Small-sample behavior of novel phase I cancer trial designs

Assaf P Oron\textsuperscript{a,b} and Peter D Hoff\textsuperscript{c}

Background  Novel dose-finding designs for Phase I cancer clinical trials, using estimation to assign the best estimated Maximum Tolerated Dose (MTD) at each point in the experiment, most prominently via Bayesian techniques, have been widely discussed and promoted since 1990.

Purpose  To examine the small-sample behavior of these ‘Bayesian Phase I’ designs, and also of non-Bayesian designs sharing the same main ‘Long-Memory’ traits of using likelihood estimation and assigning the estimated MTD to the next patient.

Methods  Data from several recently published experiments are presented and discussed, and Long-Memory designs’ operating principles are explained. Simulation studies compare the small-sample behavior of Long-Memory designs with short-memory ‘Up-and-Down’ designs.

Results  In simulation, Long-Memory and Up-and-Down designs achieved similar success rates in finding the MTD. However, for all Long-Memory designs examined, the number \( n \) of cohorts treated at the true MTD was highly variable between simulated experiments drawn from the same toxicity-threshold distribution. Further investigation using the same set of thresholds in permuted order indicates that this Long-Memory behavior is driven by sensitivity to the order in which participants enter the experiment. This sensitivity is related to Long-Memory designs’ ‘winner-takes-all’ dose-assignment rule, which grants the early cohorts a disproportionately large influence, and causes many experiments to settle early on a specific dose. Additionally for the Bayesian Long-Memory designs, the prior-predictive distribution over the dose levels has a substantial impact upon MTD-finding performance, long into the experiment.

Limitations  While the numerical evidence for Long-Memory designs’ order sensitivity is broad, and plausible explanations for it are provided, we do not present a theoretical proof of the phenomenon.

Conclusions  Method developers, analysts, and practitioners should be aware of Long-Memory designs’ order sensitivity and related phenomena. In particular, they should be informed that settling on a single dose does not guarantee that this dose is the MTD. Presently, Up-and-Down designs offer a simpler and more robust alternative for the sample sizes of 10–40 patients used in most Phase I trials. Future designs might benefit from combining the two approaches. We also suggest that the field’s paradigm change from dose-selection to dose-estimation.

Clinical Trials 2013; 10: 63–80. http://ctj.sagepub.com

\textsuperscript{a}Core for Biomedical Statistics, Seattle Children’s Research Institute, Seattle, WA, USA, \textsuperscript{b}Professional and Continuing Education, University of Washington, Seattle, WA, USA, \textsuperscript{c}Departments of Statistics and Biostatistics, University of Washington, Seattle, WA, USA

Author for correspondence: Assaf P Oron, Core for Biomedical Statistics, Seattle Children’s Research Institute, 2001 Eighth Avenue, Suite 500, Seattle, WA 98121, USA.

Email: assaf.oron@seattlechildrens.org

© The Author(s), 2013

Reprints and permission: http://www.sagepub.co.uk/journalsPermissions.nav 10.1177/1740774512469311
Introduction

Over the past two decades, numerous novel dose-finding designs employing Bayesian calculations have been developed for Phase I cancer trials, most notably the Continual Reassessment Method (CRM) and Escalation with Overdose Control (EWOC) [1,2]. The hallmark of these designs is estimation of the dose–toxicity function after each cohort, in order to assign the estimated Maximum Tolerated Dose (MTD) to the next cohort. These ‘Bayesian Phase I’ designs have been joined by novel non-Bayesian designs employing similar principles [3–5]. We will use the term ‘Long-Memory’ to refer to the family of designs assigning the estimated MTD at each cohort, regardless of whether they employ Bayesian methods.

Despite their popularity among statisticians, Long-Memory designs have struggled to enter actual practice, where the conservative ‘3 + 3’ experimental protocol [6], which has been repeatedly shown to possess poor MTD-selection properties [7–9], still dominates [10]. Ivy et al. [11], on behalf of the Clinical Trial Design Task Force of the National Cancer Institute (NCI)’s Investigational Drug Steering Committee, embraced the new designs, suggesting that ‘... members of the boards may not be convinced that novel designs are better for patients. In fact, they are’.

Even as clinicians turn from skepticism to optimism, the task of constructing a comprehensive picture of Long-Memory properties in theory and practice is far from complete. The available theoretical results on Long-Memory designs are partial, and mostly involve asymptotic behavior. Shen and O’Quigley [12] proved almost-sure convergence of designs based on a one-parameter model closely related to CRM, to an allocation behavior that assigns only the true MTD to patients. This proof has been often misquoted as evidence that CRM’s asymptotic behavior is universally robust to misspecification. However, subsequent theoretical work by several contributors [13–16] has shown that the Shen-O’Quigley proof conditions apply only to a very narrow subset of dose–toxicity curves, and that one-parameter models can only guarantee convergence to some dose level whose toxicity rate is within an ‘indifference interval’ around the target rate. The interval boundaries are a function of design choices [14]. Oron et al. [15] show that a novel non-parametric design called Cumulative Cohort Design (CCD) [5] belongs to the Long-Memory family, and prove that it converges to an interval in a very similar manner to one-parameter models. More generally, Azriel et al. [17] proved that no Long-Memory design can guarantee almost-sure convergence to the MTD on the class of all dose–toxicity functions.

Hence, it appears that with Long-Memory designs, one must settle at best for the interval guarantee, rather than expect convergence to the MTD itself [15]. Both designs mentioned here will be featured later in the article. At present, there is no proof for convergence under misspecification for multiparameter Long-Memory designs.

Much less is definitively known regarding small-sample properties of Long-Memory designs. Despite the relatively small number of published Bayesian Phase I studies, some of these have reported disturbing small-sample behavior, prompting the analysts to develop ad hoc design modifications that might mitigate it [18,19]. Numerical Phase I studies have focused almost exclusively upon ensemble-average statistics. While average values are important, for some outcomes, this is the wrong statistic to focus on because in practice one does not run an ensemble – but rather a single experiment. A case in point is the number of cohorts treated at the MTD, a statistic we shall refer to as $n^*$. The high ensemble-average values of $n^*$ obtained via Long-Memory designs have been repeatedly invoked as a compelling reason for choosing this design approach [10,20]. The study by Iasonos et al. [21], perhaps the only article to date to present a measure of $n^*$’s variability, reports large standard deviations along with these high averages (Ref. [21]; Table 2). In the ‘Numerical demonstration’, we describe the complete distribution of $n^*$ under various scenarios and designs. Long-Memory designs suffer from alarmingly high run-to-run $n^*$ variability, meaning that individual experiments might quite often end with a very low $n^*$ rather than the high value promised by method developers. This between-run variability is related to Long-Memory designs’ overarching feature, namely, the insistence upon treating every cohort with the estimated MTD at any given time. This phenomenon is presented and analyzed in detail in this article.

The article is organized as follows: in ‘Preliminaries’, we define terminology and describe Long-Memory designs’ operating principles; subsequently, detailed examples from published Bayesian Phase I experiments are presented. Following that, we numerically compare CRM with an ‘interval design’ and a short-memory ‘Up-and-Down’ design. A general discussion ends the article.

Preliminaries

Basic terminology

Consider trials carried out as sequential dose-finding experiments with $n$ cohorts, indexed $c, c = 1, \ldots, n$, each cohort comprising of $k_c \geq 1$ subjects. Except for cohort 1, the dose administered to cohort $c$ is
(generally speaking) not known until all observations up to cohort \( c - 1 \) are available. \( Y_c \), the number of dose-limiting toxicities (DLTs) observed in cohort \( c \), can be modeled as a Binomial random variable

\[
Y_c \sim \text{Binomial}(k_c, F(x_c))
\]  

(1)

where \( x_c \) is the dose administered to cohort \( c \), and \( F \) is the true (and unknown) underlying toxicity function, assumed to be a continuous strictly increasing Cumulative Distribution Function (CDF) of the response-triggering dose variable \( x \). In these terms, the experiment’s goal is to find \( F^{-1}(p) \) – the 100th percentile of \( F \). This dose is known as the experiment’s target. In Phase I cancer experiments, \( p \) is usually between 1/5 and 1/3. Doses themselves are restricted to a finite set of levels \( D = \{d_u\}, u = 1, \ldots, l \), with \( l \) usually between 4 and 10. The dose level closest to \( F^{-1}(p) \) will typically be recommended as the MTD for Phase II. Hereafter, we denote the estimated MTD as \( \hat{MTD} \).

