International Statistical Review

International Statistical Review (2020), 0, 0, 1–24 doi:10.1111/insr.12363

Performance Measures in Dose-Finding Experiments

Nancy Flournoy[1], José Moler[2] and Fernando Plo[3]

[1]*Department of Statistics, University of Missouri, Columbia, MO, USA*
[2]*Departamento de Estadística, Informática y Matemáticas, Universidad Pública de Navarra, Campus Arrosadia s/n, Pamplona 31006, Spain*
[3]*Department of Statistical Methods, Universidad de Zaragoza, Pedro Cerbuna, 12, Zaragoza 50009, Spain*
E-mail: jmoler@unavarra.es

Summary

In the first phase of pharmaceutical development, and assuming that the probability of positive response increases with dose, the main statistical goal is to estimate a percentile of the dose–response function for a given target value Γ. We compare the Maximum Likelihood and centred isotonic regression estimators of the target dose and we discuss several performance criteria to assess inferential precision, the amount of toxicity exposure and the trade-off between them for a set of some exemplary adaptive designs. We compare these designs using graphical tools. Several scenarios are considered using simulation, including the use of several start-up rules, the change of slope of the dose-toxicity function at the target dose and also different theoretical models, as logistic, normal or skew-normal distribution functions.

Key words: Adaptive designs; optimal designs; maximum tolerated dose; performance criteria.

1 Introduction

The goal of dose-finding experiments is to estimate the dose having a targeted expected proportion of positive responses by assigning doses sequentially to cohorts near the target dose. Such designs are used in many fields. For example, Lagoda & Sonsino (2004) determine failure thresholds in electrical and material engineering; Zera (2017) describes sensory threshold estimation; Chang & Ying (2009)) work in adaptive learning. Without loss of generality, we use the language of phase I clinical trials. To develop a new drug in the context of clinical trials, the first step aims to establish a dose for which the rate of toxicity reaches a pre-specified target; subsequent trials are more concerned with efficacy. Ting (2006) and Chevret (2006) provide a comprehensive list of issues to consider in dose-finding experiments, in the framework of drug development. Each topic is discussed and analysed in a separate chapter by an expert.

Conflicting desires to avoid imprecise estimates and hazardous assignments while dose-finding raises questions about how various allocation procedures handle the trade-off between these goals. This paper attempts to consolidate considerable literature on dose-finding designs by considering both goals simultaneously. One challenge to creating useful ways to examine the trade-off is that the two goals are measured on different scales, requiring one to consider how much precision a toxic event is worth. To mitigate this problem, the efficacy and toxicity

criteria adopted in this paper are both expressed in terms of numbers of subjects to assist in simultaneously comparing designs and their settings.

Simulation is the practical approach to study the performance of dose-finding designs because the exact characteristics of estimators are typically unknown under non-linear models when dependencies are generated with the application of adaptive allocation rules. A set of five papers (viz. Storer, 1989; Stylianou & Flournoy, 2002; Ivanova *et al*., 2003; Oron & Hoff, 2013; Diniz *et al*., 2019) illustrate simulated comparisons existing in the literature that have different goals. The three first papers study the performance of several designs under different parameter values assuming responses to follow a logistic curve. Storer (1989) presents an unfruitful search for confidence intervals of the target dose; he compares several methods for constructing confidence intervals as applied to several different dose allocation rules. Stylianou and Flournoy (2002) compare the performance of several estimators of the target dose based on one allocation rule, whereas Ivanova *et al.* (2003) compare different estimators following different designs, finding that the isotonic estimator was superior to others. Oron & Hoff (2013) and Diniz *et al.* (2019) also consider non-logistic response curves. The former studies the small sample behaviour of Bayesian and random walk dose-finding, concluding that the later are more robust. Finally, Diniz *et al.* (2019) studies the loss of information that comes from discretising continuous variables under two Bayesian allocation procedures. The authors conclude that more than nine doses should be used to minimise the loss of information when discretising.

Our main goal is to provide performance measures attending to several criteria, with graphical tools, to evaluate the global performance of designs. Good performance requires accurate estimation of the target dose with a small number of total toxicities.

1.1 Design Goals

Dose-finding designs can be classified in a variety of ways, and we select several (described in Section 1.3) to represent this variety in our graphical comparisons. They are long memory (using all past observations for current dose allocations) and short memory (using only recent observations to curtail the drag of early misleading outcomes). Some designs aim to allocate subjects around the target dose to provide information about the dose–response function in the neighbourhood of the target. Others seek to allocate all subjects to the target dose and may use the dose final allocation to estimate the target. Regarding the later approach, Azriel *et al.* (2011) proves that, for any adaptive allocation rule operating under any arbitrary monotonic dose–response function, the sequence of allocations cannot converge with probability 1 to the target dose, and so specifically, any sequence of treatments that are selected by estimating the target dose cannot converge almost surely to the target dose. Along the same line, in Fedorov *et al.* (2011) practitioners are warned against so-called *best-intention designs* (those that allocate the present patient to the dose believed to be the best), because allocations may converge to the wrong point with non-zero probability. Shen & O'Quigley (1996) provides sufficient conditions for convergence of allocations with the continual reassessment method (CRM, see Section 1.3.3). Cheung (2011) revisits these conditions in order to relax them.

In interval designs, the dose assignment is replicated when the estimated toxicity rate is within a prescribed interval around the target rate, whereas if the estimate is outside this interval, the dose assigned will move toward it. Oron *et al.* (2011) proves that interval design allocations will converge to the target dose if it is the only dose within the inverse interval prescribed; it will oscillate between the two doses straddling the target dose if no dose levels are in the interval; and if there are multiple doses in the interval, allocations will converge to one of them, but not necessarily to the one closest to the target. A simulation study in Oron *et al.* (2011) shows

that a small number of realistic scenarios meet the conditions of Shen & O'Quigley (1996) for convergence using the CRM or the interval cumulative cohort design (CCD, see Section 1.3.2)

Designs to optimally estimate the parameters of quantal response curves are well known, a remarkable paper is Ford *et al.* (1992) where they obtain the optimal designs for the canonical version of a wide class of generalised linear models with a sole explanatory variable. With non-linear mean response functions, these optimal designs depend on unknown model parameters which must be estimated to start the experiment and then may be updated using study data for subsequent assignments. When parameter estimates are updated as a study progresses, the design is called *adaptive optimal*. Other model-based designs also require parameter estimation to specify dose allocation probabilities. Parametric estimates of the target dose invariably involve estimating a slope parameter, and to do this well, optimal design theory prescribes the need for observations relatively far from the target. In Fedorov & Leonov (2014), constrained optimal designs to reduce the likelihood of assignments with high toxicity potential were shown useful in studies with relatively large sample sizes. However, with small sample sizes, estimating slope parameters is problematic (as is made clear later), and although we study final parametric estimates that are functions of slope parameters, we restrict this paper to designs whose treatment-allocation procedures do not depend on slope parameter estimates.

Designs may also be constructed for dose-selection rather than estimation, a distinction made explicit with notation developed in the following section.

1.2 The Model

Assume patients arrive sequentially (or in cohorts). The probability of a toxic event is an increasing function of dose, and the toxicity function is defined as $F(x) = Pr(Toxicity|Dose = x)$. The nth patient receives dose $X_n \in \{d_1 < d_2 < \cdots < d_L\}$. The L doses are called *permissible*. The nth patient's response and indicators for the dose received are, respectively,

$$Y_n = \begin{cases} 1, \text{ toxicity;} \\ 0, \text{ non toxicity.} \end{cases} \qquad \delta_{nj} = \begin{cases} 1, \ X_n = d_j; \\ 0, \text{ otherwise.} \end{cases} \quad j = 1, \ldots, L.$$

Now, the toxicity function can be written as

$$F(x) = \mathrm{E}[Y_n = 1 | X_n = x].$$

The dose with prescribed target toxicity rate Γ is $F^{-1}(\Gamma)$.

Observe that the toxicity rate at any permissible dose is unlikely to equal Γ. If the experiment's goal is dose selection, then one seeks to identify the dose in $\{d_1, \cdots, d_L\}$ that is the closest to $F^{-1}(\Gamma)$ (or maybe closest and less than); whereas for dose-estimation, one wants a good estimator of $F^{-1}(\Gamma)$. We focus on the estimation of $F^{-1}(\Gamma)$, following recommendations given in Oron & Hoff (2013).

Some designs require model assumptions to operate, and these are described when the design is described in Section 1.3. The designs' performance is studied by simulated experiments that are described in Section 3; simulations assume underlying logistic, normal and skewed normal models of the dose–response function. For estimation of the target dose, maximum likelihood estimates are calculated assuming a logistic model, while centred isotonic estimators (Oron & Flournoy, 2017) only assume an increasing response function as described in Section 1.5.

1.3 Designs and Some of Their Properties

The catalogue of phase I designs is quite large, and an exhaustive comparative study with all of them is unfeasible. In Sverdlov *et al.* (2014), a wide survey on novel adaptive designs for phase I trials is presented. This survey is based on classifying designs as algorithm based or parametric model based. Comparisons in Section 3 are limited to a few exemplary designs, but the performance measures and comparative procedures presented in this work extend to other designs. They are described briefly in this section.

All designs studied use previous allocations and patient outcomes to allocate the next patient or patient cohort. Let $\mathcal{F}_n = \sigma(Y_j, \boldsymbol{\delta}_j : j \leq n)$ be the accrued information up to the nth patient. Then, an adaptive allocation rule induces, in each new patient n, a set of allocation probabilities

$$\pi_{nj} := P(\delta_{nj} = 1 | \mathcal{F}_{n-1}) = F(d_j) \qquad j = 1, \ldots, L. \tag{1}$$

When patients are allocated in cohorts, these definitions hold after substituting the index n with an index for the cohorts.

For a faster read, one may just note the abbreviations given to each method hereafter and skip to the next section. Abbreviations used are summarised in Appendix A.

