The New England Journal of Statistics in Data Science Volume 3, 222–233 (2025)
DOI: https://doi.org/10.51387/24-NEJSDS74

Up-and-Down: The Most Popular, Most Reliable, and Most Overlooked Dose-Finding Design

Assaf P. Oron* and Nancy Flournoy

Abstract

Up-and-Down designs (UDDs) are ubiquitous for dose-finding in a wide variety of scientific, engineering, and clinical fields. They are defined by a few simple rules that generate a random walk around the target percentile. UDDs' combination of robust, tractable behavior, straightforward usage, and good dose-finding performance, has won the trust of practitioners and their consulting analysts across fields and continents. In contrast, in recent decades the statistical dose-finding design field has turned a cold shoulder towards UDDs, and it is quite possible that many younger dose-finding methods researchers are not even aware of this design approach.

We present a concise overview of UDDs and their current state-of-the-art methodology, with references for further inquiry. We also revisit the performance comparison between UDDs and novel, more complicated design approaches such as the Continual Reassessment Method and the Bayesian Optimal Interval design, which we group under the term "Aim-for-Target" designs. UDDs fare very well in the comparison, particularly in terms of robustness to sources of variability.

keywords and phrases: Adaptive designs, Dose-finding, Up-and-Down, Staircase method.

1. INTRODUCTION

Up-and-Down designs (UDDs) were developed in the 1940s, independently on two continents and in two different fields: sensory studies [65] and explosive testing [10]. They remain the dose-finding method of choice in both fields [59, 44], and are very popular in many other fields including anesthesiology [39, 52], dentistry [61], toxicology [54], materials science and engineering [25, 57], electrical engineering [71], and more. UDDs are considered a standard or recommended design in these fields by many national [2, 11, 34, 43] and international [29, 30, 42, 45] organizations.

One surprising domain where UDDs have become rather *unpopular* and mostly neglected in recent decades, is the statistical field of dose-finding methodology. Amid a veritable explosion of articles presenting, modifying, and discussing novel dose-finding designs, UDD methodology articles have dwindled to less than a trickle. This relative silence in the statistical community is surprising in several ways:

- Statisticians are the ones who had spearheaded UDD methodological development during the design's early decades;
- Those decades, followed by a rather abrupt neglect, have left behind them many key unresolved challenges, both theoretical and practical;
- Judging by the sheer number of UDD experiments taking place across such a wide array of fields, one would expect that statistical consulting needs alone would have spurred many statisticians to continue investigating and improving UDD methodology. To wit, Oron's long affair with UDDs began with a 2003 graduate-student consulting project.
- Last, but not least: when one compares UDDs' dose-finding performance with newer, far more complicated designs, and does so on a level playing field – UDDs tend to hold their own [13, 19, 48]. When robustness is examined, UDDs are generally far more robust than these newer designs.

The last point alone, a variation on Occam's Razor, should convince statisticians to take UDDs seriously again. Why invest so much in design overhead, when a simpler more straightforward method does the job at least as well? One plausible explanation for the collective overlooking of UDDs is that after several decades outside the statistical limelight, they have simply receded beyond the horizon of methods that most active and incoming statisticians are familiar with. Our aim here is to pique the reader's interest regarding UDDs, and to provide concrete information for getting started, both methodologically and in a consulting capacity. Following a brief overview of UDDs and recent methodological developments, we will present fresh simulation data comparing UDD performance with leading newer designs. The latter will be described and discussed only to the extent required for such a comparison, as we maintain the article's main focus upon UDDs. We end with a general discussion.

*Corresponding author.

2. UP-AND-DOWN OVERVIEW

2.1 Basics

Due in part to the scarcity of authoritative material, there is no single definition that distinguishes UDDs from other dose-finding designs, some of which are closely related. We prefer to define UDDs as sharing the following elements [52]:

1. The responses, $\mathbf{Y} = \{Y_i,\ i = 1,\ \ldots n\}$, are binary or dichotomized.[1] We will refer to the two options verbally as "positive" and "negative", even though they are coded numerically as 1 and 0.
2. The treatments $\mathbf{X} = \{X_i,\ i = 1,\ \ldots n\}$ (often generically known as *"doses"*) are selected from a discrete set of increasing values $\mathcal{X} = d_1 < d_2 < \cdots < d_M$, which we will call **dose levels**. We assume here that $\mathcal{X}$ is finite, without loss of generality.[2]
3. The probability of positive response is monotone over $\mathcal{X}$; without loss of generality we assume monotone increasing. The probability is usually denoted via the dose-response function $F(x)$, where x is the continuous treatment-magnitude variable. The dose levels $\mathcal{X}$ are simply specific discrete values of x. It is common and often useful to think of $F(x)$ as a cumulative distribution function (CDF) of response thresholds, but it is not required.
4. Treatments are allocated sequentially and (for each new subject or cohort) only allow for increasing by one dose level, decreasing by one dose level, or no change from the current level. Hence, the design's name "up-and-down," or (in sensory studies and materials testing) the *"Staircase Method"* [63].
5. **Dose-transition rules** are based on the treatments and responses of the most recent observations – up to k of them (with $k \geq 1$ constant), and possibly also on a few additional fixed design parameters. The rules involve no estimation.
6. UDDs have no intrinsic stopping rules, although such rules can be constructed optionally.