A generic Bayesian Phase I design can be described as one where, in order to decide which dose to assign to the next cohort, all hitherto available observations are used to estimate \( F \) via the model

\[
R_u \sim \text{Binomial}(n_u, G(d_u, \theta)), \quad u = 1, \ldots, l
\]  

(2)

where \( n_u \) is the number of available observations at dose \( d_u \), \( R_u \) is the number of those among the \( n_u \) who exhibit toxicities; and \( G \), the model curve, is a CDF belonging to a parametric family \( \mathcal{G} \) indexed by a parameter vector \( \theta \) (which usually has a prior distribution with additional, fixed parameters). Input data to the model can be summarized as the observed toxicity rates at the dose levels

\[
\hat{F}_u = \frac{R_u}{n_u}, \quad u : n_u > 0
\]  

(3)

which are also the sufficient statistics for a nonparametric model of \( F \).

According to Rogatko et al. [10], most Bayesian Phase I experiments published through 2006 had used a one-parameter CRM model, most often of the generic form

\[
G(d_u) = \phi_0^u, \quad \phi_1 < \phi_2 < \cdots < \phi_l, \quad \phi_u \in (0, 1) \forall u, \theta > 0
\]  

(4)

The \( \phi_u \), a sequence of constants supplied by the user, are known as the model’s ‘skeleton’. This model form requires, beside the single data-estimable parameter \( \theta \), the specification of \( l \) fixed parameters defining the skeleton, as well as additional fixed parameters involved in \( \theta \)’s own prior distribution.

After cohort \( c \), Bayesian Phase I designs assign the next dose to \( \hat{MTD} \), via Bayesian posterior estimation of \( G \) at the dose levels and some optimization criterion. The most common criterion is choosing the dose that minimizes \( |G(p) - \hat{F}\rangle \), although variations exist. For example, EWOC assigns to the dose closest to the \( a \) posterior quantile of \( G^{-1}(p) \), with \( a = 0.25 \) most often used [2].

Equation (2) is also applicable to frequentist Long-Memory designs, and even to nonparametric ones. The latter directly use the \( F \), which can be viewed as a special case of \( G \). Therefore, we treat any design that allocates successive cohorts to some variation on \( MTD \) via estimation of a model of the general form (2), as belonging to the Long-Memory family, regardless of whether it employs nonparametric, parametric, or Bayesian methods.

Long-Memory designs were called ‘designs with memory’ by some researchers [23]. We prefer the term ‘Long-Memory’ because of the contrast with short-memory designs that only use recent observations. Fedorov et al. [24] suggest the name ‘Best-Intention Designs’, highlighting the \( MTD \) dose allocation inherent to the Long-Memory paradigm. There exist other Bayesian designs that maintain a long memory but employ different principles for dose assignment [25,26]. Our definition of the term ‘Long-Memory’ excludes these designs, but they will be briefly mentioned in the ‘Discussion’.

**Long-memory designs’ operating principle**

The Long-Memory dose-assignment process is akin to fitting a regression curve, constrained by the model family \( \mathcal{G} \), through the points \( \{\{d_u, \hat{F}_u\}\} \). These points are the ‘X’s in Figure 1, displaying data at the end of a published CRM experiment [27]. The regression is fit by a weighted combination of the prior and the likelihood, which is itself weighted by the number of observations at each dose. The experiment’s goal is finding the dose closest to the place where the true \( F \) crosses the horizontal \( y = p \) dashed line in Figure 1. Long-Memory designs allocate each cohort to the best current candidate dose, according to the fitted \( \hat{G} \) curve. If the observed toxicity rate at that dose increases, the corresponding ‘X’ mark will move higher, pulling \( \hat{G} \) with it and eventually mandating dose de-escalation, and vice versa. This is Long-Memory designs’ basic self-correction mechanism, with the underlying intuition that the empirical proportions \( \hat{F} \) will eventually converge to their true \( F \) values, as indeed has been recently proven for generic sequential dose-finding designs [15].

\[\text{http://ctj.sagepub.com}\]
These two elements – self-correction in the assumed direction of target and consistency of observed toxicity rates – form the ‘engine’ driving Long-Memory designs. These elements are so simple that a model \( G \) for \( F \) is not even needed in order to construct the ‘engine’. For example, interval designs such as CCD [5] have no model. Instead, they mandate dose escalation if \( ^\wedge F \) at the current dose is below some ‘tolerance interval’ around \( p \), and vice versa.

If a model is used, the operating principles dictate the relationship between its slope and experimental trajectories. Shallow model curves will shift the crossing point more dramatically as \( ^\wedge G \) changes. Hence, they are associated with more volatile dose allocations, and vice versa for steep curves. Convex one-parameter \( G \) skeletons, shallow to the left and steep to the right, are rather popular in practice. They are quick to descend but more conservative when escalating. Multiparameter models can fit the observed frequencies more closely, but they might suffer from overfitting, especially early in the experiment.

It is important to note that the model-determined degree of volatility is unrelated to the actual rate of convergence to the MTD. The latter is paced by the convergence rate of \( F \), that is, root-\( n \) [12,15]. This is a very slow rate compared with typical Phase I sample sizes of 10–40 patients. If \( G \) correctly specifies \( F \), then all data are pooled to consistently estimate \( \theta \), providing the fastest possible convergence within the root-\( n \) constraints. In the more likely case of misspecification, this pooling affords little help. Convergence is then constrained by the convergence of the empirical toxicity rates observed around the true MTD. Rather often, Long-Memory convergence to the correct MTD is not guaranteed at all. As mentioned in the ‘Introduction’, some Long-Memory designs can guarantee convergence to an ‘indifference interval’ around \( p \), which might contain several levels – or none at all. Other Long-Memory designs provide no MTD-convergence guarantee under misspecification.

**Experimental examples**

We present in this section four published Bayesian Phase I experiments. Each experiment is accompanied by a figure, in which the left-hand frame describes the experiment’s trajectory – that is, each cohort’s administered dose levels and the number of toxic and nontoxic responses observed for each, arranged in chronological order – and the right-hand frame presents the evolution of posterior model curves. For brevity’s sake, some model details are relegated to Supplement A. Several additional experiments are described in Supplement B.

**Dougherty et al.’s anesthesiology experiment [28]**

This study (Figure 2) was not, strictly speaking, a Phase I trial but rather a CRM design applied to an anesthesiology dose-finding experiment [28]. Instead of toxicity, a positive response indicates pain. The target pain rate was 0.2, and there were 25 patients treated one at a time. Chevret’s [22] one-parameter logistic model was used. There were four levels in this design, with the convex skeleton \( d = (0.1, 0.2, 0.4, 0.8) \). The Goodman et al. [29] constraint, forbidding escalation by more than one level between successive cohorts, was in effect. According to its bottom line, the experiment was an astounding success: 18 of 25 patients were treated at the MTD \( (d_2) \), with a cumulative pain rate of 3 out of 18 – almost as close to target as possible (4 of 18 would have been slightly closer).