1.3.1 The k-in-a-row design

The k-in-a-row designs (kRDs) are Markov chain-based designs that were introduced to sensory studies by Wetherill (1963), Wetherill *et al.* (1966) and Wetherill & Levitt (1966) where they are widely used. They go by a variety of names in the literature including *transformed* and *geometric rules*. If a toxicity is observed at a permissible dose, immediately lower dose is allocated next; otherwise, the same dose is administered until k consecutive non-toxic responses are observed in which case the next higher dose is assigned to the immediately individual. This rule allocates patients, asymptotically, unimodally around the target dose with the most patients assigned to one of the doses straddling the target quantile $\Gamma = 1 - (1/2)^{(1/k)}$ of the toxicity function (Oron & Hoff, 2009). So the kRD with $k = 2$ is adopted to estimate the dose having toxicity rate $\Gamma = 0.293$. Oron and Hoff also show that dose-specific allocation probabilities using the kRD converge faster to their stationary values than those of other Markovian up-and-down designs with the same target toxicity rate.

1.3.2 The cumulative cohort design

The first *interval dose-finding design* is the CCD by Ivanova *et al.* (2007). Sample size and dose dependent "no-change" intervals are determined to cluster allocations around the target using the theory of Markovian-based group up-and-down designs; all dose-specific data to date are used to make treatment decisions, rather than just those from the current cohort of subjects.

Let R_{nj} denote the observed proportion of toxic responses among N_{nj} subjects that have been assigned to dose d_j up through the nth patient. If j is the last dose used, and

$$\begin{aligned} \text{if } R_{nj} \leq \Gamma - \Delta_{Lnj}, &\text{the next subject is given dose } d_{j+1}; \\ \text{if } R_{nj} \geq \Gamma + \Delta_{Unj}, &\text{the next subject is given dose } d_{j-1}; \\ \text{otherwise}, &\text{the next subject is again given } d_j. \end{aligned}$$

The no-change limits $\{\Delta_{Lnj}, \Delta_{Unj}\}$ are solutions to the equations $c_{Lnj} = (\Gamma - \Delta_{Lnj})N_{nj}$ and $c_{Unj} = (\Gamma + \Delta_{Unj})N_{nj}$, where c_{Lnj} and c_{Unj} satisfy the so-called *balance equation* which equates the probability of increasing and decreasing the dose:

$$P(W_{nj} \leq c_{Lnj}) = P(W_{nj} \geq c_{Unj}) \qquad \Gamma < 0.5. \tag{2}$$

and W_{nj} is a binomial random variable with parameters (N_{nj}, Γ). If the rule prescribes a dose outside the range $[d_1, d_L]$, the same dose is administered again. When there is no exact solution for (2) given Γ, a solution for a binomial random variable having a toxicity rate close to Γ is used.

The study (Liu *et al.*, 2013) comparing six up-and-down designs shows that the CCD has the best overall performance. In Oron *et al.* (2011), the CCD is shown to meet the criteria for convergence more often than the continual reassessment method (which is discussed next).

1.3.3 Continual reassessment method

The CRM was introduced in O'Quigley *et al.* (1990) as follows. Consider the dose–response *skeleton* model

$$\psi(x, a) = [(\tanh x + 1)/2]^a \quad a > 0,$$

where x represents the dose and a is an unknown parameter. The a priori distribution of a is exponential with parameter one. Bayes theorem is applied as data become available to obtain and successively update the a posteriori distribution of a. The mean of the a posteriori distribution is denoted by $\hat{a}_n$ when $n-1$ patients have participated in the experiment. The first patient is assigned the permissible lowest dose. When $n-1$ patients have been allocated, the next patient will receive the dose $x \in d_1, \ldots, d_L : |\psi(x, \hat{a}_n) - \Gamma|$ is minimal.

This initial CRM model has been modified to improve its performance; Sverdlov *et al.* (2014) provides a selected review of the CRM modifications, but a deeper presentation can be found in the book by Cheung (2011) which is focused on CRM variations and their properties. In general, the skeleton is a strictly monotone sequence of prior toxicity probabilities for the L permissible doses that initiates the CRM. Simulations in Section 3 use the *step-up skeleton* of James *et al.* (2016) in which toxicity probability increments are slow (i.e. 0.05) until the prior median is reached; beyond the prior median, increments are 0.1. James *et al.* (2016) showed that this skeleton performed well in a case study.

Even though Shen & O'Quigley (1996) conjectured conditions under which CRM allocations converge to the permissible dose closest to $F^{-1}(\Gamma)$, and Azriel *et al.* (2011) proved them to be true, Oron *et al.* (2011) showed the conditions to be extremely restrictive, and Azriel *et al.* (2011) proved that CRM convergence cannot be guaranteed in general under monotonic sequences of toxicity probabilities.

1.3.4 Escalation with overdose control

The escalation with overdose control (EWOC) was introduced in (Babb *et al.*, 1998) with the goal of allocations quickly approaching the $F^{-1}(\Gamma)$ under the constraint that the predicted proportion of patients treated above $F^{-1}(\Gamma)$ be equal or less than a bound α. EWOC is driven by Bayesian updates like the CRM but is constrained to decrease the exposure to highly toxic doses. Following (Babb *et al.*, 1998), a two-parameter logistic model for the dose response curve drives the design, and $\alpha = 0.25$. When the $(n-1)$th patient is allocated, the posterior cumulative distribution function of the target, say π_n, is obtained, and the next patient is allocated in the dose $x : \pi_n(x) = \alpha$. So the nth patient receives the 25th percentile of the posterior distribution of the target.

Comparative simulation studies in Babb *et al.* (1998) show that EWOC ant CRM have similar estimation efficiency, but EWOC treats fewer patients on doses greater than $F^{-1}(\Gamma)$.

1.3.5 Benchmark designs

The benchmark designs used in this manuscript are common designs that are well characterised for non-sequential experiments but are unethical in the dose-finding environment. They provide comparative standards in some simulations.

Uniform design (UND): a dose is randomly selected for each subject. So with L permissible doses, each is applied with probability $1/L$. The allocation probabilities do not change; there is no learning.

D-optimal design (OD): doses are selected to maximise the determinant of the information matrix. The D-optimal design for logistic dose–responses was obtained by Wetherill (1963) and Minkin (1987). It prescribes equal numbers of subjects be assigned to the 0.176th and the 0.824th quantiles of the logistic function, that is, to $(\pm 1.5434 - \alpha)/\beta$. Later, Ford *et al.* (1992) found optimal designs for a variety of dose–response models, and Biedermann *et al.* (2006) addressed restricted design spaces.

1.4 Start-Up Allocation Rules

Used with procedures that require the next dose to be near the current dose, start-up rules reduce the number of patients allocated to inefficacious doses by accelerating dose assignments into a neighbourhood of $F^{-1}(\Gamma)$. Procedures that assign the next subject as close as possible to the predicted target dose are unreliable early in the study, and for these, start-up rules mitigate the likelihood of early allocations to highly toxic doses (Cheung, 2011). We consider the following start-up rules:

Escalate until first toxicity (ETk): starting at d_1, ETk assigns cohorts of size k to escalating doses until one or more toxicities appear. ET1 was studied in Ivanova *et al.* (2003).

k-in-a-row (kR): starting at d_1, kR escalates doses only after k consecutive non-toxicities are observed and stops when a toxicity appears. This contrasts with the primary k-in-a-row design (kRD) which moves to the next lowest dose when a toxicity appears and continues on from there. The 1R and ET1 rules coincide.

3+3: patients are treated in cohorts of size 3 starting at the lowest dose d_1 and escalating without skipping any permissible doses. At any dose, if no toxicities are observed among the first three patients, the next three are allocated to the next higher permissible dose; if one out of three toxicities is observed, the next three patients are treated at the same dose; otherwise, the three next patients are treated at the next lower permissible dose. When toxicities are observed in more than one third of the subjects at a dose, 3+3 stops, and the largest permissible dose that has a toxicity frequency no greater than one third is chosen as the initial dose for the primary design.

The 3+3 plays a crucial role in the phase I of clinical trials because some reviews confirm that more than 90% of trials use it or some variation (Hansen *et al.*, 2014). However, many criticisms have been raised about 3+3 designs. Note the procedure is independent of the target. Reiner *et al.* (1999) completely enumerate possible allocation sequences for the 3+3 and compute exact estimates of the target dose. For a variety of scenarios, they find that the probability of selecting an incorrect dose is excessively high. Ivanova (2006) finds toxicity rates of the selected dose vary from 0.17 to 0.26, far from the target 0.33 . But because the 3+3 rule remains common in practice, we evaluate it as a start-up rule.

1.5 Estimators

The performance of estimators of the target $F^{-1}(\Gamma)$ is our primary interest. We denote an arbitrary estimate of the target by $\tilde{F}^{-1}(\Gamma)$. We consider both the maximum likelihood estimator

and the centred isotonic regression estimator of $F^{-1}(\Gamma)$, which are abbreviated simply as MLE and CIRE, respectively. We introduce these estimators here and briefly discuss some problems that may appear when calculating them. The seriousness of these problems varies considerably by design and by the slope of F as we demonstrate in Section 3.

The MLE of the target dose is denoted by $\widehat{F}^{-1}(\Gamma)$. It depends on an assumed parametric model. Although simulations assume different underlying models drive the designs, maximum likelihood estimates are found always assuming the logistic model:

$$F(x;\alpha,\beta) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}, \quad n \geq 1. \tag{3}$$

Thus, $\widehat{F}^{-1}(\Gamma) = [logit(\Gamma) - \widehat{\alpha}]/\widehat{\beta}$, where $logit(\Gamma) = log[\Gamma/(1-\Gamma)]$, and $\widehat{\alpha}$ and $\widehat{\beta}$ are the MLEs of α and β, respectively.

Silvapulle (1981) provides necessary conditions for $\widehat{\alpha}$ and $\widehat{\beta}$ to exist. The literature contains kludges to "fix" the problem of non-existent MLEs, but we take the position that failure of Silvapulle's conditions to hold implies failure of the experiment: either more observations are needed or a new study is needed with a different set of permissible doses. That is, algorithmic fixes should not be applied to force the existence of MLEs; important information is contained in this failure. Hence, we consider the failure of Silvapulle's conditions to hold to be an important design performance criterion. In addition, even when Silvapulle's conditions hold, common algorithms used to obtain the MLE may fail (Heinze & Ploner, 2003).