Using this terminology, a dose-finding experiment's typical goal is estimating the **target percentile** (also known as the "target dose" or simply "the target") $F^{-1}(\Gamma), \Gamma \in (0, 1)$.

Elements 1–3 in the list above are common to dose-finding designs in many fields, and define the dose-finding task's characteristic constraints. Element 4 has become a widely (though not universally) accepted guideline across most dose-finding designs. The remaining two elements turn a dose-finding design on a grid, into a UDD. With UDDs, $\mathbf{X}$ is a random walk over $\mathcal{X}$. It is also a regular random walk, meaning that the distribution of $\mathbf{X}$ over $\mathcal{X}$ converges to a stationary distribution $\boldsymbol{\pi}$.

[1] Some ordinal forms of $\mathbf{Y}$ may also be possible; see Discussion.
[2] Preferably, dose levels are uniformly spaced in an algebraic or geometric sequence, but this is not required.

UDD dose-transition probabilities depend only upon $F(x)$ and the design's specific rules. If the 'up' transition probability decreases with increasing $F(x)$ and vice versa for the 'down' probability, then the UDD generates a random walk with a central tendency [14, 27], and $\boldsymbol{\pi}$ is sharply peaked around $F^{-1}(\Gamma)$ – or more precisely, around the **UDD balance point** $x^* \equiv F^{-1}(p^*)$ [50]. The balance point can be determined from the specific UDD chosen by solving the equation

$$\Pr\left(up \mid F(x) = p^*\right) = \Pr\left(down \mid F(x) = p^*\right). \qquad (2.1)$$

Specifically, the dual monotonicity conditions on the dose-transition probabilities guarantee that $\boldsymbol{\pi}$'s mode is at one of the two dose levels straddling x^*. The conditions are known as *the Durham-Flournoy conditions* after the researchers who first spelled them out [14, 12, 50]. Without meeting these conditions, a design might still be considered a UDD – that is perhaps a matter of semantics – but it is unlikely to work well as a *dose-finding* UDD.

The balance-point equation (2.1) holds for all UDD variants described in Section 2.2; one only needs to plug the correct transition probabilities into the formula. In general, design parameters should be chosen so that $p^* \approx \Gamma$.

The convergence of $\mathbf{X}$ towards stationary behavior happens at a very rapid, geometric rate, meaning that within a few dozen observations and usually sooner, a contiguous "slice" of $\mathbf{X}$ will be essentially equivalent to a sample out of $\boldsymbol{\pi}$ [9].

Before we continue, a few words about robustness, a term mentioned frequently in this article. Dose-finding is a small-to-moderate sample affair; in most fields n is rarely over 50, and in many fields it is usually ≤ 25 [28]. Each of these observations is binary, so the experiment provides a few dozen bits of information at best. Thus, the overall signal-to-noise ratio cannot be very high, particularly when observations are obtained from live subjects rather than, say, industrially-produced units. Even under idealized simulated conditions in which all response thresholds are drawn from a single well-defined $F(x)$ and there are no experimental mishaps, challenging situations are common. For example, the target $F^{-1}(\Gamma)$ might be situated relatively far from the starting dose x_1, or, very commonly, different parts of the experiments might encounter "streaks" of relatively high or low response thresholds compared with the population average, so that experimental behavior seems erratic and the target percentile might not be clearly discernible from the data.

In this context, a design or estimator being robust means that its dose-allocation behavior and dose-finding performance show little degradation under such more challenging situations. Conversely, some design approaches are intrinsically oriented towards capitalizing upon well-behaved conditions, but falter under moderate deviations from such conditions. In the terminology we use here, this indicates lack of robustness.

2.2 Popular Types of UDDs

The original UDD has the simplest of rules: escalate when $Y_i = 0$ and vice versa. Therefore, $\Pr(up) = 1 - F(x)$ and $\Pr(down) = F(x)$. Whether by plugging this into (2.1) or simply by symmetry, evidently $p^* = 0.5$. To date this is the most commonly and widely used UDD. Below we list three popular straightforward extensions that enable targeting other percentiles, while remaining only once removed from the original UDD and meeting all six criteria listed in the UDD definition above, as well as the Durham-Flournoy conditions.