**Figure 2.** Experimental trajectory and dose–response curves (right) from the Dougherty et al. [28] experiment. In the left frame, subjects are shown in chronological order plotted against the administered levels; each empty circle represents a single negative (no-pain) response, and each filled circle represents a positive response. In the right frame, final empirical pain-rates (f) are shown in ‘X’ marks, whose size is proportional to the number of observations. The piecewise-linear curves represent posterior-based toxicity estimates, with the number indicating the last subject before the update. The zero-symbol curve is the prior, and the symbols A, B, and C stand for estimates after the 16th, 18th, and 25th subject, respectively. The dashed horizontal line indicates the target response rate, in this case 0.2.

**Pisters et al. [30] and Mathew et al. [31]**

A pair of experiments conducted at the M.D. Anderson Center and published in 2004 targeted \( p = 0.3 \), using a one-parameter ‘power’ model CRM [30,31]. The former followed the single-level increment constraint [29], and had four dose levels with a skeleton nearly identical to Dougherty et al.’s: \( \phi = (0.05, 0.20, 0.40, 0.80) \) (Figure 3, top). After an unplanned single patient at \( d_1 \) and the first three-patient cohort at \( d_2 \) with no DLT’s observed in either, all 23 remaining patients (eight cohorts) were assigned \( d_3 \). The observed DLT rate at this dose (7/23) was the closest possible to the target with 23 observations; not surprisingly \( d_3 \) was the final MTD.

The story was different for the second experiment [31], which neglected to follow the single-escalation constraint. The design called for six-person cohorts and had six levels with a relatively shallow skeleton \( \phi = (0.07, 0.16, 0.30, 0.40, 0.46, 0.53) \), beginning at \( d_3 \) (Figure 3, bottom). After zero toxicities observed on the first cohort, allocation jumped directly to \( d_6 \) – where three out of four toxicities forced the experimenters to cut the cohort short and de-escalate to \( d_4 \). At that level, five toxicities out of six were observed, so the experiment descended back to \( d_1 \), where now three of six experienced DLTs. This dose, with a cumulative toxicity rate of 0.25, was recommended as the MTD but not before half the patients in the study (11 of 22) experienced DLTs. More disturbingly, a recalculation of \( G \) according to the model indicates that the final MTD estimate should have been \( d_2 \), with a posterior \( G = 0.28 \) compared to 0.43 for \( d_3 \) (Figure 3, bottom right, curve marked ‘4’). This level had never been assigned during the experiment. Moreover, \( d_2 \), rather than \( d_3 \), should have been assigned to the last cohort as well (\( G = 0.25 \) and 0.40, respectively; curve marked ‘3’).  

**Neuenschwander et al. [18]**

This experiment began as a one-parameter ‘power’ CRM, with a large number of levels, \( l = 15 \) (Figure 4). The starting dose was \( X_1 = d_1 \), and the single-level escalation restriction was initially in effect. The predictive prior indicated \( MTD = d_{10} \), creating an immediate tension between posterior recommendations and dose-escalation restrictions. After four cohorts with 16 patients, cumulatively, yielded no toxicities, \( MTD \) was \( d_{12} \) and researchers agreed to skip from \( d_4 \) to \( d_7 \). The next two patients both experienced DLT’s, but CRM still recommended jumping from \( d_7 \) to \( d_9 \) rather than de-escalating. This

---

1 We inquired with the consulting statistician to this study, and he could not recall the circumstances surrounding the decisions to overrule \( d_2 \) with \( d_3 \) for the last cohort and for the final estimate.
type of dose-transition recommendation (i.e., escalation following toxicity or vice versa) was called ‘incoherent’ by Cheung [32]. Cheung proved that if the current dose is also the \( MTD \), one-parameter CRM designs cannot make ‘incoherent’ recommendations. However, if \( X_1 \neq MTD \), then ‘incoherence’ might be encountered during the experiment’s initial phase until the gap between the two is closed – which is what happened to the experiment in question.

After the ‘incoherent’ recommendation, the trial was put on hold, and intensive simulation and theory work was carried out in order to modify the design [18]. The authors replaced the one-parameter model with a two-parameter logistic, and modified the decision rule to penalize toxicity more heavily. These changes resulted in \( d_6 \) (i.e., a one-level de-escalation) being recommended for the trial’s continuation. All three remaining cohorts were administered.

\[ \text{This Phase I-specific use of the term ‘coherence’ can be applied to any Phase I experiment, not just those using Bayesian Phase I designs. It should not be confused with the term ‘coherence’ as used in Bayesian literature [33].} \]
that dose, which eventually became the final MTD with two toxicities observed on nine patients. Interestingly, even at the experiment’s end, the original one-parameter model still estimated $d_{MTD} = d_8$, rather than $d_6$ (Figure 4, right, curve marked ‘8’).

**Numerical demonstrations**

**Overview and methods**

In this section, we numerically examine some aspects of Long-Memory behavior, compared with short-memory designs belonging to the Up-and-Down family [34]. The simulated sample size was $n = 32$ subjects in cohorts of 2 on $l = 6$ dose levels, for all designs. Several 6-tuples of $F$ values on $D$, hereafter called scenarios, were chosen using different parametric curve forms for $F$, hence the scenario names as they appear later. For each scenario, $M = 1000$ sets of (pseudo-)random toxicity thresholds, representing patients with different toxicity sensitivities, were drawn from that scenario’s exact parametric model for $F$. For each design, each of the $M$ simulated experiments (hereafter: runs) compared each cohort’s assigned doses to its toxicity thresholds, and subsequently applied the design’s transition rules. The group of $M$ runs drawn from the same distribution is known as an ensemble. All designs were tested under the exact same conditions, encountering the same ensemble of toxicity thresholds. The experiment’s formal target was $p = 0.3$ throughout the simulation. We show results from six scenarios, calibrated so that each of the six dose levels was the true MTD for one scenario.

Up-and-Down, a family of designs routinely used in a wide variety of scientific and engineering fields, is often conflated in Phase I methodological articles (e.g., Refs [10] and [18]) with the 3+3 protocol. However, the two diverge in several important respects – first and foremost, the fact that Up-and-Down is a statistical design family with clear theoretical properties, while 3+3 is not. See Supplement C for a more detailed list of differences. Up-and-Down designs generate Markov chains over the dose space $D$, with asymptotic visit frequencies peaking near $F^{-1}(p)$ [35,36]. Their convergence rate to this asymptotic behavior is geometric. Recent methodological work on Up-and-Down designs has explored its properties as a nonparametric design family, and developed novel variations, extensions, and estimation methods [37–41]. However, the overall number of recent methodological Up-and-Down publications is at least an order of magnitude smaller than analogous work on Long-Memory designs.

For the simulations illustrated here, we used a group Up-and-Down design [36,42]. The allocation rule is

1. Start at a predesignated dose, and treat cohorts of size 2
2. After cohort $c$ is treated at $X_c = d_u$, set $X_{c+1}$ to
   \[
   \begin{cases}
   d_{u+1} & \text{If 0 of 2 are toxicities} \\
   d_{u-1} & \text{otherwise}
   \end{cases}
   \]

Boundary conditions replace nonexistent levels ($d_0$ or $d_{l+1}$) with the existing levels $d_1$ and $d_l$. 

![Figure 4. Trajectory (left) and posterior model curves (right) of the Neuenschwander et al. [18] experiment. The dashed line after cohort 5 in the left-hand frame indicates the original allocation to cohort 6 using the one-parameter model.](http://ctj.sagepub.com/clinicaltrials.2013.10.63-80)
respective, whenever the former are mandated. The design converges to an asymptotic allocation distribution peaked near $F^{-1}(0.29)$.

For CRM, the simulation used the ‘power’ model with a skeleton similar to that of Flinn et al. (Figure 1). Those authors’ skeleton was $\phi = (0.05, 0.10, 0.20, 0.30, 0.50, 0.65, 0.80)$ with $l=7$, and ours is $\phi = (0.05, 0.11, 0.22, 0.40, 0.60, 0.78)$ with $l=6$. The prior on $\theta$ was log-Normal, the one most commonly used in practice, and was calibrated so that initial responses are ‘coherent’ in the sense defined in the description of Neuenschwander et al. [18]. The single-level escalation constraint was universally used, in both the upward and downward directions.

