Method: A Likelihood Approach." *Biometrics*, **52**, 673–684.

O'Quigley J, Zohar S (2006). "Experimental Designs for Phase I and Phase I/II Dose-Finding Studies." *British Journal of Cancer*, **94**(5), 609–613.

Paoletti X, Kramar A (2009). "A Comparison of Model Choices for the Continual Reassessment Method in Phase I Cancer Trials." *Statistics in Medicine*, **28**(24), 3012–3028.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Shen L, O'Quigley J (1996). "Consistency of Continual Reassessment Method under Model Misspecification." *Biometrika*, **83**(2), 395–405.

Spiegelhalter DJ, Thomas A, Best NG, Lunn D (2003). ***WinBUGS*** *Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge. URL http://www.mrc-bsu.cam.ac.uk/bugs/.

Storer BE (1989). "Design and Analysis of Phase I Clinical Trials." *Biometrics*, **45**(3), 925–937.

Sturtz S, Ligges U, Gelman A (2005). "**R2WinBUGS**: A Package for Running **WinBUGS** from R." *Journal of Statistical Software*, **12**(3), 1–16. URL http://www.jstatsoft.org/v12/i03/.

Thomas A, O'Hara B, Ligges U, Sturtz S (2006). "Making BUGS Open." *R News*, **6**(1), 12–17. URL http://CRAN.R-project.org/doc/Rnews/.

Tourneau CL, Lee JJ, Siu LL (2009). "Dose Escalation Methods in Phase I Cancer Clinical Trials." *Journal of the National Cancer Institute*, **101**(10), 708–720.

US Department of Health and Human Services (2010). *Guidance for Industry. S9 Nonclinical Evaluation for Anticancer Pharmaceuticals*. Food and Drug Administration.

Venier J (1999). ***CRM*** *Version 1.0*. URL https://biostatistics.mdanderson.org/SoftwareDownload/SingleSoftware.aspx?Software_Id=1.

Whitehead J (1997). "Bayesian Decision Procedures with Application to Dose-Finding Studies." *International Journal of Pharmaceutical Medicine*, **11**, 201–208.

Affiliation:

Michael Sweeting
MRC Biostatistics Unit
Institute of Public Health
University Forvie Site
Robinson Way
Cambridge
CB2 0SR, United Kingdom
E-mail: michael.sweeting@mrc-bsu.cam.ac.uk
URL: http://www.mrc-bsu.cam.ac.uk/personal/michaels/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 54, Issue 13 *Submitted:* 2012-05-01
August 2013 *Accepted:* 2013-03-22