A simple extension that can target any percentile is known as Biased-Coin UDD [12]. For $\Gamma < 0.5$, following $Y_i = 0$ one draws a random number to determine whether to escalate or repeat the same dose. In contrast, $Y_i = 1$ mandates a de-escalation. Setting the random (*"biased coin"*) escalation probability to $\Gamma/(1-\Gamma)$ ensures that $p^* = \Gamma$ exactly. For $\Gamma > 0.5$ the roles of Y_i are reversed, and the biased-coin probability is inverted. The `bcoin` utility function in the R package `upndown` provides the required coin probability to achieve a given Γ. The utility also returns a verbal description of transition rules, to clarify how the result is to be used:

```
> bcoin(0.3)
After positive response, move DOWN.
After negative response, 'toss a COIN':
   - with probability of approximately 0.43 move UP
   - Otherwise REPEAT same dose.
```

Another simple UDD extension replaces the random draw with a requirement for a run of k contiguous negative (positive) responses at the same dose level before escalation (de-escalation), to target $\Gamma \leq 0.5$ ($\Gamma \geq 0.5$). This UDD is extremely popular in sensory studies, to which it was introduced in the 1960s by its developer G.B. Wetherill [67, 68]. It has been known by various names; we prefer the rather straightforward name "k-in-a-row UDD" [32]. Dose allocation behavior can be described either as a k-th order random walk, or as a random walk with internal states [22, 50]. For $\Gamma > 0.5$ (typical of sensory studies), the balance point is $p^* = 0.5^{1/k}$, with mirror-image balance points for $\Gamma < 0.5$ (adequate for toxicity studies). Thus, for toxicity studies the $k = 2, 3, 4$ balance points are very close to the 30th, 20th and 15th percentiles, respectively. The `k2targ` utility function in `upndown` provides the balance point for given k. The reverse utility `ktargOptions` provides plausible values of k given Γ, together with a verbal description of the rules analogous to the `bcoin` output shown above.

For both k-in-a-row and Biased-Coin UDDs, the non-median balance point is achieved by rendering one transition direction "slow", while the opposite direction retains the original UDD's "fast" transitions. Beginning an experiment from the "slow" end (e.g., starting from the lowest dose in toxicity studies) might incur a substantial delay and reduced performance if the true target is not close. A common modification, introduced already in the 1960s [67], is to start the experiment with original-UDD rules, until at least one observation of each type is encountered. In toxicity studies, this would mean escalating after every observation until the first toxicity, then reverting to the intended k-in-a-row or Biased-Coin rules. Barring extreme exceptions, this modification is highly recommended.

Lastly, the Group UDD (GUD) evaluates cohorts of fixed size $s > 1$ simultaneously, escalating with l or fewer positive responses and de-escalating with u or more [64, 23]. All members of the same cohort receive the same dose. Somewhat similarly to k-in-a-row, GUDs can be described either as an s-th order random walk, or as first-order with a twist; in this case, moving from binary Y to size-s Binomial. Balance points can be determined from symmetry when $l + u = s$ (in which case $p^* = 0.5$), by solving (2.1) analytically for some other specific GUD sub-families, and otherwise by solving (2.1) numerically from Binomial distribution probabilities. Similarly to the k-in-a-row utilities, The `g2targ` utility function in `upndown` provides the balance point for given (s, l, u). The reverse utility `gtargOptions` provides plausible (s, l, u) trios for a given Γ. See the following example:

```
> gtargOptions(0.3, maxsize = 5, tolerance = 0.05)
For each design, if positive responses <= Lower, move up
                 if positive responses >= Upper, move down
otherwise repeat same dose
    (relevant only when Upper - Lower > 1).

  Cohort Lower Upper BalancePoint
1      2     0     1    0.2928932
2      3     0     2    0.3472963
3      4     0     2    0.2663668
4      5     0     3    0.3019788
5      5     1     2    0.3138095
```

GUDs may have inspired the '3+3' escalation design [6], which is notorious in the phase I cancer trial design literature for its enduring popularity despite volumes of evidence for its poor dose-finding performance. The transition rules after '3+3's first visit to a new dose-level resemble a $\mathrm{GUD}_{(3,0,2)}$, listed in the second row of the `gtargOptions` output above. However, '3+3' stops the experiment before any dose level sees more than 6 observations, and completely disallows re-escalation to a previously visited dose. This prevents any possibility for a target-centered random walk, and therefore denies '3+3' the attendant UDD performance-beneficial properties. To emphasize: despite occasional misidentification in literature, '3+3' is not a UDD.

When these UDD variants are compared for estimation of the same target percentile, k-in-a-row converges somewhat faster to its stationary behavior [50]. This translates into an estimation-efficiency advantage, which has however become more nuanced with improvements to UDD estimation methods that have enhanced all variants' performance [49, 17]. k-in-a-row's advantage depends on it having a balance point close enough to the experiment's designated target (say, within ~ 5 percentage points).