Another Long-Memory design examined here is Ivanova et al.’s [5] nonparametric CCD, mentioned earlier in the article. CCD repeats the same dose $d_n$ as long as $F_n$ falls inside a tolerance interval around $p$, escalates if $F_n$ is below the interval, and vice versa. Here, we used the interval $(0.2, 0.4)$, recommended in Ref. [5] for $l=6$.

All runs started at $d_2$. The code and subsequent analysis were implemented in R [43]. Further simulation details are presented in Supplement D.

Between-run variability and the order effect

Dose-finding simulation summaries are usually statistics of average ensemble performance, for example, $\overline{M}$ – the ensemble average number of cohorts per run that were administered the true MTD. Many Long-Memory designs tend to perform well on this statistic, which also happens to be one of Up-and-Down's weakest aspects, being a random-walk design that inevitably spreads allocations over several levels around $F^{-1}(p)$.

Rather than just report the average, Figure 5 displays the ensemble distribution of $n^*$ (excluding the arbitrary first cohort), enabling a glimpse into run-to-run variability. The ensemble average $\overline{M}$ is visible as the bold vertical line in the middle of each histogram. One-parameter CRM (left) is compared here with group Up-and-Down (right). Three of the scenarios are shown in Figure 5 (top to bottom).

The most dramatic feature in Figure 5 is CRM's between-run variability. Even under the Normal scenario (top left), where the ensemble mode is at a spectacular 11 MTD-allocated cohorts out of 15 and the average is around 8.5 cohorts, 14% of the CRM runs ended with $n^* \leq 2$ – meaning that fewer patients were treated at the true MTD than would have under a fixed uniform allocation rule across the dose levels. Under the Gamma scenario (middle left), CRM's most common $n^*$ outcome allocates zero cohorts to the MTD during the experiment. It should be noted that for the Gamma scenario, the MTD was actually the starting dose ($d_2$), meaning that in one-fifth of the runs, CRM immediately veered away from its starting dose, never to return – despite $d_2$ being the correct MTD. Finally, the log-Normal scenario (bottom left) generates strongly divergent behavior, with very low or very high values of $n^*$ more common than intermediate outcomes.

With Up-and-Down (Figure 5, right-hand frames), between-run and between-scenario differences are far smaller. Due to its random-walk nature, group Up-and-Down cannot allocate more than roughly half the cohorts to any single level except on the boundary. However, in all scenarios, the modal outcome is reasonably close to this limit at five to six cohorts per run, with the vast majority of runs producing $n^*$ values within $\pm 2$ of the mode.

CRM's $n^*$ variability indicates a sensitivity to variations in the input data. Sensitivity to model fit definitely exists, as can be deduced from the differences between scenarios. However, this does not explain the considerable within-scenario sensitivity between different runs since these runs were all drawn from the same $F$. Two remaining sources of variability are as follows:

1. Each run's empirical sample moments. For example, if a single run's set of 32 thresholds is uncharacteristically high (low) on the average, the run will likely have a smaller (larger) number of toxic responses, respectively, unless the design self-corrects upward (downward). This is analogous to two groups of patients who by random chance are on the average more or less liable to respond toxicity to the treatment, despite belonging to the same patient population.

2. The order in which toxicity thresholds are encountered during the run. For example, would a group of low thresholds followed by a group of high ones generate different behavior than if the two groups' order was reversed?

To help pinpoint which of the two sources is more responsible for CRM's $n^*$ variability, we replaced the randomly generated thresholds with fixed sets. For each scenario, a 'perfect set' consisting of the percentiles $F^{-1}(1/33), \ldots, F^{-1}(32/33)$, was slightly modified by ‘knocking out’ two thresholds in the vicinity of $F^{-1}(p)$, one on each side, and replacing them with replicas of $F^{-1}(1/33)$ and $F^{-1}(32/33)$, respectively (the original 'perfect set' would be unrealistically well behaved). We then generated $M=1000$ runs from each scenario, each run using the exact same set of thresholds, but with the order in which they appear randomly permuted. Figure 6 shows the distributions of $n^*$ from these runs; it is impressively similar to Figure 5. This establishes that CRM's run-to-run variability in $n^*$ is driven primarily by variations in sampling order, that is, the order in which participants enter the experiment.
Variability in \( n^r \) and sensitivity to sampling order are properties of all Long-Memory designs, not just Bayesian ones. Figure 7 repeats the same exercise of Figures 5 and 6 (pseudorandom draws, then permutations of a fixed threshold set), using the nonparametric CCD for dose allocations. Between-scenario variability in \( n^r \) is smaller than with CRM, but between-run variability within each scenario is as great or greater.

**MTD-selection performance and effect of prior**

Table 1 presents the percent of runs in which the MTD was correctly selected for the three methods under each scenario, halfway through the experiment (left three columns) and at its end (right three columns). We chose scenarios for which the MTD is unambiguous: its true \( F \) value is always very close to 0.3, and the \( F \) values of neighboring levels are no closer than approximately 0.2 or 0.4 (see details in Supplement D). For Up-and-Down and CCD estimation, we used a variant of isotonic regression, recently recommended as a robust and reasonably efficient nonparametric estimator [39]. If the \( F \) are monotone increasing, then this estimator simply chooses the level whose \( F \) value is closest to 0.3 as the MTD. If there is a monotonicity violation, the violating values are replaced by a single weighted average, and the MTD is determined by linear

---

**Figure 5.** Between-run and between-scenario variability. The histograms depict the ensemble distribution of \( n^r \), excluding the first cohort. The ensemble size is 1000 runs. Scenarios are Normal (top), Gamma (middle), and Lognormal (bottom); designs are CRM one-parameter ‘power’ (left) and Group Up-and-Down (right). The bold vertical lines in each histogram pinpoint the ensemble average, \( \overline{F} \).

---

CRM: Continual Reassessment Method; MTD: Maximum Tolerated Dose; U&D: Up-and-Down.
Figure 6. Similar to Figure 5, except that rather than draws out of a simulated distribution, the runs are order permutations of the same set of 32 thresholds, as described in section 'Between-Run Variability and the Order Effect'.

CRM: Continual Reassessment Method; MTD: Maximum Tolerated Dose; GU&D: Group Up-and-Down.

Table 1. Bulk performance comparison between ‘power’ CRM, CCD, and group Up-and-Down (abbreviated as U&D in column headings). For each of the six scenarios, the table shows the proportion of runs in which the correct MTD was found, after 8 (left) and 16 (right) cohorts, respectively. CRM is estimated as the next dose allocation; Up-and-Down and CCD were estimated using centered isotonic regression [44]. All numbers are in percent.

| Scenario     | MTD level | After 8 cohorts | After 16 cohorts |
|--------------|-----------|-----------------|-----------------|
|              | CRM | CCD | CRM | CCD | CRM | CCD | U&D | U&D |
| ‘Uniform’    | 1   | 50.2 | 57.1 | 54.0 | 62.1 | 64.1 | 60.8 |
| ‘Gamma’      | 2   | 36.6 | 44.2 | 40.8 | 47.4 | 53.2 | 51.2 |
| ‘Normal’     | 3   | 57.8 | 54.2 | 56.4 | 67.5 | 67.1 | 63.0 |
| ‘Lognormal’  | 4   | 46.7 | 34.0 | 33.0 | 59.3 | 46.2 | 48.4 |
| ‘Weibull’    | 5   | 39.0 | 28.1 | 38.1 | 47.2 | 42.6 | 45.0 |
| ‘Logistic’   | 6   | 26.0 | 30.0 | 32.2 | 29.3 | 48.5 | 54.6 |

CCD: Cumulative Cohort Design; CRM: Continual Reassessment Method.
interpolation between the new values [44]. The code used for isotonic-regression estimation is available in Supplement F.