Isotonic estimation only requires that the dose–response function increases monotonically, which implies $F(d_1) < \cdots < F(d_L)$. Let N_{ni} and T_{ni} denote the number of subjects and the number of toxic responses observed on treatment d_i up through the nth patient, respectively; and let $R_{ni} := T_{ni}/N_{ni}$ denote the observed proportion of toxicities. The isotonic regressors, denoted $\{\breve{F}(d_i)\}$, minimise the weighted least squares expression $\sum[R_{ni} - \breve{F}(d_i)]^2 N_{ni}$ subject to $\breve{F}(d_1) < \cdots < \breve{F}(d_L)$. The Pool Adjacent Violators Algorithm (PAVA) produces the isotonic estimates at the permissible doses (Robertson *et al.*, 1988). Traditionally, these estimates are connected over the range $[d_1, d_L]$ via a step function. Centred isotonic regression modifies PAVA by forcing strict monotonicity and using the allocation frequencies, $\{N_{ni}\}$, to locate increases in the dose–response function estimate between permissible doses (for details, see Oron & Flournoy; 2017). We use $\breve{F}(x)$ to denote the CIRE of $F(x)$ going forward. $\breve{F}^{-1}(\Gamma)$ will exist by inverse estimation methods unless $(\breve{F}(d_1), \breve{F}(d_L))$ fails to span Γ. In this general setting, there is no closed form expression for the mean or variance of CIRE.

The rest of the paper is organised as follows. Section 2 introduces performance measures (in addition to estimators' existence rates) that are used to evaluate the designs. In Section 3, general simulation procedures are presented along with an analysis of the results.

2 Design Performance Measures

2.1 Inferential Measures

2.1.1 The root-mean-square error

The mean squared error is a standard measure of the inferential performance of an estimator. Its square root is useful because it has the same units of measurement as does the estimator and its standard deviation. However, because many acceptable doses may have toxicity rates close to the target Γ, any estimator having $F[(\breve{F}^{-1}(\Gamma)]$ close to Γ may be acceptable. Therefore,

we evaluated both root-mean-square error (RMSE)$[\tilde{F}^{-1}] = E\{[\tilde{F}^{-1}(\Gamma) - F^{-1}(\Gamma)]^2]\}$ and RMSE$[F(\tilde{F}^{-1})] = E\{[F(\tilde{F}^{-1}(\Gamma)) - \Gamma]^2\}$ for both MLE and CIRE. However, rescaling to the toxicity scale does not change the ordering of the designs' performance , and we did not obtain additional insight rescaling to RMSE$[F(\tilde{F}^{-1})]$. So only results using RMSE$[\tilde{F}^{-1}]$ are shown.

2.1.2 D-optimal efficiency

The design points $\{d_1, \ldots, d_L\}$ and the allocation frequencies $\{N_{n1}/n, \ldots, N_{nL}/n : \sum N_{ni} = n\}$ together comprise a design ξ_n which is said to be optimal and denoted by ξ^* when it maximises (or minimises) a criterion function. A fixed design ξ_n under likelihood $\mathcal{L}_n \equiv \mathcal{L}_n(\theta)$ has information $M(\xi_n) \equiv \mathrm{E}[(\partial \log \mathcal{L}_n/\partial\theta)(\partial \log \mathcal{L}_n/\partial\theta^T)|\xi_n]$. In this paper, $\mathcal{L}_n = \prod_{i=1}^{L} F_i^{T_{ni}}(1 - F_i)^{(N_{ni}-T_{ni})}$, where $\theta = (\alpha, \beta)$. Classical optimality criteria are expressed in terms of a concave (convex) function $\phi[M(\xi_n)]$ of information matrices. Our benchmark D-optimal design (Section 1.3.5) maximises the determinant of $M(\xi_n)$.

Using information in the D-optimal design for reference, a measure of ξ_n's inferential quality is $\gamma(\xi_n) \equiv \phi[M(\xi_n)]/\phi[M(\xi^*)]$, $\gamma(\xi_n) \in [0, 1]$, and the closer to 1 the better. Because the D-optimality criterion is positive homogeneous [i.e. for any information matrix M; $\phi(\delta M) = \phi(M)/\delta$], $\gamma(\xi_n)$ is the fraction of the sample size needed using ξ^* to get the same inferential precision as using ξ_n. In other words, the percentage increase in the number of patients needed using ξ_n instead of ξ^* to obtain the same criterion value is *the percentage loss of information*, $[1 - \gamma(\xi_n)] * 100$.

When a random rule is used to allocate n patients to doses $d_1, \ldots, d_L$ sequentially or in cohorts, the allocation proportions $\{N_{n1}/n, \ldots, N_{nL}/n\}$, the information $M(\xi_n)$ and the efficiency $\gamma(\xi_n)$ are stochastic processes. We also define *relative mean a posteriori efficiency* as $\gamma_n := E[\gamma(\xi_n)]$.

2.2 Ethical Measures

2.2.1 The overall proportion of toxicities (p_n) and its standard deviation $\sigma(p_n)$

From an ethical point of view, minimising the frequency of toxic responses is a natural criterion. Its standard deviation measures the precision of a design with respect to its expected toxicity rate, a measure of ethical reliability so to speak. We evaluate the overall proportion of observed toxicities through the nth patient, $p_n \equiv n^{-1} \sum_{i=1}^{n} Y_i$, and its standard deviation, $\sigma(p_n)$, as ethical criteria.

2.2.2 Allocation measures ($\mathbf{g_n}$)

An *allocation measure* (g_n) summarises the closeness of the allocations to the target dose. Criteria based on an allocation measure are surrogate ethical criteria because they imply a toxicity that occurs far from the target is worse than a toxicity that occurs at the target. The use of allocation measures as ethical criteria asserts that a design with patients closer to the target is ethically preferable even if it produces exactly the same overall observed toxicity rate as a design with more diverse dosing.

An example of a reasonable measure that increases with the distance of a patient's allocation from the target (on the toxicity rate scale) but penalises overdosing is

$$g_n \equiv \sum_{d_i \leq \text{target}} \frac{N_{ni}}{n}[F(d_i) - \Gamma]^2 + \sum_{d_i > \text{target}} \frac{N_{ni}}{n}[F(d_i) - \Gamma].$$

Although popular in the literature, we do not use allocation measures as performance criteria because they presume that the dose at which the toxicity occurs is more important than the event.

3 Comparing Designs And Target Dose Estimators

Now, tables and graphs summarising simulation data are provided to contrast the performance of the designs with respect to their inferential and ethical performance.

3.1 The Simulation Setup

The graphs and tables presented in Section 3.2 use summary statistics obtained by simulating patients' responses under the *primary designs* (i.e. 2RD, CCD, CRM and EWOC) as described in Section 1.3. A complete simulated clinical trial is called a *run*. The dose–response functions used to simulate subjects' responses to treatment are called *generating* models or generating functions (in contrast to models integral to allocation decisions and models assumed for analyses). Four parametric generating models of $F(x)$ were considered: the logistic and normal cumulative distribution functions, and the skew-normal with shape parameters 3 and -3.

For each primary design and each generating model,

(a) one thousand clinical trial runs, each with 100 patients, are simulated and called an *ensemble* as in Oron & Hoff (2013);
(b) the target dose is arbitrarily fixed between d_7 and d_8 at $F^{-1}(\Gamma) = 7.25$; so $\Gamma = F(7.25) = 0.2929$. This condition induces a link between the parameters of $F(x)$. For the logistic model, $\alpha = logit[\Gamma] - \beta * F^{-1}(\Gamma)$. For the normal model with parameters μ and σ, $F^{-1}(\Gamma) = \mu + \sigma * z_\Gamma$, where z_Γ is the Γ-quantile of the N(0,1) distribution. For the skew-normal with location and scale parameters μ and σ, and a fixed λ_0 shape parameter, the relationship is $F^{-1}(\Gamma) = \mu + \sigma * w_\Gamma$, where w_Γ is the Γ-quantile of the skew-normal with parameters 0, 1 and λ_0;
(c) to create a varied mixture of scenarios in each ensemble, the generating model's slope is chosen randomly for each run. Specifically, the slope of $F(x)$ at the target dose is $\eta := f(F^{-1}(\Gamma)) = tan(\theta)$, where $f(x) := dF(x)/dx$ and θ denote the angle between $F(x)$ at the target $[x = F^{-1}(\Gamma)]$ and the dose-axis x. For each run, θ is randomly chosen from the range (0, 35). Other model parameters are obtained from Γ and η by solving the following equations: $\eta = \beta * \Gamma * (1 - \Gamma)$ for the logistic model; $\eta = \phi(z_\Gamma)/\sigma$ for the normal model, where $\phi(x)$ is the density function of the N(0,1); and $\eta = g(w_\Gamma)/\sigma$ for the skew-normal model, where $g(x)$ is the density function of the skew-normal with parameters 0, 1 and λ_0. To illustrate the range of scenarios considered, Figure 1 and Table 1 display a sample of logistic generating functions and gives their toxicity probabilities for selected values of θ, at d_6, d_7 and d_8 together with their corresponding β values; and
(d) the set of permissible doses is arbitrarily set to be $\{1, 2, 3, \ldots, 13\}$.

Following simulations with random slopes, this general setup is modified in two ways: (i) in Section 3.3, separate simulations are carried out for fixed values of θ to show the effect of the slope at the target dose on a design's performance, and (ii) in Section 3.4, the effect of using a start-up rule prior to the primary design is studied.

In this paper, simulations of both CRM and EWOC were performed with the R-package *bcrm* which is profusely explained in Sweeting *et al.* (2013). R code to implement both the kRD and the CCD can be obtained from the corresponding author upon request.