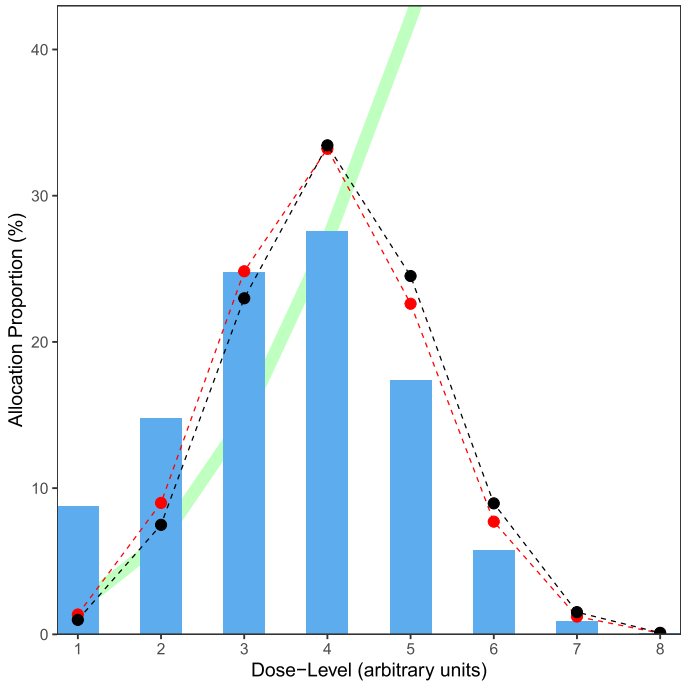

Figure 1: Illustration of UDD dose-allocation distribution and allocation convergence. Details are described in the text.

Figure 1 illustrates a UDD allocation distribution and provides insight into the somewhat elusive topic of UDD allocation convergence. The vertical bars show the expected accumulated dose-allocation distribution after $n = 30$, under k-in-a-row with $k = 2$ and the first $F(x)$ curve in the 500-curve ensemble used in the next section's simulations. Given knowledge of $F(x)$ (shown in the background as a faint green band) and the starting point (here assumed to be d_1), the distribution can be calculated analytically and was derived via the `upndown` utility `cumulvec`. The stationary or asymptotic random-walk distribution $\boldsymbol{\pi}$ (connected black dots; calculated via `pivec`) is independent of the starting point. The bars' heights are not too far removed from it, but one can see the low starting point's effect. The expected marginal distribution of additional doses halfway through the experiment at $n = 15$ (connected red dots; calculated via `currentvec`) is hardly distinguishable from $\boldsymbol{\pi}$. Around $n = 25$, the distribution of additional doses becomes visually indistinguishable from $\boldsymbol{\pi}$ at this scale. This demonstrates the meaning of dose-allocation convergence. Note that $F(x)$ is on the same scale as the allocation probabilities: it crosses 29.3% (the balance point) and 30% (the experiment's official target rate) shortly after d_4, where indeed the peaks of all depicted distributions are located.

Other UDD variants beyond the four described here have been published, some of them extending the possibilities via additional biased coins [e.g., 8, 12, 18, 24]. Generally speaking, designs with biased coins applied to both the 'up' and 'down' transitions do not provide additional practical benefit to justify the added complication, and have rarely if ever been implemented in practice. One UDD variant that does enjoy popularity in sensory studies, uses different step sizes for the up and down transitions [20]. This innovation "violates" either criterion element 4 (if the ratio between step sizes is rational) or 2 (otherwise), and hence narrowly speaking might not be considered a UDD as it does not generate a random walk on $\mathcal{X}$. However, it does generate a target-centered Markov chain (either discrete or continuous-state) that shares many properties with "proper" UDDs.

The R package `upndown` has additional utilities, such as estimation functions and even a fast-running ensemble simulation framework. We recommend using the package's development version, available via GitHub at `"assaforon/upndown"`.

2.3 Estimation

2.3.1 Regression Estimators

The estimator we recommend for UDD is Centered Isotonic Regression (CIR) [49]. Using regression for dose-finding begins by calculating the observed dose-specific response rates, $\mathbf{R} = (R_1, \ldots, R_M)$:

$$R_m \equiv \frac{T_m}{N_m}, \quad m = 1, \ \ldots M, \tag{2.2}$$

where N_m is the sample size at d_m and T_m is the number of responses (e.g., toxicities) observed among them. The rates $\mathbf{R}$ (shown as 'x' marks in Figure 2) are used to estimate the dose-response curve $F(x)$, ultimately "reading" $F^{-1}(\Gamma)$ off of the regression curve. See for example in Figure 2, how CIR's 90th percentile estimate (purple dot) is the value of x where the CIR curve (blue) crosses $y = 90\%$.

Isotonic regression methods are a good match for UDDs, as both are non-parametric and both tend to demonstrate robustness to experimental mishaps and to variations in the dose-response relationship. We prefer CIR specifically, because it offers a considerable performance improvement over straightforward interpolation of ordinary isotonic regression, an estimator introduced to UDD by Stylianou and Flournoy [60]. CIR produces more realistic dose-response curves by avoiding the characteristic flat stretches produced by ordinary isotonic regression. Figure 2 illustrates CIR and isotonic regression, with data from an anesthesiology experiment that targeted the 90th percentile using biased-coin UDD [21].