In Table 1, we used a visual convention of emphasizing only outcomes for which the design in question is substantially different (by 5% or more) from the others. Stronger results are in boldface, and weaker results are in italics. Overall, the performance differences between these three very different designs are remarkably small. It is CRM that falls most conspicuously behind under the 'Logistic' scenario (bottom row), in which it shows nearly no improvement during the experiment’s second half.

Table 2 summarizes the MTD-selection performance of the same CRM skeleton with three different sets of prior parameters for θ (stronger results are in boldface, and weaker results are in italics). The prior used to produce Figures 5 and 6 and Table 1 is labeled ‘A’. It represents a modest amount of scientific knowledge and priorities: it assumes that the middle of the dose range is somewhat more likely to contain the MTD, and the highest doses are less likely or desirable than the lowest ones (Table 2, left column). Prior B, which encourages dose escalation (e.g., $d_5$ has more prior-predictive weight than $d_2$ or $d_3$), is commonly recommended by CRM researchers as ‘uninformative’. It is the default prior in the ‘crm’ function of the R package ‘dfcrm’ described in Cheung’s recent book [45]. Prior C reflects a strong belief that the MTD is in the lower half of the dose range, or (equivalently) a reluctance to prefer higher doses until overwhelming evidence has accumulated. All priors used the log-Normal distribution.

Overall, the performance variability when using the same CRM model and the same prior distribution family, but with different prior parameters, is as great or greater than the variability between

---

**Figure 7.** Distribution of $n^*$ using the exact same random draws of Figure 5 (left) and Figure 6 (right), under the nonparametric interval design CCD [5].

MTD: Maximum Tolerated Dose; CCD: Cumulative Cohort Design.

---
completely different methods seen in Table 1. Furthermore, the relative performance in Table 2 closely mirrors the MTD’s relative prior-predictive weight under each prior, or more precisely, each level’s predictive weight compared with its immediate neighbors. The performance improvement from 16 to 32 subjects is around 10%–15% in most scenarios regardless of prior; however, it is substantially slower under the Weibull and Logistic scenarios – the two scenarios for which this model fails to converge to the MTD.

'Settling' and estimation success

The phenomenon of Long-Memory experiments settling fairly early on a single dose is well known; see, for example, the first two experiments in 'Experimental examples'. O’Quigley [46], Rogatko et al. [10], and many others view it as a strength. The rationale is that such a settling indicates the Long-Memory self-correction mechanism needs little further information to determine the MTD. No Long-Memory convergence proof uses any mechanism faster than root-\( n \), and therefore, it is hard to argue that model estimates based on a sample of 10–40 binary observations are already guaranteed to be close to their true asymptotic values. However, the question remains whether early settling in a Long-Memory experiment is an encouraging sign for good outcomes, or not.

Here, we consider a run to have ‘settled’ once the same dose has been assigned five consecutive times (excluding the arbitrary starting dose). Some Long-Memory studies had suggested a similar settling criterion as a stopping rule [47]. More sophisticated stopping-rule approaches, such as calculating the posterior probability of future dose transitions, are not far removed from this simple rule. Figure 8 divides the bulk summaries of CRM’s MTD-selection performance (with Prior A) in each scenario into four groups, according to the time at which settling is first encountered – before the ninth cohort, after 9–12 cohorts, after 13–16 cohorts, or not at all. Recall that the simulation had 16 cohorts of size 2. Bar lengths are proportional to group sizes, and the shaded regions represent the runs pointing to the correct MTD at the time of settling.

Figure 8 presents little evidence that early settling is associated with better MTD performance, if anything, the contrary. Furthermore, while (as shown in Figure 5 and Table 1) CRM performance varies widely between scenarios, its settling behavior itself is remarkably uniform across scenarios. Under all scenarios, roughly half the runs encounter five consecutive identical allocations by cohort 8, and 80%–90% of the runs display this phenomenon by cohort 12. All in all, Figure 8 suggests that early settling is a universal CRM design side effect, rather than a sign of quick convergence to the true MTD.

Simulations with random \( F \)

The simulations described above follow the common approach used in the field: fix a handful of scenarios for \( F \), then simulate ensembles of random samples from each scenario. This approach is useful for isolating sources of variability but is deficient as a performance evaluation tool. Inevitably, any small set of consciously selected scenarios is favorable to some methods and unfavorable to others. Moreover, such simulations under-represent the amount of variability encountered in practice.

For this purpose, and also to establish the broadness of the numerical results of Figures 5 to 8, we present a brief summary of simulation results under an ensemble of random scenarios of \( F \), using the methodology presented in Ref. [15]. Random \( F \) simulations better approximate real-life situations, in the sense that each Phase I experiment examines the effect of a different treatment for a different

Table 2. Similar to Table 1, but only with CRM, using the same skeleton and three different priors labeled A, B, and C. The first three columns show each prior’s predictive MTD distribution.

| Scenario/MTD       | Prior weight | After 8 cohorts | After 16 cohorts |
|--------------------|--------------|-----------------|-----------------|
|                    | A | B | C | A | B | C | A | B | C |
| 'Uniform'/\( d_1 \) | 0.25 | 0.26 | 0.33 | 50.2 | 53.3 | 50.2 | 62.1 | 63.1 | 64.5 |
| 'Gamma'/\( d_2 \)  | 0.14 | 0.10 | 0.22 | 36.6 | 34.4 | **44.3** | 47.4 | 45.9 | **54.0** |
| 'Normal'/\( d_3 \)  | 0.20 | 0.15 | 0.25 | 57.8 | 52.6 | **63.2** | 67.5 | 66.0 | **72.6** |
| 'Lognormal'/\( d_4 \) | 0.22 | 0.18 | 0.16 | 46.7 | 46.6 | 45.0 | 59.3 | 55.4 | 56.4 |
| 'Weibull'/\( d_5 \) | 0.14 | 0.17 | 0.04 | 39.0 | 39.9 | 23.0 | 47.2 | 50.6 | 29.7 |
| 'Logistic'/\( d_6 \) | 0.05 | 0.15 | 0.002 | 26.0 | **34.6** | 0.0 | 29.3 | **41.0** | 7.2 |

CCD: Cumulative Cohort Design; CRM: Continual Reassessment Method; MTD: Maximum Tolerated Dose.
disease, on different patient populations. Other earlier attempts to generate $F$ randomly for Phase I simulations include Refs [48] and [49].

Rather than the relatively generous $n = 32$ of the fixed-scenario simulations, our random-scenario simulations used $n = 25$, more in line with sample sizes of experiments described in this article. Simulated ‘patients’ were treated one by one. We used $l = 7$ and $l = 4$ levels, enabling us to directly incorporate the model skeletons of Flinn et al. [27] and Pisters et al. [30], respectively. The former was used with the default Cheung prior, while the latter used the published prior that had mean 0 and variance 1.8. Cheung’s ‘crm’ function was used to calculate dose transitions and final estimates. The target toxicity rate $p$ was left unchanged at 0.3. For the CCD interval design, we used a width of $0.1\times$ as before for both values of $l$.

The leading Up-and-Down design for one-at-a-time treatments and $p=0.3$ is ‘k-in-a-row’ [40,41], used extensively in sensory studies under the name ‘forced-choice fixed staircase’. Its allocation rule is

1. Start at a predesignated dose, and treat one patient a time
2. After patient $c$ is treated at $X_{c} = d_{u}$, set $X_{c+1}$ to

$$
\begin{cases}
    d_{u+1} & \text{if } Y_{c} = Y_{c-1} = 0 \text{ and } X_{c} = X_{c-1} = d_{u} \\
    d_{u-1} & \text{if } Y_{c} = 1 \\
    d_{u} & \text{Otherwise}
\end{cases}
$$

In words, escalation is mandated only after two consecutive nontoxicities at the current dose, while de-escalation is mandated after every toxicity. Besides the usual boundary conditions, there is a

---

**Figure 8.** CRM estimation performance, by scenario and ‘settling’. Bar length is proportional to the number of runs in each ‘settling’ stage. The shaded portion represents those runs ‘settling’ on the true MTD at that time.