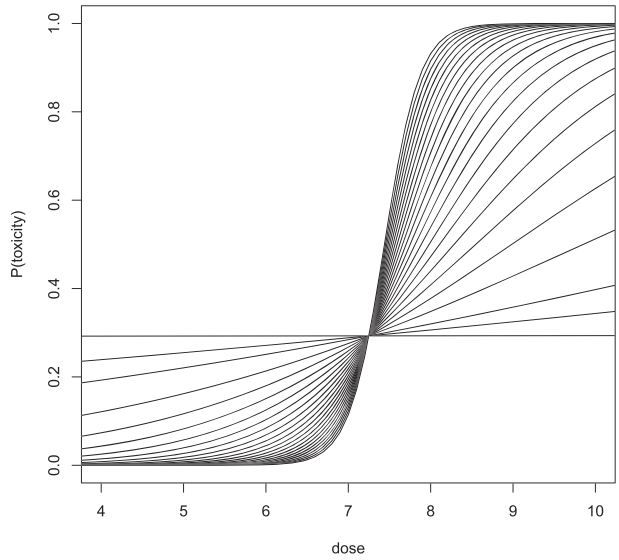


Figure 1. *Logistic functions for which* $F(7.25) = 0.2929$ *and* $\tan(\theta) = F'(7.25)$.

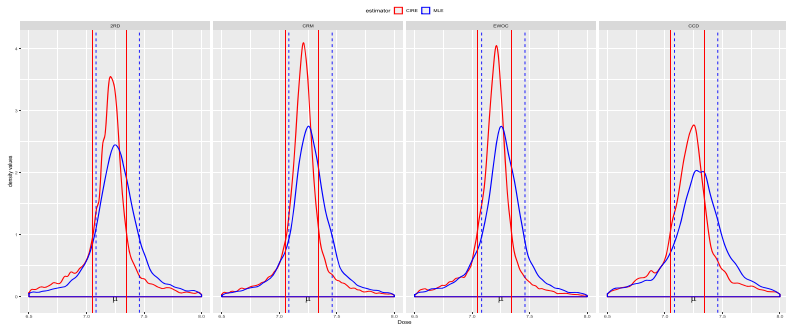
Table 1. *Values of* β *and toxicity probabilities* $F(x)$ *for doses* $x = \{6, 7, 8\}$ *are given for selected angles* θ.

θ	β	Dose 6	Dose 7	Dose 8
0.01	0.0008	0.2927	0.2928	0.2930
1.01	0.09	0.2714	0.2885	0.3063
2.01	0.17	0.2510	0.2842	0.3199
4.01	0.34	0.2134	0.2757	0.3481
8.01	0.68	0.1505	0.2590	0.4081
12.01	1.02	0.1029	0.2427	0.4723
18.01	1.57	0.0550	0.2186	0.5735
24.01	2.15	0.0274	0.1948	0.6752
28.01	2.57	0.0164	0.1790	0.7398
32.01	3.02	0.0094	0.1630	0.7994

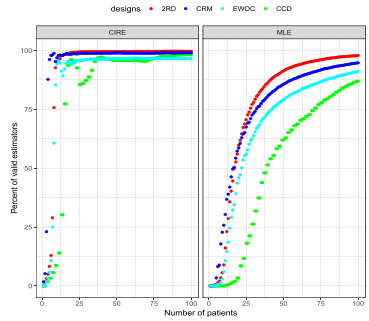
3.2 A Snapshot Comparison of Centred Isotonic Regression and Maximum Likelihood Estimates of the Target Dose

Detailed design comparisons for ethical and inferential criteria are given in Section 3.3. Preliminary to these comparisons, Figure 2 motivates the study of CIR estimators in addition to the more traditional ML estimators of the target dose (simulated wtih a logistic generating function with random slopes). Figure 2a displays density estimates of CIREs and MLEs for the moderately large sample size $n = 75$ for each of the primary designs: 2RD, CRM, EWOC and CCD. Both estimators cluster around the target dose but the CIREs are less variable than MLEs for all designs. MLEs are also more skewed. These observations hold for $n = 50$ and 25 (not shown), but curves spread and become increasingly rough.

As discussed previously, not all runs produce valid estimators, and it needs to be kept in mind that only valid estimators appear in Figure 2a and subsequent tables and graphs. Figure 2b shows the sample size required to obtain estimators. Observe that valid CIREs are much more likely than are valid MLEs. The CIRE can be obtained with 30 patients almost 100% of time for



(a) Density plots for CIREs (red) and MLEs (blue) with their Q1 and Q3 quartiles (vertical lines) for sample size 75 assuming a logistic generating function. The logistic model also is assumed for ML estimation. μ marks the target dose. Designs (from left to right) are 2RD, CRM, EWOC and CCD.



(b) Percent of runs with valid estimators by sample size for CIRE (left) and MLE (right) assuming a logistic generating function. The logistic model also assumed for ML estimation. Designs are 2RD (red), CRM (blue), EWOC (cyan) and CCD (green).

Figure 2. *Centred isotonic regression and maximum likelihood estimation of the target dose. [Colour figure can be viewed at wileyonlinelibrary.com]*

any design and generating model (data not shown). By $n = 50$, valid CIREs can be produced from all designs with probability close to 100, whereas obtaining valid MLES is still problematic at $n = 100$. Consequently, statistics formed with MLEs in Figures 3 and 4 are made with fewer observations than are statistics formed from CIREs.

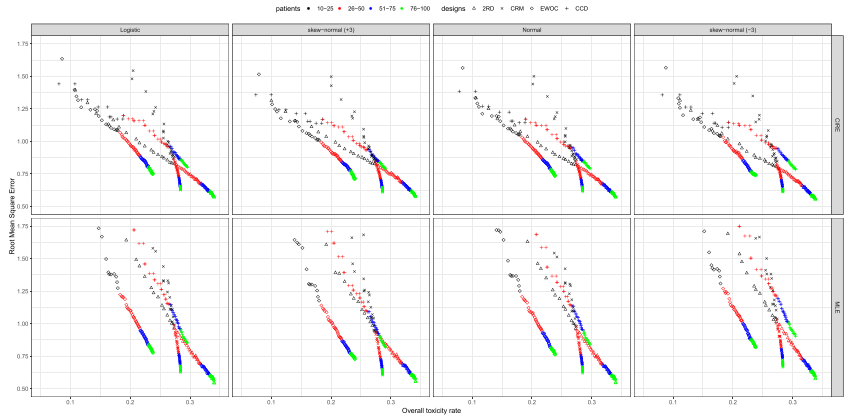
The 2RD is most likely to produce valid MLEs, followed by the CRM. We elaborate on components required to obtain estimators in Section 3.4.

3.3 Compound Ethical and Inferential Criteria

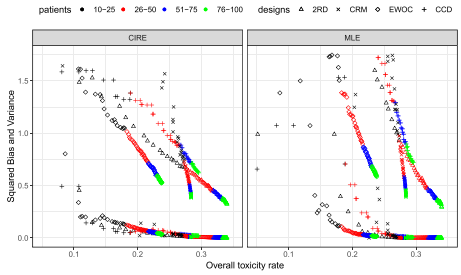
As stated in the introduction, a good design would provide accurate estimates with a small number of total toxicities. However, inferential and ethical criteria usually compete, so that improving one of them entails a worsening of the other. Approaches proposed for balancing opposing criteria involve different strategies for allocating patients to doses. We present a graphical evaluation of the trade-offs between ethics and inference in a design.

3.3.1 Toxicity-efficicacy trade-off headlines from inspecting Figures 3 and 4

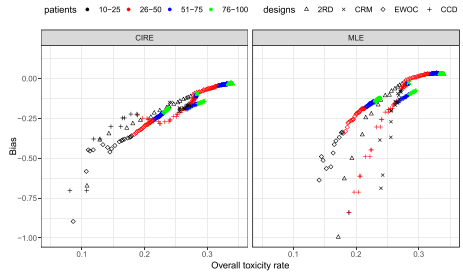
Figure 3 displays cumulative average values of two competing criteria that are obtained from the simulated experiments with random slopes as the sample size n evolves from 10 to 100 for



(a) RMSE sequences for CIR (above) and ML (below) estimates of the target dose assuming a logistic analysis model with logistic, skew-normal (+3), normal and skew-normal (-3) generating functions.



(b) Variance (upper sequences) and Squared Bias (lower sequences) by Toxicity Rate for CIRE (left) and MLE (right) assuming a logistic generating and analysis functions.



(c) Bias by Toxicity Rate for CIRE (left) and MLE (right) assuming logistic generating and analysis functions.

Figure 3. *Trade-off between ethics and estimation for primary designs as the sample size increases. Colours distinguish blocks of sample sizes: 10–25 (black), 25–50 (red), 50–75 and 75–100 (green). Symbols represent designs: 2RD (△), CRM (×), EWOC and CCD (+) [Colour figure can be viewed at wileyonlinelibrary.com]*

each of the primary designs introduced in Section 1.3. The x-axes in all subplots are values of a single ethical criterion, namely, the overall toxicity rate. In Figure 3a, the subplots' y-axes are RMSE values for the CIRE (first row) and MLE (second row); plots in each column are produced under different generating dose–response models; the y-axis in Figure 3b is squared bias and variance, and the y-axis in Figure 3c is bias. Designs in Figure 3 are represented by different symbols. Colours change to distinguish batches of sample sizes, which are black, red, blue and green for $n \in \{10-25\}, \{26-50\}, \{51-75\}$ and $\{76-100\}$, respectively. Designs' performance, with respect to the two chosen criteria, are contrasted by examining graphs of the trajectories of these pairs of values as n evolves.

The box plots in Figure 4 use the same data as were used to produce the averages in Figure 3 but restricted to $n \in \{25, 50, 75\}$. These box plots provide marginal summaries at fixed sample sizes.