CIR also includes an accompanying confidence interval with adequate coverage, beginning with an interval for $F(x)$ based on an analytical formula for ordered binary data by Morris [40], then using a localized delta-method-like inversion to obtain an interval for $F^{-1}(\Gamma)$ [49]. It should be noted that isotonic regression has historically lacked an adequate small-sample interval. Therefore, CIR available via the R package `cir`, offers solutions relevant for dose-response and dose-finding applications far beyond UDD

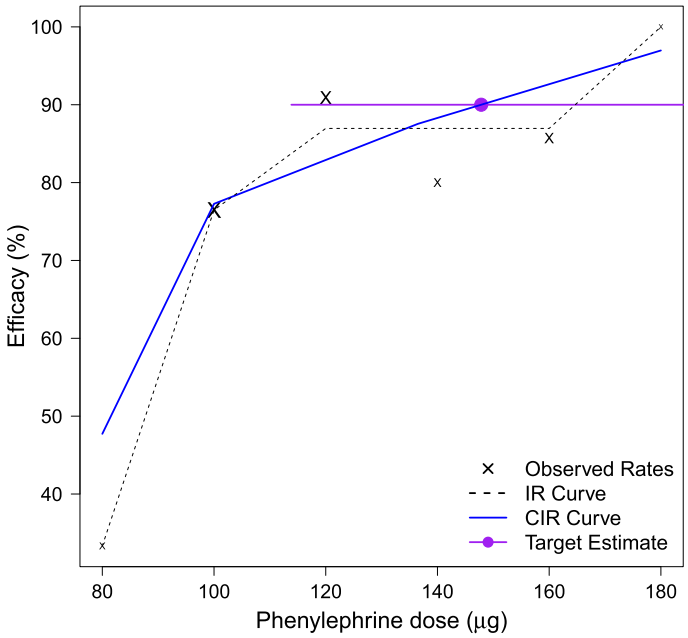

Figure 2: Example of isotonic regression and CIR using data from an anesthesiology UDD experiment with $n = 45$ that targeted the 90th percentile [21]. Isotonic regression as adapted by Stilyanou and Flournoy (black dashes) interpolates between the observed response rates $\mathbf{R}$ ('x' marks; size proportional to n_m), replacing regions of decrease with flat stretches. CIR (solid blue) collapses the flat stretches to single points, ensuring strict monotonicity. CIR also incorporates the bias-mitigation formula (2.3). The CIR target estimate and 90% confidence interval are shown in purple. The figure was generated via the `upndown` package utility `drplot`.

alone. Our confidence-interval method is compatible with both CIR and ordinary isotonic regression. The convenience function `udest` in `upndown` offers a CIR target estimate preconfigured for UDD datasets.

One might wonder why we do not recommend parametric regression, e.g., Logistic or Probit. Such methods can be found in some UDD experimental reports, but we generally advise against them because of poorly-characterized performance under model mis-specification, and the considerable chance for non-existent estimates [58, 19]. If the latter problem is circumvented via use of Bayesian priors, performance might depend too strongly upon them.

A note of caution regarding regression estimators: in dose-finding it is customary to assume that the observed rates $\mathbf{R}$ are equivalent to Binomial random variables, and therefore constitute unbiased estimates of F on $\mathcal{X}$. The assumption is wrong, not only for UDDs but for all adaptive dose-finding designs, because of the dependence between numerator and denominator in (2.2) [26]. We recently described the typical form this bias takes in dose-finding. In the target's vicinity the bias is nearly zero, and therefore it has little affect upon designs' dose-finding performance. Away from target, the bias "flares out" in both directions, making observed rates seem more extreme than the underlying values of F and therefore producing exaggerated slopes [17]. The bias tends to be stronger for non-UDD designs such as the Continual Reassessment Method (CRM) [46], because the dependence there is stronger.

Inspired by Firth [15] and informed by the shape of the bias, we developed a simple ad-hoc bias mitigation formula that shrinks $\mathbf{R}$ towards Γ:

$$\widetilde{R}_m = \frac{T_m + \Gamma}{N_m + 1} = \frac{N_m R_m + \Gamma}{N_m + 1}. \tag{2.3}$$

When $\Gamma = 0.5$, this formula is identical to the commonly used correction for calculating the empirical logit in the presence of zero cell counts [69, 1]. The bias mitigation is an option in `cir`, and implemented as default in `upndown::udest`. It tends to improve CIR interval coverage for the target-dose estimate, via making the slope of F less exaggerated. The CIR curve in Figure 2 incorporates the bias-mitigation formula.

Due to the bias, we generally advise against off-target estimates with adaptive dose-finding designs, e.g., estimating the 95th percentile using data from a median-targeting UDD. Note that many safety dose-exclusion rules implemented in other dose-finding designs rely upon such off-target estimates.

2.3.2 Dose-Averaging Estimators

Historically, dose-averaging estimators appeared before regression estimators, and are still very popular, particularly in non-medical fields. These are averages of a subset of the sequence of allocated doses, $\mathbf{X}$. The rapid convergence of $\mathbf{X}$ to stationary behavior and $\boldsymbol{\pi}$'s relative symmetry provide the basic rationale for dose-averaging estimators. A deeper justification is that nearly all the experiment's information is encoded in $\mathbf{X}$ via the dose-transition rules. One can even add a "phantom" X_{n+1} to the average, because when the experiment ends the next treatment allocation can be predetermined without need to observe Y_{n+1} [5]. Both the original Dixon-Mood UDD estimator, and the estimator developed by Wetherill and Leavitt upon UDDs' introduction to sensory studies, are dose-averaging estimators [10, 68]. The latter is likely still the single most popular UDD estimation approach when all fields are considered. It averages only the subset of doses at points where $\mathbf{X}$'s trajectory changes direction from 'up' to 'down' or vice versa.