CRM: Continual Reassessment Method; MTD: Maximum Tolerated Dose.
start-up condition preventing escalation before two observations are available. This design is closely related to the cohort-2 design used earlier [41], and similarly converges to an allocation distribution peaked near $F^{-1}(0.29)$.

Ensemble size was 2000 runs for each value of $l$, each run having a different, randomly generated $F$. The ensembles are stratified samples from a larger number of random scenarios, to ensure a uniform distribution of true-MTD levels (except for somewhat lower counts on boundary levels). The $l=7$ runs began at $d_2$, and $l=4$ runs at $d_1$. Further simulation details are in Supplement D, and the R function for generating random scenarios is in Supplement F.

A figure displaying ensemble distributions of $n'$ is relegated to Supplement D due to space concerns. With $l=7$, the patterns are very similar to Figures 5 to 7, except that between-run variability tends to increase for all designs. With $l=4$, there is a marked improvement, mostly because identifying the correct MTD from among four options is easier than from among seven.

Table 3 presents summary statistics, with relatively overperforming outcomes in bold and underperforming ones in italics. Overall MTD-selection performance (top row) is remarkably similar across designs, and improves quite substantially from $l=7$ (left) to $l=4$ (right). Up-and-Down succeeds in matching Long-Memory performance, despite being handicapped by waiting two patients for each dose escalation (the Long-Memory designs are allowed to escalate at each turn). Interestingly, CRM falls somewhat behind with $l=4$. The cause was traced to the model skeleton: the skeleton’s jump from 0.4 at $d_5$ to 0.8 at $d_6$ is disproportionately large, and the predictive prior MTD distribution is uneven (0.33, 0.17, 0.28, 0.22). Changing the skeleton’s $d_4$ value to 0.6 and tightening the prior variance to produce a nearly uniform predictive distribution, increased the MTD-selection rate by 2% and reduced the proportion of low- $n'$ runs by 3% compared with the values in Table 3.

Speaking of $n'$, we replaced the commonly reported ensemble average with two threshold-based summaries. The desirable-outcome statistic (second row) is the proportion of runs in which at least half of the allocations to patients 2–25 were assigned to the true MTD. The detrimental outcome of low $n'$ runs (third row) was defined as the proportion of runs in which the number allocated to patients 2–25 is smaller than $(n-1)/l$ (i.e., <4 patients for $l=7$, and <6 patients for $l=4$) – the number obtained by simple nonsequential uniform dose allocation. The two Long-Memory designs perform very similarly on these statistics, and the contrast between them and Up-and-Down is stark: for Long-Memory designs and $l=7$, only one-quarter of runs have an intermediate value of $n'$, with the high and low tails divided roughly 4 to 3, respectively. For Up-and-Down and $l=7$, three-quarters of runs have an intermediate $n'$, with the tails divided roughly equally. The high/low ratios substantially improve for all designs with $l=4$.

In addition, we examined the proportion of runs with many toxicities observed in patients 2–25 (>9 toxicities for $l=7$, and >10 for $l=4$; fourth row). With $l=7$, Up-and-Down is safest while CCD somewhat underperforms having >10% high-toxicity runs. With $l=4$, the performance is more similar across designs.

Runs containing an ‘incoherent’ transition (escalation following toxicity or vice versa) were also counted (bottom row). While Up-and-Down is hard-wired to disallow such transitions, and the two CRM skeletons appear to be well-calibrated to prevent this during the first few cohorts until the experiment reaches MTD, CCD displays at least one ‘incoherent’ transition in most runs, doubtlessly related to its tendency to sometimes oscillate between adjacent levels.

**Discussion and recommendations**

Our simulations indicate that the leading Bayesian Phase I method – one-parameter CRM – delivers MTD-selection success rates similar to those of a simpler nonparametric Long-Memory design (CCD), and the even simpler ‘Up-and-Down’ designs.
Finding the MTD is usually Phase I’s main goal. Given the general preference for parsimony in science and in statistical modeling, this raises the question whether Long-Memory designs in general and Bayesian Phase I designs in particular should still be recommended by statisticians.

Long-Memory designs do maintain an advantage in the expected number of patients treated at the MTD during the experiment itself, approximated by the simulation ensemble averages $\bar{n}$. However, this comes at the price of dramatically increasing $n^*$’s variability, in particular, the probability of having even fewer subjects treated at the MTD than by random chance or fixed allocation. It is unclear whether clinicians, once cognizant of this trade-off, will see the increase in $\bar{n}$ as worth the added risk and complexity, especially since the MTD-selection success rate does not substantially improve, and toxicity rates are as high or higher than with Up-And-Down. We suggest that studies of future designs include an examination of $n^*$’s distribution, and report the ensemble proportions of various adverse outcomes as done in Table 3. The habit of reporting only the ensemble average for nonbinary outcomes such as $n^*$ and the toxicity rate masks the variability between simulation runs. The average does not provide researchers with information more immediately relevant on a practical level, that is, how likely is their particular experiment to produce a desirable or adverse outcome.

Bayesian Phase I designs, and in particular CRM, might offer a sense of control and versatility, and a shot at pinpointing the MTD early on by settling on it. But we have no good way of knowing whether the experiment indeed settled on the right dose, before it is too late. These designs necessitate intricate tweaking and calibration that Lee and Cheung [14] described as a time-consuming search in multidimensional space. Even in recent years, some of the best-known CRM design experts still occasionally find themselves stumped by its unexpected behavior [31,19] (the latter is described in Supplement B). Last but not least, it is rather difficult to avoid perceiving the settling behavior as a sign of convergence, especially when accompanied with reassuring model-based terminology, and it is rather tempting to stop the experiment prematurely when settling occurs.

The nonparametric CCD offers MTD-selection performance and convergence properties similar to CRM’s – without any need for model skeletons, finely tuned parameter distributions, or other sophisticated design tools. Only an interval and a set of dose levels are required. CCD experimental trajectories are somewhat quirky, tending to oscillate between temporary barriers and to occasionally generate ‘incoherent’ transitions (Table 3). If one chooses an interval design such as CCD, care must be taken to communicate to clinicians that such oscillating behavior is likely. Furthermore, CCD suffers from the same sensitivity to sampling order, and from the associated $n^*$ variability common to all Long-Memory designs. To complete the Long-Memory picture, Supplement E presents $n^*$ simulation distributions from a two-parameter Bayesian Phase I design. The variability is just as high as with the one-parameter and nonparametric designs.

Using long memory or likelihood models for dose-finding can be a useful principle. Rather than the long memory itself, Long-Memory designs’ core vulnerability is their ‘winner-take-all’ decision rule: attempting to allocate the MTD itself at each step. After a handful of observations had accumulated, the incremental changes to the likelihood between successive cohorts become too small to force a dose transition under the winner-take-all paradigm, and these increments continue to diminish as the experiment progresses (see the right-hand sides of Figures 2 to 4). However, the $F$ values underpinning the likelihood are still rather imprecise at these small samples. This explains why ‘settling’ is a poor predictor of the correct MTD selection (see section “Settling” and estimation success’ and Figure 8), and the resultant sensitivity to sampling order. Since Bayesian Phase I designs weigh the likelihood with a prior, the impact of new data is even smaller for them compared with non-Bayesian designs, settling is earlier and more pervasive, and outcomes become sensitive to both order and predictive-prior weight (see Table 2, and also an experiment described in Supplement B).