(a) The relative performance of designs does not depend on the generating dose–response model: looking across the columns in Figure 3a, one sees similar patterns for all generating

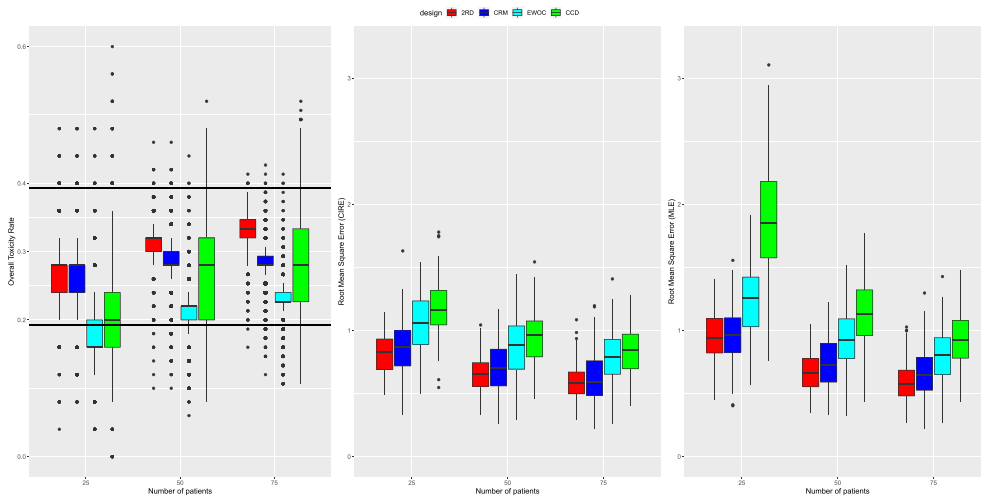


Figure 4. *Box plots for the toxicity rate (left), RMSE of CIRE (middle) and RMSE of MLE (right) assuming a logistic generating function for designs 2RD (red), CRM (blue), EWOC (cyan) and CCD (green). The logistic model is also assumed for ML estimation. [Colour figure can be viewed at wileyonlinelibrary.com]*

models. Figure 3a demonstrates that the comparative behavior of designs can be satisfactorily evaluated with only one generating model. While the skew-normal (−3) generating model produces some slightly noticeable variations in this figure and in other statistics studied, they are not sufficient to merit replications of tables and figures by generating function. Thus, only results using the logistic generating model are shown going forward.

(b) RMSE(MLE) and RMSE(CIRE) behave similarly as functions of observed toxicity rates: variability of target dose estimates decreases as observed toxicity rates increase: Figures 3a are quite similar for both estimators. As expected when starting at a low dose, early batches of subjects (black in Figure 3a) have relatively low toxicity rates with relatively high RMSE. But RMSE(MLE)s start higher than RMSE(CIRE)s. Hence, for small sample sizes, RMSE(MLE)s tend to be slightly larger than RMSE(CIRE)s.

The evolution of each design generally moves from 'left and up' to 'down and right' in this figure. Except for a little early wiggle with the CCD, sequences do not change direction (left-right-left or down-up-down) as sample sizes increase. 2RD's sequences cross the CRM's and CCD's. Toxicity rates, in general, increase as the RMSEs decrease. This suggests a dependence between the precision of target dose estimates and the frequency of toxic events. With large sample sizes (blue and green) however, toxicity rates are less associated with RMSEs particularly for the CRM.

(c) Toxicity rates and inferential precision change little after samples sizes reach 50 for all the designs: in Figure 3a, green and blue points (marking statistics for $n \geq 50$) become increasingly compressed for all the designs almost overlapping in the blue colour. This suggests that expected toxicity rates converge and little is to be gained regarding inferential precision (as measured by RMSE) as the sample size continues to increase above 50.

(d) The bias contributes negligibly to the RMSE: Figure 3b displays separate sequences for the squared bias and the variance, again as functions of the observed toxicity rate. After $n = 25$, sequences of squared biases lie close to zero whereas sequences of variances asymptote at much higher values. This demonstrates that MSEs are dominated by the variances after $n = 25$ regardless of the estimator, CIR or ML.

Zooming in, Figure 3c graphs bias versus observed toxicity rates. CIRE bias is negative for all the designs, growing with sample size and with toxicity rate to converge near zero. Observe that CIRE bias moves in a very short range of values for the CRM. MLE bias behaves similarly except that with the 2RD, it becomes sightly positive after $n = 25$.

(e) The 2RD has the best inferential performance, while EWOC settles on the smallest expected toxicity: observe in the subplots of Figure 3a that, for any batch of patients, the 2RD symbol (triangle) is closer to the horizontal axis (i.e. zero RMSE) than any other symbol. For instance, RMSEs for sample sizes 26–50 (red) in the 2RD trajectory are contained within CRM's RMSE values with sample sizes 51–75 (green).

In the top row of Figure 3a, one sees EWOC's complete sequences with RMSE(MLE) falling to the left of other designs', and in the bottom row with RMSE(CIRE), most of EWOC's sequences are to the left of other designs' demonstrating almost uniformly less expected toxicity. However, EWOC's RMSE values are compromised relative to other designs. Note, for example, that with large samples (green and blue), EWOC's RMSE values are exceeded only by the CCD's and are markedly higher than the 2RD's and CRM's. Furthermore, RMSE values in EWOC sequences for moderate sample sizes (red) barely overlap if at all with 2RD's and CRM's. There is considerable overlap among the sequences for small sample sizes (black).

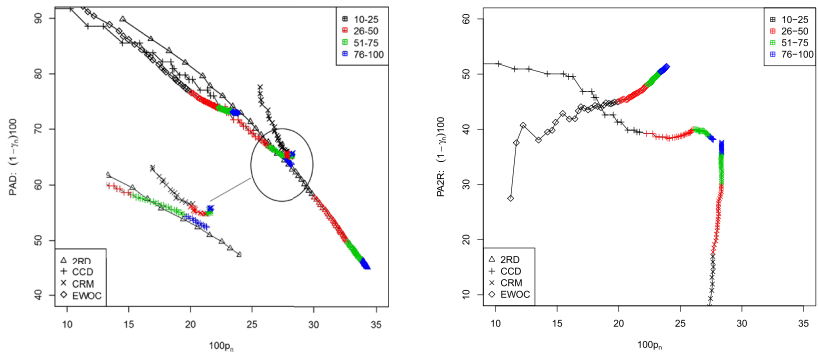
This trade-off is also seen in the box plots of Figure 4 for $n = 25$, 50 and 75. Note that these box plots are each constructed from 10 000 points; so if 1% of the data were beyond the whiskers, there would be 100 such points. Following (Diniz *et al.*, 2019) , solid bars are placed at $\Gamma \pm 0.1$ across the panel of box plots of toxicity rates. At $n = 25$, EWOC's toxicity rate box is almost entirely below this line, and while it rises with sample size, it remains well below Γ. For $n = 75$, EWOC's box is below 0.25 while 2RD's box is above 0.3. However, 2RD's and CRM's boxes and whiskers remain within the $\Gamma \pm 0.1$ bounds for $n = 25$, 50 and 75, while CCD's box comes within bounds at $n = 50$, but CCD's whiskers extend outside even at $n = 75$. Asymptotically, 2RD places more subjects on the two doses straddling the target than on any other doses (Oron & Hoff, 2009).

In contrast for RSMEs, CCD's boxes are highest followed by EWOC's. 2RD's and CRM's boxes overlap substantially with 2RD's generally being lower. EWOC's RSME boxes at $n = 50$ come to be centred at magnitudes comparable to 2RD's and CRM's at $n = 25$, but with slightly larger spread.

3.3.2 Designs' efficiency headlines from inspecting Figure 5

Figure 5a plots the percentage of additional patients needed to have the same inferential precision as the benchmark D-optimal design [(1-γ_n)*100, abbreviated PAD] versus toxicity rates for $n = 10, \ldots, 100$. Figure 5b plots the percentage of additional patients needed to have the same inferential precision as the 2RD (abbreviated PA2R) versus toxicity rates for $n = 10, \ldots, 100$. Symbols and colours in Figure 5 are the same as in Figure 3. These plots are obtained under logistic generating and analysis models with random slopes. As observed in Figure 5, the patterns reported are similar to those seen using other generating functions (not shown).

(a) The PAD is lowest using the 2RD; PADs generally decrease as the toxicity rate increases. Figure 5a shows general reductions in PADs with increasing toxicity rates and sample sizes. Reductions are steady with 2RDs and CCDs. PADs decrease dramatically over the small sample size segment (black) of 2RD experiments. This decrement continues progressively more slowly in subsequent batches of patients. 2RD attains efficiencies with



(a) Percentage of additional patients needed to have the same inferential precision as the benchmark D-optimal design versus the percent of toxic responses.

(b) Percentage of additional patients needed to have the same inferential precision as the 2RD versus the percent of toxic responses.

Figure 5. *Samples sizes required to equalise precision of target dose estimates (Note: y-axis differ in the two figures). [Colour figure can be viewed at wileyonlinelibrary.com]*

small sample sizes (black) that are higher that the highest efficiencies obtained by the other designs. For all batches, CRM's PADs are smaller than either CDD's or EWOC's.

(b) Strangely, in the larger segments (green and/or blue), CRM's and EWOC's PAD value slightly increases as the number of patients increases. The range of variation under CRM, both in PAD and in toxicity rate, is very small. Surprisingly, trajectories under the CRM and EWOC designs have a turn point after which their PAD increases (albeit slightly) with increasing sample size; in other words, their MLEs become less accurate. This effect can be seen in the enlargement within Figure 5a.

(c) CCD, EWOC and CRM require substantially more subjects in order to match estimation precision of 2RDs. Because the 2RD was found to outperform other designs with respect to the D-optimality criterion and because D-optimal designs are unattainable benchmarks, other designs are compared to the 2RD in Figure 5b. Substantially, more subjects required in CRM and EWOC experiments to match the precision of 2RD target dose estimates. EWOC's PADs increase with toxicity rate and sample size. CRM's PADs increase with sample size, but it is virtually invariant to toxicity rates. CCD's PADs slightly decrease over small sample sizes but are fairly invariant to sample sizes and toxicity rates for $n > 25$ (red, green and blue segments). Figure 5b shows that EWOC requires around 45–55% more patients than 2RD to achieve the same precision estimates of the target dose. Additional patients required by the CRM rockets from 10% for very small sample sizes to over 30% for $n \geq 50$. CCD requires about 40% more patients for $n \geq 20$.

(d) The uniform design clearly is outperformed by the other designs according to both ethical and inferential criteria. For the sake of clarity, UND's trajectories were not included in Figures 3 to 5. UND's mean toxicity rate sticks to 40% from the first patient, as expected. This high toxicity rate, very far from the target, does not bring better inference: the UND's RMSE and loss of information are outperformed by the other designs. The expected loss of information remains in the same level, 90%, from the first patient. The RMSE decreases slowly from 0.76 with 50 patients to 0.73 with 100, which is outperformed by 2RD and CRM. Observe that the main difference between the UND design and the other designs is

that any dose can be applied to any patient with no learning from previous responses and allocations; this explains its poor toxicity performance.