The simplicity of averaging and the relatively low variance of using an average for estimation in general are appealing, but we have found that dose-averaging approaches tend to lack robustness. A plethora of biases counter-balances the low-variance advantage; some are very difficult or impossible to mitigate [52]. For example, a starting-point bias may be observed if the target is far from the starting dose, and a

boundary bias takes place when the target is near the edge of $\mathcal{X}$.

In addition, none of the dose-averaging confidence intervals in current use offers sufficient and robust coverage, mostly because all require a standard-error estimate, and those are hard to obtain reliably when the data are so discrete. Our `upndown` package offers a bootstrap-based interval that comes close to passable coverage for some dose-averaging estimators, but generally still falls short by at least $\sim 5\%$.

3. UP-AND-DOWN – AND OTHER APPROACHES

3.1 Background

Because dose-finding is a generic challenge that resurfaces in many contexts, a variety of approaches have been developed to address it. For comparison with UDD, we focus on the most prominent family of approaches in recent literature, one that utilizes repeated estimation. The use of estimation to guide the next treatment's placement can be traced back at least to the 1950s, nearly as old as UDDs [56, 41, 35]. More recently, in the context of dose-finding on a discrete grid $\mathcal{X}$, most estimation-based approaches have coalesced around the following outline:

- After each observation, estimate $F(x)$ – either the entire curve or the value at the current dose-level;
- Place the next treatment at the dose-level deemed "closest to target", according to these estimates and the design's specific optimization criterion.

The optimization criterion could be, e.g., the dose-level with the smallest $|\hat{F} - \Gamma|$, or – in the case of so-called "interval designs", simply that the current dose-level's estimate of F is within some tolerance interval around Γ. When the former criterion is used, the target estimation method at the experiment's end is usually identical to the dose-allocation method during the experiment.

The first dose-finding design we are aware of to follow this outline explicitly, was a parametric Bayesian design for sensory studies published in 1983 under the acronym QUEST [66]. While QUEST has gained considerable traction in its own field, it was a 1990 publication of another parametric Bayesian design that has caught mainstream statistics' attention: the aforementioned CRM [46].[3] Catering to phase I cancer trials, which is the dose-finding application receiving the most method-development resources nowadays, CRM was soon followed in that field by methods bearing acronyms such as EWOC [3], or more recently interval designs such as CCD [31], mTPI [33], and BOIN [38]. This is a very partial list.

[3]Wu also presented such an approach independently in 1985, but it seems that the reach of his work has remained confined mostly to methodological discussions of stochastic approximation and related designs [70].

Such designs have often been named "long-memory" because they incorporate information going back to (x_1, y_1), but there are other approaches with long memory that do not follow the outline as described above. Here we suggest the provisional name **Aim-for-Target** designs, which seems more specific and descriptive. Plausibly, one can also describe them as greedy algorithms [7, Ch. 15]. Our impression is that Aim-for-Target designs have taken up nearly all the oxygen in the statistical dose-finding-literature room, with attempts to dethrone or modify '3+3' in the phase I realm accounting for most of the balance. The practical needs of other fields that use dose-finding have been largely ignored in this recent body of methodological literature. UDDs are mentioned in passing, if at all, and often in a misguided manner.

Oron and Hoff demonstrated a decade ago that Aim-for-Target designs suffer from a disturbing, structural lack of robustness which, even more disturbingly, has gone almost completely under the radar of all this novel methodological activity [48]. In a nutshell, Aim-for-Target designs tend to lock onto a perceived optimum early in the trial. In case this "early bet" misses the true optimum, these designs take very long to self-correct, because their self-correction mechanism operates at a root-n rate, with new information accumulating rather slowly since it consists of dependent binary data.

In Oron and Hoff's work, UDDs were shown to attain similar dose-finding performance overall, and to have far better robustness, than Aim-for-Target. Since the evidence presented there has not become common knowledge, and since some time has passed with new designs and new developments, we have thought it appropriate to revisit the comparison with new, broader simulations.

3.2 Comparative Performance Simulations

3.2.1 Methods

We present here results for designs targeting the 30th percentile, a common phase I cancer target, and the 90th percentile, popular in anesthesiology. We refer to the $Y = 1$ outcome in the former case as dose-limiting toxicity (DLT), and in the latter as efficacy, even though both are simulated via very similar computer code. More simulations details are provided below.

We generated parametric random F curves using a 3-parameter Weibull (shape, scale, lateral shift). In order to enable separate looks into the effect of curve properties (slope, shape, etc.) and of the relationship between starting dose and target location, for each target a single ensemble of $B = 500$ was generated, with each curve having different Weibull parameter values but with all curves crossing target near the middle of $\mathcal{X}$. Extremely steep or extremely shallow curves were excluded as being less "interesting" for the dose-finding task.