As briefly mentioned in ‘Preliminaries’, there exist Bayesian dose-finding approaches that attempt to optimize information collection during dose assignment, rather than optimizing treatment. They are known as Bayesian Decision Procedures [25,26]. Due to space limitations, and since these designs have been rarely discussed or put to the test, we have not included them in this article.

Interestingly, under a one-parameter model, the optimization of treatment also optimizes information collection about $F^{-1}(p)$, and therefore, at face value, one-parameter CRM is ‘properly’ Bayesian, optimizing both utilities of interest. This is a slight-of-hand: proper Bayesian methods assume that the model is a reasonable attempt to approximate reality. By contrast, CRM’s one-parameter approach is universally described as a ‘working model’, rather than a realistic approximation of $F$. Researchers trying to construct a one-parameter CRM model around a realistic view of $F$, and to incorporate prior scientific knowledge into it, are likely to encounter challenges similar to those reported by Neuenschwander et al. [18]. This is because such attempts inevitably abandon some crucial element of the delicate balance between conflicting constraints upon operating
characteristics, which is the real purpose toward which CRM has retooled Bayesian machinery.

Some interesting recent attempts to modify Bayesian Phase I dose-assignment rules [50–52] will probably not resolve order sensitivity, unless the underlying loss function is modified to discourage a winner-take-all solution. Bartroff and Lai [53] and Azriel et al. [17], both writing about ‘the treatment vs. experimentation dilemma’, each offered a new design that attempts to resolve this dilemma. We have been able to examine the Azriel et al. design, and it does not alleviate the variability in \( n^* \) (see figure in Supplement E).

This brings us back to Up-and-Down, a design family that has received scant attention in recent Phase I literature. Regardless, Up-and-Down is used every day in dozens of similar applications, including dose-finding anesthesiology experiments. A recent didactic article, co-authored by an anesthesiologist and a statistician, has helped to substantially update and improve the application practices of Up-and-Down in that field [54].

The recent common wisdom in Phase I methodological literature has suggested that Long-Memory designs, in particular Bayesian Phase I designs, deliver far better MTD-selection success rates than any other available approach including Up-and-Down. In view of results presented in ‘Numerical demonstrations’, this notion has been misguided. One can trace its origins to studies that inadvertently used the last assigned dose to represent Up-and-Down’s ‘estimate’ [55], or to studies that chose scenarios highly favorable to the Bayesian Phase I models examined. But most often, Long-Memory designs have been compared only to each other or to ‘3 + 3’, whose MTD-selection performance is poor.

Up-and-Down does have its limitations, such as a relatively rigid design and no ‘natural’ standard estimator. The latter problem has been largely mitigated with the adaptation of isotonic regression to Up-and-Down [39,44]. If statisticians invest more resources in nonparametric dose-finding estimators or in Up-and-Down, this design family’s performance is likely to improve even further.

The well-known ‘two-stage’ Long-Memory approach [7,21,56] starts with a stage of single-patient cohorts, escalating until the first toxicity. This stage’s dose-transition rules are identical to a median-targeting Up-and-Down [34]. However, the first-toxicity transition to Long-Memory would occur too early to avoid the side effects described in this article. A simultaneous hybrid Up-and-Down/Long-Memory approach with interesting properties was developed by Narayana in the 1950s, and recently rediscovered and discussed by Ivanova et al. [40]. A newer hybrid design incorporating Up-and-Down in a role analogous to the sequential probability ratio test’s \( continue sampling \) option [57] was presented in Oron’s [44] dissertation. It succeeds in increasing \( \pi^* \) compared with Up-and-Down, while retaining low variability and somewhat improving MTD-selection performance (see Supplement E). The Narayana design can be seen as a simple special case of this hybrid design family. This is an area of ongoing research.

Beyond debating and improving designs, there are simpler recommendations methodologists can agree upon. First and foremost, defining a Phase I study as an exercise in dose-selection rather than dose-estimation further degrades the amount of information, in an application whose main challenge is information scarcity. In other words, it will be beneficial for the field if we succeed in convincing clinicians to accept an estimate of \( F^{-1}(p) \) on a continuous scale as the MTD for Phase II, rather than try to pinpoint the best candidate from among a small number of arbitrary dose levels. This is especially true when there is no good MTD candidate to be found, or when two dose levels are nearly equally suitable. It should be pointed out that the MTD-selection success rates presented in section ‘Numerical demonstrations’ are somewhat optimistic, since the simulations excluded scenarios belonging to these two categories because the notion of ‘true MTD’ is not very meaningful for them.

As long as the experiment’s goal is still defined as dose selection rather than estimation, methodologists should highlight the trade-off between \( n \) and \( l \) (more patients whenever possible and less dose levels to improve MTD detectability), and align expectations based on the final choice of \( n \) and \( l \). For example, our simulations as well as conceptual considerations indicate that with \( n=20 \) and \( l \geq 6 \), one should not expect better than roughly even odds of finding the true MTD, regardless of the design. Therefore, under the dose-selection paradigm, methodologists should strive to make sure there are at least five to six patients per dose level to ensure reasonable prospects for MTD-selection success.

Acknowledgments

Both authors are deeply grateful to the Editor, the Associate Editor, and the two anonymous reviewers for their thoughtful suggestions and, at times, lively debate. Oron thanks Yuehui Wu, Jay Bartroff, and Jimmy Doi for inviting him to present some of the ideas summarized in this article at conferences, and is also grateful to many colleagues, in particular Vladimir Dragalin, for fruitful and entertaining discussions. A special thanks is extended to Nancy Flournoy, for her mentorship and moral support through the years.
Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest
None declared.