3.4 *Influence of the Dose–Response Slope on Designs' Performance*

As can be seen in Table 1, varying the slope of the logistic function covers a wide range of toxicity response patterns, including extreme situations. For example, $\theta = 0.01$ corresponds to a dose–response curve that is almost flat, whereas $\theta > 30$ reflects a large change in toxicity rates between doses 6 and 8. In simulations reported in this section, the slopes of the generating logistic functions are not randomly chosen. Instead, separate simulations are performed at several fixed slope values.

In Figures 7 and 6, the upper x-axis $\beta = \tan(\theta)/[\Gamma(1-\Gamma)]$ has tick marks corresponding to $\theta = (0.01 + 2 * i)^o$, $i = 1, \ldots, 23$ on the lower x-axis. In Figure 7, the y-axis represents the first value n for which at least 90% of runs provide a valid MLE (Figure 7a) and CIRE (Figure 7b). The horizontal dotted line at $n = 20$ is plotted for reference. In Figure 6, two different values are represented on the y-axis: along the top of the graph are filled circles which are the proportion of runs for which Silvapulle's conditions hold before the end of an experiment with $n = 100$. The lines are the expected first patient for which Silvapulle's conditions are fulfiled (when they hold before $n > 100$). Table 2 summarises the mean (standard deviation) sample size required for Silvapulle's conditions to hold for selected slopes.

Figures 8 and 9 show plots of the RMSE(CIRE) and the toxicity rate, respectively, as a functions of n. Subfigures are produced from simulations with $\theta = 1.01$, 4.01 and 8.01 selected to represent small, moderate and large slopes at $F^{-1}(\Gamma)$ (Table 1). Figure 9 includes $\pm$ standard deviation bars at $n = 10$, 25 and 45 which are plotted with a slight delay to avoid overlapping bars.

3.4.1 *Headlines about the effect of slope on inference and toxicity*

(a) To have a high probability of successfully calculating the MLE for small and large slopes the sample size must be large. Figure 6 shows that, as the slope grows, the expected number

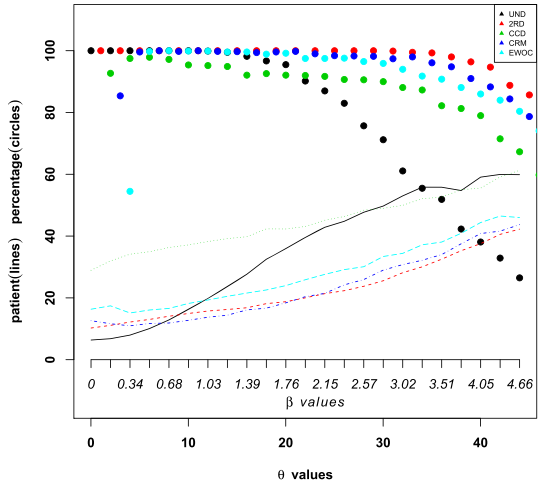
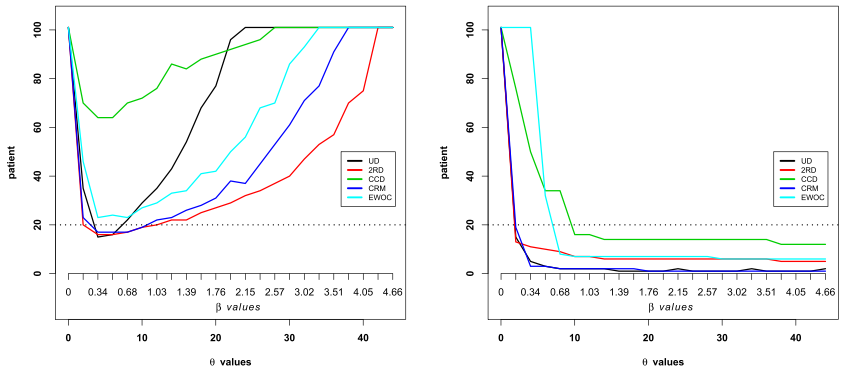


Figure 6. *Existence of MLEs. Filled circles are the percent of runs in which Silvapulle's conditions hold before the 100th patient. Lines provide the expected first patient for which Silvapulle's conditions holds. The x-axis is θ. [Colour figure can be viewed at wileyonlinelibrary.com]*

Table 2. *Existence of MLEs.*

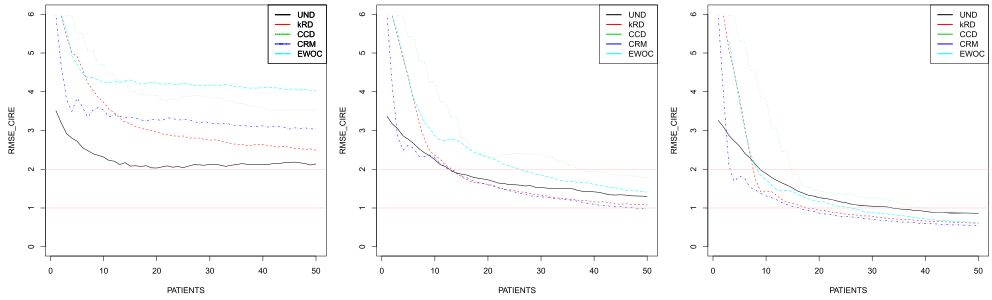
θ	***UND***	2RD	CCD	CRM	EWOC
0.01	6.4	10.2	28.9	12.6	16.3
	(3.3)	(4.7)	(17.6)	(10.9)	(14.2)
2.01	6.8	11.1	32.0	11.6	17.4
	(3.9)	(4.3)	(16.6)	(8.4)	(11.2)
8.01	12.9	14.1	36.3	11.9	16.6
	(9.0)	(4.9)	(16.7)	(7.5)	(7.4)
12.01	20.0	15.8	38.2	13.7	19.3
	(13.7)	(5.6)	(17.1)	(10.0)	(10.4)
16.01	27.7	16.8	39.8	16.1	21.6
	(19.4)	(6.7)	(17.9)	(12.1)	(12.2)
40.01	54.7	35.2	55.0	37.5	40.9
	(25.8)	(19.9)	(20.9)	(23.4)	(22.3)

Mean (SD) of the first patient for whom Silvapulle's conditions hold.



(a) Valid MLE Frequencies. **(b)** Valid CIRE Frequencies.

Figure 7. *Expected first patient for whom at least 90% of trials provide a valid estimator in trials of size 100. [Colour figure can be viewed at wileyonlinelibrary.com]*



(a) $\theta = 0.01$: flat toxicity response. **(b)** $\theta = 4$: intermediate toxicity response. **(c)** $\theta = 8$: steep toxicity response.

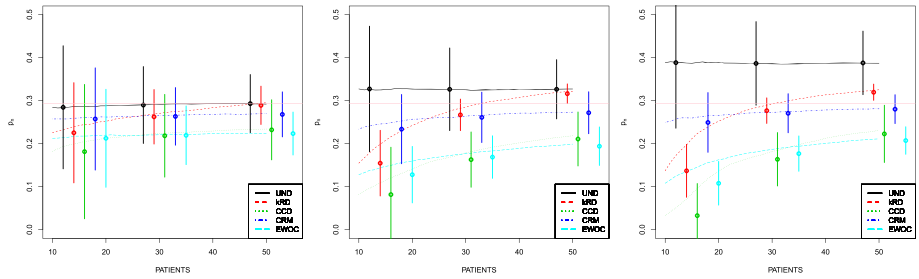
Figure 8. *RMSE of the CIRE for sample sizes up to 50 and increasing slopes in the target of the toxicity response. [Colour figure can be viewed at wileyonlinelibrary.com]*

of patients required on-study before MLEs exist grows, as does its standard deviation. In addition, the proportion of runs with MLE existing when the experiment ends at $n = 100$ decreases as θ increases. This was expected because, as the slope grows, the variation in toxicity rates between consecutive doses also grows, and then observed non-toxicities and toxicities tend to separate on the dose scale and Silvapulle's conditions are more likely to fail.

Even if existing, MLEs may not be obtained because the optimisation algorithm prescribes a dose out of range or it fails. Several alternative procedures for estimating the logistic regression parameters have been considered in the literature and implemented in R (R Core Team, 2019), for instance, *glm2* based on Marschner (2011), the penalised likelihood algorithm (Firth, 1993), exact logistic regression (Mehta & Patel, 1995) and Markov Chain Monte Carlo methods for Bayesian analysis (Hamra *et al.*, 2013). We have carried out a simulation study to find out if one of them was more effective than R's basic *glm* command in obtaining the MLE. Even though these procedures mitigate the problem of convergence, they do not solve it completely. In fact, we have confirmed that it is not unusual to obtain estimates out of the prescribed dose range when the slope of the logistic generating function is small, regardless the procedure used for estimation. We adopted R's *glm2* command for maximising the likelihood. The failure of estimates to exist provides a critical warning to investigators that follow-up studies should be performed with more doses in range where separation, or near separation, occurred. Figure 7a shows the combined effect of all three types of failure. Observe that for very small and high slopes, a large number of patients is required to have at least 90% of probability of obtaining a valid MLE. For valid MLEs, one requires the smallest number of subjects on-study with 2RD for any slope, followed by the CRM; and both are quite competitive when $2 < \theta < 12$.

(b) Calculating the CIRE is only problematic when slopes are very small. In Figure 7b, observe that using UND, CRM and 2RD, the CIRE is obtained with probability at least 90% with less than 20 patients for slopes $\theta \geq 2$. Besides this, one expects to obtain the CIRE with the CRM and UND with fewer patients than with the other designs. This is because the only condition required to successfully calculate the CIRE is that the isotonic regressors (Section 1.5) cross the target toxicity rate. Only designs that escalate dose assignments without restrictions are likely to fulfil this condition early in the experiment.