Then, the simulation setting was varied by shifting the entire ensemble right or left, or by changing the starting dose. This approach follows in the footsteps of earlier randomized-F simulations [51, 48]. In its specific details it is quite similar, but somewhat more sophisticated, than the curve ensembles shown in the supplement of reference [52].

For the 30th percentile we used $M = 8$ dose levels and $n = 30$ observations, and 4 different settings using the same 500-curve ensemble. In 3 settings the starting dose was d_1 as is common in toxicity studies, and the target location was in the middle $[x^* \in (d_4, d_5)]$, low $[x^* \in (d_2, d_3)]$ or high $[x^* \in (d_6, d_7)]$. The fourth setting had the target in (d_4, d_5) and started at d_4. The 90th percentile simulations were fairly similar, except for using $M = 12$ dose levels and $n = 50$, and with 3 of the 4 settings varying the starting point (high, middle, low) rather than the target location. We kept one particularly "hard" setting, the one starting at d_1 and having a high target. The 30th percentile simulation had one set of comparisons with single-patient dose allocation decisions, and one set with cohorts of 3, a cohort size used very commonly in phase I cancer trials. The 90th percentile simulation only had a single-patient set.

For the 30th percentile cohort-allocation simulations, we used $\text{GUD}_{(3,0,2)}$ ($p^* \approx 0.35$): escalate after zero-toxicity cohorts, repeat the same dose with one toxicity, and de-escalate otherwise. For the single-patient allocation simulations, we used the k-in-a-row UDD with $k = 2$ and $k = 6$ for the 30th (below-median rules, $p^* \approx 0.29$) and 90th (above-median rules, $p^* \approx 0.89$) percentiles, respectively. For k-in-a-row we used the quick start-up modification described in Section 2.2. For UDD estimation, CIR was used including the bias-mitigation formula (2.3).

As to Aim-for-Target designs, we used three CRM variants for each target, all generated via the `getprior` function in the `dfcrm` package. This function provides a "skeleton" of F values on $\mathcal{X}$, determined by the user's choice of an "indifference interval" around target, and by the prior-predictive mode location of the target dose [36, 37]. For the 30th percentile we used an "indifference interval" half-width of 0.05, except for one variant with 0.1 half-width. For the 90th percentile, we found by trial-and-error that these intervals needed to be half as wide. Prior distributions of the estimated parameter were kept at defaults. The narrower-interval variants varied by prior-predictive mode location (high vs. low), while the wider-interval variant had its prior mode near the middle of $\mathcal{X}$. CRM dose transitions were limited to a single dose-level upwards or downwards.

We also used two interval designs: the Cumulative Cohort design (CCD) [31] and the Bayesian Optimal Interval design (BOIN) [38]. For the 30th percentile, CCD was used with an interval width of ± 0.1 as recommended by the authors, and BOIN used the transition and dose-exclusion look-up table generated via the `get.boundary` function in the `BOIN` package, using the function's defaults. For the 90th percentile, CCD's interval width was halved, and BOIN software did not permit calculation of the design rules. For estimation with both interval designs, CIR was used including the bias-mitigation formula.

All simulated experiments were run using the `dfsim` simulation utility in `upndown`, currently (fall 2024) available only in the package's development version, but eventually to become available in the CRAN version as well. Post-processing and visualization were done using auxiliary code in R version 4.3.3. The entire simulation's scripts can be found on Github under the `assaforon/UpndownBook` repository, in the folders `P2_Estimation` and `P3_Practical`.

3.2.2 Results: Main Metrics

We follow the phase I field's conventions, and rather than evaluate point estimates of $F^{-1}(\Gamma)$ using continuous metrics, we identify the dose-level in $\mathcal{X}$ with the smallest $|\hat{F} - \Gamma|$, often known as the Maximum Tolerated Dose (MTD) estimate. Phase I simulation studies usually examine what proportion of the ensemble's runs had the correct MTD estimate (i.e., the MTD estimate was indeed the dose-level with smallest $|F - \Gamma|$), or whether this estimate falls on a dose whose F value is within an "acceptable window" around Γ. Here we adopted the latter criterion; for the 30th percentile we used an "acceptable window" of $F \in [0.2, 0.4]$. We made sure that every $F(x)$ curve in the ensemble has at least one dose-level within the "acceptable window", but no more than three. The Supplementary Material includes analogous summary plots (Figures S1, S2), with the narrower criterion of correct-MTD identification for the 30th percentile simulations.

The proportion of single-patient 30th percentile simulation runs whose MTD estimate fell within the window, is plotted on the y-axis of Figure 3. The x-axis shows the ensemble-average DLT rate during the experiment. Since 30% is the target rate (marked as a dashed vertical line), rates below, or not far above 30%, should be acceptable. Thus, the desirable region of the plot is high and to the left, or at least not too far to the right. For each design we show the mean (dot) and range (lines extending from it) across the 4 starting-point and target-location settings described in Section 3.2.1. Designs more robust to changes in settings will have shorter lines extending from the mean.