References
1. O’Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase I clinical trials in cancer. Biometrics 1990; 46(1): 33–48.
2. Babj J, Rogatko A, Zacks S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. Stat Med 1998; 17: 1103–20.
3. Leung DHY, Wang VG. Isotonic designs for phase I trials. Control Clin Trials 2001; 22(2): 126–38.
4. Yuan Z, Chappell R. Isotonic designs for phase I cancer clinical trials with multiple risk groups. Clin Trials 2004; 1(6): 499–508.
5. Ivanova A, Flournoy N, Chung Y. Cumulative cohort design for dose-finding. J Stat Plan Inference 2007; 137: 2316–27.
6. Carter SK. Study design principles in the clinical evaluation of new drugs as developed by the chemotherapy programme of the National Cancer Institute. In Staquet MJ (ed.), The Design of Clinical Trials in Cancer Therapy. Editions Scientifique Europe, Brussels, 1973, pp. 242–89.
7. Storer BE. Design and analysis of phase I clinical trials. Biometrics 1989; 45(3): 925–37.
8. Reiner E, Paoletti X, O’Quigley J. Operating characteristics of the standard phase I clinical trial design. Comput Stat Data An 1999; 30: 303–15.
9. Lin Y, Shih WJ. Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. Biostatistics 2001; 2(2): 203–15.
10. Rogatko A, Schoeneck D, Jonas W, et al. Translation of innovative designs into phase I trials. J Clin Oncol 2007; 25(31): 4982–96.
11. Ivy SP, Siu LL, Garrett-Mayer E, Rubinstein L. Approaches to phase I clinical trial design focused on safety, efficiency, and selected patient populations: A report from the clinical trial design task force of the National Cancer Institute Investigational Drug Steering Committee. Clin Cancer Res 2010; 16(6): 1726–36.
12. Shen LZ, O’Quigley J. Consistency of continual reassessment method under model misspecification. Biometrika 1996; 83(2): 395–405.
13. Cheung YK, Chappell R. A simple technique to evaluate model sensitivity in the continual reassessment method. Biometrics 2002; 58(3): 671–74.
14. Lee SM, Cheung YK. Model calibration in the continual reassessment method. Clin Trials 2009; 6: 227–38.
15. Oron AP, Azriel D, Hoff PD. Dose-finding designs: The role of convergence properties. Int J Biostat 2011; 7(1): Article 39.
16. Azriel D. A note on the robustness of the continual reassessment method. Stat Probab Lett 2012; 82: 902–06.
17. Azriel D, Mandel M, Rinott Y. The treatment versus experimentation dilemma in dose finding studies. J Stat Plan Inference 2011; 141(8): 2759–68.
18. Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. Stat Med 2008; 27: 2420–39.
19. Resche-Rigon M, Zohar S, Chevret S. Adaptive designs for dose-finding in non-cancer phase II trials: Influence of early unexpected outcomes. Clin Trials 2008; 5: 595–606.
20. Zohar S, Resche-Rigon M, Chevret S. Using the continual reassessment method to estimate the minimum effective dose in phase I dose-finding studies: A case study. Clin Trials 2011; doi: 10.1177/1740774511411593 (Epub ahead of print).
21. Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR. A comprehensive comparison of the continual reassessment method to the standard 3+3 dose escalation scheme in phase I dose-finding studies. Clin Trials 2008; 5: 465–77.
22. Chevret S. The continual reassessment method in cancer phase I clinical trials: A simulation study. Stat Med 1993; 12: 1093–108.
23. O’Quigley J, Zohar S. Experimental designs for phase I and phase I/II dose-finding studies. Br J Cancer 2006; 94(5): 609–13.
24. Fedorov VV, Flournoy N, Wu Y, Zhang R. Best intention designs in dose-finding studies. Cambridge: Isaac Newton Institute for the Mathematical Sciences, 2011. Available at: http://www.newton.ac.uk/preprints/NI11065.pdf
25. Whitehead J, Brunier H. Bayesian decision procedures for dose determining experiments. Stat Med 1995; 14: 885–93.
26. Whitehead J, Patterson S, Webber D, Francis S, Zhou Y. Easy-to-implement Bayesian methods for dose-escalation studies in healthy volunteers. Biostatistics 2001; 2: 47–61.
27. Flinn IW, Goodman SN, Post L, et al. A dose-finding study of liposomal daunorubicin with CVP (COP-X) in advanced NHL. Ann Oncol 2000; 11: 691–95.
28. Dougherty TB, Porche VH, Thall PF. Maximum tolerated dose of nalmefine in patients receiving epidural fentanyl and dilute bupivacaine for postoperative analgesia. Anesthesiology 2000; 92: 1010–16.
29. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. Stat Med 1995; 14: 1149–61.
30. Pisters PW, Patel SR, Prieto VG, et al. Phase I trial of preoperative doxorubicin-based concurrent chemoradiation and surgical resection for localized extremity and body wall soft tissue sarcomas. J Clin Oncol 2004; 22(16): 3375–80.
31. Mathew P, Thall PF, Jones D, et al. Platelet-derived growth factor receptor inhibitor imatinib mesylate and docetaxel: A modular phase I trial in androgen-independent prostate cancer. J Clin Oncol 2004; 22(16): 3323–29.
32. Cheung YK. Coherence principles in dose-finding studies. Biometrika 2005; 92(4): 863–73.
33. Lindley DV. Theory and practice of Bayesian statistics. Statistician 1983; 32: 1–11.
34. Dixon WJ, Mood AM. A method for obtaining and analyzing sensitivity data. J Am Stat Assoc 1948; 43: 109–26.
35. Derman C. Non-parametric up-and-down experimentation. Ann Math Stat 1957; 28: 795–98.
36. Tsutakawa RK. Random walk design in bio-assay. J Am Stat Assoc 1967; 62: 842–56.

http://ctj.sagepub.com
37. Durham SD, Flournoy N. Up-and-down designs. I: Stationary treatment distributions. In Flournoy N, Rosenberger WF (eds). Adaptive Designs (vol. 25 of IMS Lecture Notes Monograph Series). Institute of Mathematical Statistics, Hayward, CA, 1995, pp. 139–57.

38. Gezmu M. The Geometric Up-and-Down Design for Allocating Dosage Levels. American University, Washington, DC, 1996.

39. Stylianou M, Flournoy N. Dose finding using the biased coin up-and-down design and isotonic regression. Biometrics 2002; 58(1): 171–77.

40. Ivanova A, Haghighi AA, Mohanty SG, Durham SD. Improved up-and-down designs for phase I trials. Stat Med 2003; 22: 69–82.

41. Oron AP, Hoff PD. The k-in-a-row up-and-down design, revisited. Stat Med 2009; 28: 1805–20.

42. Gezmu M, Flournoy N. Group up-and-down designs for dose-finding. J Stat Plan Inference 2006; 136(6): 1749–64.

43. R Development Core Team. R: A Language and Environment for Statistical Computing. R Development Core Team, Vienna, 2011.

44. Oron AP. Up-and-Down and the Percentile-Finding Problem. University of Washington, Seattle, WA, 2007, Available at: http://arxiv.org/abs/0808.3004

45. Cheung YK. Dose Finding by the Continual Reassessment Method. Chapman and Hall/CRC, Boca Raton, FL, 2011.

46. O’Quigley J. Theoretical study of the continual reassessment method. J Stat Plan Inference 2006; 136(6): 1765–80.

47. Zohar S, Chevret S. The continual reassessment method: Comparison of Bayesian stopping rules for dose-ranging studies. Stat Med 2003; 20: 2827–43.

48. O’Quigley J, Paoletti X, Maccario J. Design efficiency in dose finding studies. Biostatistics 2002; 3: 51–56.

49. Paoletti X, O’Quigley J, Maccario J. Design efficiency in dose finding studies. Comput Stat Data Anal 2004; 45: 197–214.

50. Ji Y, Li Y, Bekele BN. Dose-finding in phase I clinical trials based on toxicity probability intervals. Clin Trials 2007; 4(3): 235–44.

51. Ji Y, Li Y, Bekele BN. A modified toxicity probability interval method for dose-finding trials. Clin Trials 2007; 7: 653–63.

52. Yin G, Yuan Y. Bayesian model averaging continual reassessment method in phase I clinical trials. J Am Stat Assoc 2009; 104: 954–68.

53. Bartroff J, Lai TL. Approximate dynamic programming and its applications to the design of phase I cancer trials. Stat Sci 2010; 25: 245–57.

54. Pace NL, Stylianou MP. Advances in and limitations of up-and-down methodology: A précis of clinical use, study design, and dose estimation in anesthesia research. Anesthesiology 2007; 107(1): 144–52.

55. O’Quigley J, Chevret S. Methods for dose finding studies in cancer clinical trials – A review of a Monte-Carlo study. Stat Med 1991; 10(11): 1647–64.

56. Storer BE. An evaluation of Phase I clinical trial designs in the continuous dose-response setting. Stat Med 2001; 20: 2399–408.

57. Wald A. Sequential tests of statistical hypotheses. Ann Math Stat 1945; 16(2): 117–86.

Glossary

| Mathematical (Latin) | Definition |
|----------------------|------------|
| $d_1, \ldots, d_l$  | Dose levels indexed 1 through $l$. |
| $\bar{F}_i, F$      | The observed toxicity rate at dose level $i$; the set of all observed toxicity rates. |
| $G$                 | In a model-based design, the dose–toxicity model curve that approximates $F$. |
| $l$                 | The number of distinct dose levels available in a specific dose-finding design. |
| $M$                 | The simulation ensemble size: the number of virtual ‘experiments’ run under comparable conditions. |
| $n$                 | The sample size (number of cohorts). |
| $X_i$               | A random variable representing the dose assigned to cohort $i$, with $i$ being the cohort’s order in the experiment. |
| $Y_i$               | A random variable representing the number of toxic responses in cohort $i$. |

| Mathematical (Greek) | Definition |
|----------------------|------------|
| $\theta$             | In a model-based design, the parameter(s) that need to be estimated from the data. |
| $\phi$               | In a ‘power’ Continual Reassessment Method design, the vector of $l$ prior toxicity-rate estimates fixed by researchers and not estimated from the data. |