(c) The RMSE(CIRE) decreases slowly after 30 patients on-study, but at different levels depending on the slope and the design. Two horizontal lines at $y = 1$ and $y = 2$ are plotted in Figure 8 for reference. Observe that, as the number of patients grows, the RMSE(CIRE) stabilises under 1 for large slopes ($\theta > 8$); RMSE(CIRE) stabilises in the range (1, 2) for



(a) $\theta = 0.01$: small slope. **(b)** $\theta = 4$: intermediate slope. **(c)** $\theta = 8$: large slope.

Figure 9. *Overall toxicity rates* $[p_n \pm \sigma(p_n)]$ *for sample sizes up to 50 from logistic generating functions with small, intermediate and large slopes at the target dose. [Colour figure can be viewed at wileyonlinelibrary.com]*

intermediate slopes; and over 2 for small slopes. Something similar happens with the RMSE for the MLE when it is available (figures not shown). The rate of change in the RMSE(CIRE) decreases from 30 patients onwards.

Designs' performance rankings change with the slope. For small slopes, EWOC has the poorest inferential performance, while UND has the best followed by the 2RD which outperforms the CRM near the 10th patient. For intermediate slopes, EWOC's performance improves with RMSE(CIRE) values similar to CCD's. Other differences between UND, 2RD and CRM are negligible as the number of patients grows. Finally, for large slopes, the UND's performance deteriorates, while the 2RD and CRM have similar trajectories, outperforming the EWOC and CCD.

(d) Variability in the overall observed toxicity rate decreases as the slope grows, for each n. The expected toxicity rate changes little from 30 patients upwards, but its value depends on the design. 2RD has an expected toxicity rate slightly over target, but it is least variable and hence most predictable. Figure 9 includes a horizontal line at the target toxicity rate for reference. As expected, the UND produces the highest overall toxicity rate. Also, regardless of slope, variability in overall toxicity rates slowly decreases as the number of patients grows and ordering of designs remains constant, except for the 2RD whose toxicity rate is less than that of the CRM for small samples and gets larger at about $n = 30$. Observed overall toxicity rates with the EWOC and CCD remain far below the target, whereas the 2RD and CRM stabilise slightly over and under the target, respectively. Observe that the 2RD has the smallest variability for all examined slopes, indicating that it is more likely to perform as expected than the others. This coincides with the comments on Figure 3.

(e) Penalising for overdosing does not change performance rankings. The g_n measure behaves in the same way as the p_n measure, so they provide the same ranking of designs (figures not shown).

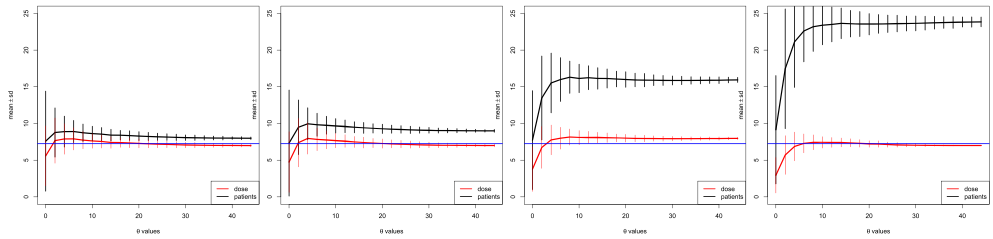
3.5 The Influence of a Start-Up Rule

In this section, we evaluate the use of several start-up rules before the primary design is applied. In addition, the interaction between the start-up rule and the slope is considered. Let DSU denote the dose where the start-up rule finishes and let NSU denote number of patients required to complete the start-up rule. The performance of the start-up rules are evaluated according to the following criteria:

(a) The closer DSU is to the target dose the better, because it mitigates the use of inefficacious doses. DSU close to the target provides a good starting dose for kRD and CCD and improves the prior for the CRM and EWOC.
(b) The smaller NSU is the better.
(c) A start-up rule that is more likely than another to provide valid MLE and CIRE calculations is better.

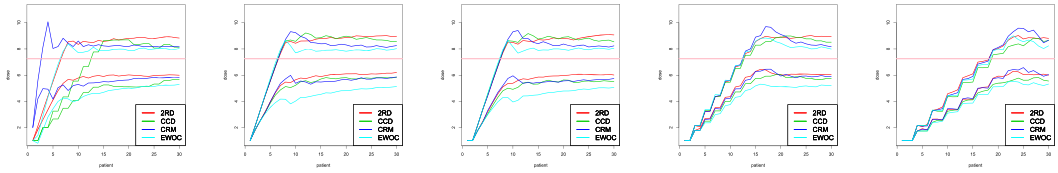
3.5.1 Headlines from inspecting Figures 10 and 11 and Table 3

Figure 10 plots mean values of DSU (black) and NSU (red) with $\pm$ standard deviation bars by slope for each start-up rule described in Section 1.4. The same scale is used for both DSU and NSU on the y-axis. The blue horizontal line marks the target dose as a reference for the DSU values. Table 3 provides a numerical summary of these graphs and includes two start-up rules not plotted. Figure 11 contains five plots, one for each start-up rule described in Section 1.4 plus one where a start-up rule is not applied. For each, interval bounds (mean $\pm$ standard deviations) for the dose assigned to the nth patient are plotted by $n = 1, \ldots, 30$ for each primary design.



(a) ET1 or 1RD start up rule. **(b)** 2R start up rule. **(c)** ET2 start up rule. **(d)** 3+3 start up rule.

Figure 10. *Mean number of patients required to complete the start-up rule (NSU) and the dose where the start-up rule finishes (DSU) ±standard deviation bars for selected θ from 0 to 46^o at the target dose (horizontal line). [Colour figure can be viewed at wileyonlinelibrary.com]*



(a) No start up. **(b)** ET1-1R start up. **(c)** 2R start up. **(d)** ET2 start up. **(e)** 3+3 start up.

Figure 11. *Mean ± standard deviation intervals for early dose allocations. [Colour figure can be viewed at wileyonlinelibrary.com]*

Table 3. *Mean and standard deviation of number of patients required to complete the start-up rule (NSU) and the dose where the start-up rule finishes (DSU) for selected slopes and start-up rules.*

θ	ET1/1R	ET2	2R	ET3	3R	3+3
	NSU					
1.01	4.2 (2.9)	5.0 (3.3)	4.3 (3.3)	6.0 (4.0)	4.5 (3.3)	5.9 (3.6)
2.01	4.9 (3.0)	6.2 (4.0)	5.4 (3.3)	5.8 (3.6)	4.7 (3.4)	7.3 (4.6)
4.01	6.3(2.7)	9.2 (4.6)	6.8 (3.0)	8.1(5.1)	6.1(3.5)	11.5 (5.9)
8.01	7.4 (2.0)	12.5 (3.8)	8.3 (3.0)	17.0 (5.3)	9.4 (2.1)	17.1 (5.2)
12.01	7.8 (1.4)	14.2 (2.6)	8.9 (1.3)	19.8 (4.1)	9.8 (1.4)	20.0 (4.0)
24.01	8.0 (0.8)	15.2 (1.4)	9.1 (0.8)	21.3 (2.7)	10.0 (1.0)	22.1 (2.2)
	DSU					
1.01	3.3 (2.6)	2.5 (1.7)	2.8 (2.7)	2.0 (1.3)	2.5 (2.5)	1.8 (1.2)
2.01	4.0 (2.8)	3.1 (2.0)	3.8 (2.9)	1.9 (1.2)	2.7 (2.6)	2.3 (1.5)
4.01	5.4 (2.7)	4.6 (2.3)	5.0 (2.8)	2.7 (1.7)	3.7 (2.9)	3.7 (1.9)
8.01	6.4 (2.0)	6.3 (1.9)	6.4 (2.0)	5.7 (1.7)	6.4 (2.0)	5.5 (1.7)
12.01	6.8 (1.4)	7.1 (1.3)	6.9 (1.3)	6.6 (1.4)	6.8 (1.4)	6.4 (1.2)
24.01	7.0 (0.8)	7.6 (0.7)	7.1 (0.8)	7.0 (1.0)	7.0 (1.1)	6.9 (0.6)

Bounds were obtained using the basic simulation setup, but each design was warmed up with a start-up rule.

(a) The slope strongly influences the behavior of the start-up rule. Figure 10 and Table 3 show that the NSU grows as the slope increases for all start-up rules, while the DSU gets closer to the target dose. The variability of both the DSU and the NSU decrease with the slope.

(b) The 1R/ET1 and 2R have the best properties. With ET1 and 2RD start-up rules, NSU has the smallest expected value and variability, regardless the slope. Moreover, ET1 and 2RD start-up rules have expected DSUs closest to the target. The variability of DSU is quite similar for all the start-up rules. It may be significant that ET1 and 2R allocate doses to patients one at a time, whereas the other designs use cohorts of two or three patients.
(c) The frequency of obtaining a valid CIRE improves using ET1 or 2R start-up rules with CCD and EWOC primary designs. When using the ET1 or 2R as start-up rules, the CCD and EWOC require less patients to obtain the CIRE with a 90% probability. But start-up rules do not improve this probability for the CRM and 2RD primary designs (figure not shown).
(d) The CCD benefits from the use of a start-up rule by reducing the assignment of patients to very low doses. Figure 11a was produced without a start-up procedure. In this graph, primary designs start, with no variability, at the first dose except with the CRM which always starts at the second dose. Allocation intervals become centred around the target dose (pink horizontal line) fastest with the CRM, but the CRM's interval width has the highest early variability. Allocation intervals with the 2RD and EWOC primary designs centre around the target with a few more patients but also with smaller variability. Observe also that, as outlined before, ET1 and 2R start-up designs produce similar allocation behavior in the primary designs; with the 3+3 start-up, more patients are required for intervals to bound the target. Using the CCD without a start-up, 15 patients are required to centre around the target dose; the use of the ET1 or 2R start-up rules saves 10 patients. When we repeat the study of Sections 3.1 and 3.2 with the ET1 as start-up rule, the performance of CCD improves, but its rankings relative to the other designs does not change. Nevertheless, trials with less than 15 patients seem unusual in practice; and with at least 15 patients, regardless of the start-up chosen, the allocation intervals have similar ranges for all primary designs.

4 Discussion

This paper demonstrates useful assessment measures and graphical tools for evaluating the global performance of dose-finding designs. Good performance requires accurate estimation with a small number of total toxicities. Specifically, the trade-off between ethics and inference is contrasted among several designs selected from the literature to have a variety of dose allocation features: (i) long and short memory, (ii) parametric model and Markov chain theory driven and (iii) with and without dose changes restricted to nearest neighbours. Assuming a logistic model for inference, comparisons were found to be substantially invariant to the generating dose–response model be it logistic, normal or skew-normal (± 3). Summary simulated statistics were created by randomly selecting the slope of the generating model at the target dose, while slope-specific statistics reveal substantial differences in expected variability of target dose estimates among designs. Finally, a start-up rule is found useful to warm up long-memory designs.

Centred isotonic regression estimators of the target dose are shown to be much more likely to be valid and consistently less variable than maximum likelihood estimators. Reliance on the use of MLEs is not advised with a small number of patients or when high dose–response slopes are anticipated because they are difficult to obtain. MLEs may be useful given they can be obtained (particularly for forming confidence intervals), but attention to the performance of statistical software packages is needed.

Several global conclusions regarding the designs are

(a) among the primary designs, 2RD and CRM are quite competitive and outperform the other designs. 2RD has slightly higher toxicity rates with larger sample sizes, while EWOC's toxicity rates are consistently lowest, but these could be equalised by adding a small random

hold to 2RD dose changes and/or by increasing the EWOC parameter that makes allocations conservative. The 2RD's toxicity rate is least variable, and hence, this design is most predictable, ethically speaking. It also has better inferential properties than the other designs studied, especially using the CIRE;

(b) start-up rules kR, $k = 1, 2$ would warm up the trial nicely, producing fewer patients on inefficacious doses and providing allocations' quick arrival to doses close to the target dose. When the study in Section 3.1 is repeated with 1R and 2R as start-up rules, global comparative conclusions change very slightly, even though CCD's properties improve substantially;

(c) consider sample size 15–50. In this case, a start-up rule will not be helpful (except if using the CCD); and the CIRE will be available with a high probability, especially if using the 2RD or CRM. Using more than 50 patients does not provide substantial gains, either from the ethical or inferential points of view; and

(d) the performance measures RMSE, p_n and $\widetilde{\gamma}_n$ provide complementary information. Mapping RMSEs to the toxicity scale provided no additional illumination , and overall toxicity rates p_n are easier to interpret than other toxicity measures such as g_n. It is useful to examine the efficiency measure $\widetilde{\gamma}_n$ together with toxicity rates because both can be interpreted in terms of numbers of patients.

ACKNOWLEDGEMENTS

Fernando Plo and Jose Moler acknowledge the financial support received from the projects Ministerio de Economía y Competitividad (MTM2014-53340-P) and Secretaría de Estado de Investigación, Desarrollo e Innovación (MTM2016-77015-R). The authors thank the editor and the reviewers for their helpful comments that have greatly improved this paper.

References

Azriel, D., Mandel, M. & Rinnot, Y. (2011). The treatment versus experimentation dilemma in dose finding studies. *J. Stat. Plan. Inference*, **141**, 2759–2768.

Babb, J., Rogatko, A. & Zacks, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.*, **17**, 1103–1120.

Biedermann, S., Dette, H. & Zhu, W. (2006). Optimal designs for dose–response models with restricted design spaces. *J. Am. Stat. Assoc.*, **101**(474), 747–759.

Chang, H. -H. & Ying, Z. (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Ann. Stat.*, **37**(3), 1466–1488.

Cheung, Y. K. (2011). *Dose Finding by the Continual Reassessment Method* Boca Raton: Chapman and Hall/CRC press.

Chevret, S. (ed). (2006). *Statistical Methods for Dose-finding Experiments.* West Sussex: Wiley.

Diniz, M. A., Tighiouart, M. & Rogatko, A. (2019). Comparison between continuous and discrete doses for model based designs in cancer dose finding. *PLOS one*, **14**, e0210139. https://doi.org/10.1371/journal.pone.0210139.

Fedorov, V., Flournoy, N., Wu, Y. & Zhang, R. (2011). Best intention designs in sose-finding studies. Technical Report, Isaac Newton Institute for the Mathematical Sciences. Cambridge, UK http://www.newton.ac.uk/preprints/NI11065.pdf.

Fedorov, V. V. & Leonov, S. L. (2014). *Optimal Design for Nonlinear Response Models.* Boca Raton: CRC Press, Taylor and Francis Group.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Ford, I., Torsney, B. & Wu, C. F. J. (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *JRSS-B*, **54**, 569–583.

Hamra, G., Maclehose, R. & Richardson, D. (2013). Markov Chain Monte Carlo: An introduction to epidemiologists. *Int. J. Epidemiology*, **42**, 627–634.

Hansen, A. R., Graham, D. M., Pond, G. R. & Siu, L. L. (2014). Phase 1 trial design: Is 3+3 the best? *Cancer control*, **21**(3), 200–208.

Heinze, G. & Ploner, M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Comput. Methods Progr. Biomed.*, **71**, 181–187.
Ivanova, A. (2006). Escalation, group and A+B designs for dose-finding trials. *Stat. Med.*, **25**, 3668–3678.
Ivanova, A., Flournoy, N. & Chung, Y. (2007). Cumulative cohort design for dose finding. *J. Stat. Plan. Inference*, **137**, 2316–2327.
Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G. & Durham, S. D. (2003). Improved up-and-down designs for phase I trials. *Stat. Med.*, **22**, 69–82.
James, G., Symeonides, S., Marshall, J., Young, J. & Clack, G. (2016). Continual reassessment method for dose escalation clinical trials in oncology: A comparison of prior skeleton approaches using AZD3514 data. *BMC Cancer*, **16**, 703.
Lagoda, T. & Sonsino, C. (2004). Comparison of different methods for presenting variable amplitude loading fatigue results. *Mat. Wissen Werkstoff*, **35**(1), 13–20.
Liu, S., Cai, C. & Ning, J. (2013). Up-and-down designs for phase I clinical trials. *Contemp. Clin. Trials*, **36**(1), 218–227.
Marschner, I. C. (2011). glm2: Fitting generalized linear models with convergence problems. *R J.*, **3**(2), 12–15.
Mehta, C. & Patel, N. (1995). Exact logistic regression: Theory and examples. *Stat. Med.*, **14**, 2143–2160.
Minkin, S. (1987). Optimal designs for binary data. *J. Am. Stat. Assoc.*, **82**(400), 1098–1103.
O'Quigley, J., Pepe, M. & Fisher, L. (1990). Continual reassessment method: A practica design for phase I clinical studies in cancer. *Biometrics*, **46**(1), 33–48.
Oron, A., Azriel, D. & Hoff, P. (2011). Dose-finding designs: The role of convergence properties. *Int. J. Biostat.*, **7**(1), 1–17.
Oron, A. & Flournoy, N. (2017). Centered isotonic regression: point and interval estimation for dose–response studies. *Stat. Biopharmaceutical Res.*, **9**, 258–267. https://doi.org/10.1080/19466315.2017.1286256.
Oron, A. P. & Hoff, P. D. (2009). The k-in-a-row up-and-down design, revisited. *Stat. Med.*, **28**, 1805–1820.
Oron, A. & Hoff, P. (2013). Small-sample behavior of novel phase I cancer trial designs. *Clin. Trials*, **10**, 63–80.
R Core Team. (2019). *R: A Language and Environment for Statistical Computing* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Reiner, E., Paoletti, X. & O'Quigley, J. (1999). Operating characteristics of the standard phase I clinical trial design. *Comput. Stat. Data Anal.*, **30**, 303–315.
Robertson, T., Wright, F. T. & Dykstra, R. L. (1988). *Order Restricted Statistical Inference.* New York: Wiley.
Shen, L. & O'Quigley, J. (1996). Consistency of continual reassessment method under model misspecification. *Biometrika*, **83**, 395–405.
Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. R. Stat. Soc. Ser. B (Methodological).*, **43**(3), 310–313.
Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, **45**(3), 925–937.
Stylianou, M. & Flournoy, N. (2002). Dose finding using the biased coin up-and-down design and isotonic regression. *Biometrics*, **58**, 171–177.
Sverdlov, O., Wong, W. K. & Ryeznik, Y. (2014). Adaptive clinical trial designs for phase I cancer studies. *Stat. Surveys*, **8**, 2–44.
Sweeting, M., Mander, A. & Sabin, T. (2013). Bcrm: Bayesian continual reassessment method designs for phase I dose-finding trials. *J. Stat. Software*, **54**(13), 1–26.
Ting, N. (ed). (2006). *Dose Finding in Drug Development* New York: Springer.
Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *J. R. Stat. Soc. Ser. B*, **25**(1), 1–48.
Wetherill, G. B., Chen, H. & Vasudeva, R. B. (1966). Sequential estimation of quantal response curves: A new method of estimation. *Biometrika*, **53**, 439–454.
Wetherill, G. B. & Levitt, H. (1966). Sequential estimation of on a psychometric function. *British J. Math. Stat. Psychology*, **18**, 1–10.
Zera, J. (2017). Application of Wald sequential test in staircase up-down adaptive procedures. *J. Acoustical Soc. Am.*, **141**, 4028–4028.

[*Received February 2019, accepted February 2020*]

Appendix A

In order to summarise the variety of designs and measures, we include the following scheme which can help the reader to visualise the simulation study.

Table A1. *Simulation scenarios.*

Start-up rule	Primary designs	Dose–response models	Slopes
• ETk	• kRD	• Logistic	• Small ($\theta < 2$)
• kR	• CCD	• Normal	• Intermediate ($2 < \theta < 8$)
• 3+3	• CRM	• Skew-normal (-3)	• Large($\theta > 8$)
• None	• EWOC	• Skew-normal (+3)	• Random ($\theta \sim Unif(0, 35^o)$)
	• UND		

Each simulation setup combine one item in each column.

Table A2. *Performance measures.*

Inferential for ML and CIR estimators	Ethical	Start-up rule
• RMSE		
• Squared Bias	• Overall toxicity p_n	• DSU
• Bias		
• Variance	• Allocation g_n	• NSU
• Coverage		
• D-optimal efficiency		