On the combination of Figure 3's three metrics – dose-finding performance, toxicity and robustness – the k-in-a-row UDD (dark red) is among the best, and arguably even the single best overall. The CRM variant with wider "indifference interval" (steel blue) does well on performance. However, its considerable spread suggests less robustness, and there is more to this story as we shall see soon. Aim-for-Target designs that show similar robustness to k-in-a-row are generally lower on performance. CRM with a high prior-predictive MTD has the highest overall DLT rate, as expected. The newest design, BOIN (orange), had disappointingly low performance, and yet does not achieve lower overall DLT rate than UDD.

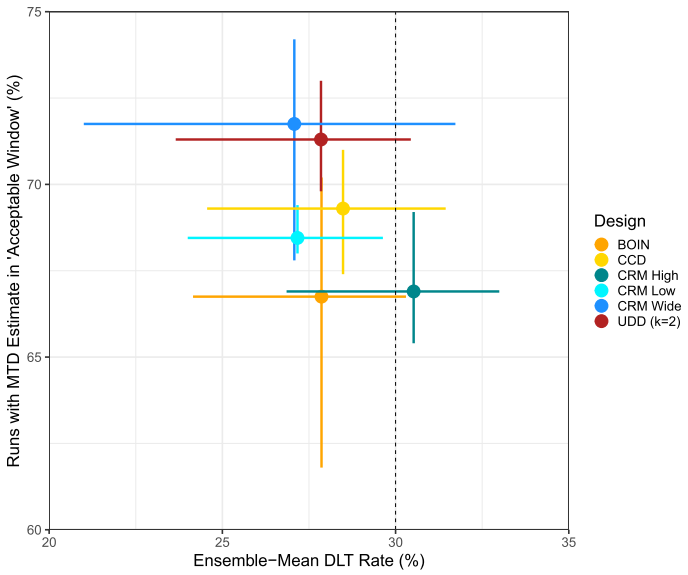

Figure 3: Main performance plot from the 30th percentile target simulations with single-patient dose allocations. Additional details are in the text.

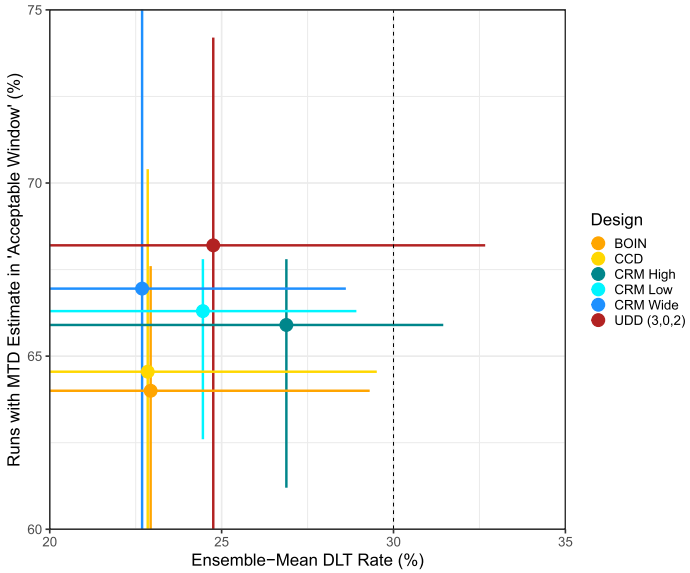

Figure 4: Main performance plot from the 30th percentile target simulations with cohorts of size 3. Additional details are in the text.

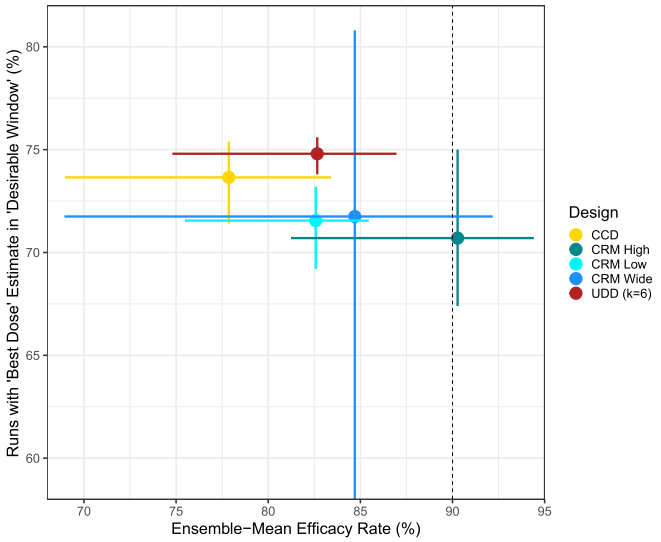

Figure 5: Main performance plot from the 90th percentile target simulations. Additional details are in the text.

Figure 4 shows summaries under identical settings except for the use of 3-patient cohorts, a common practice in phase I trials. We have retained the same plot boundaries as Figure 3, so the first thing to note is a substantial loss of dose-finding performance compared with the single-patient simulations. Much of the loss is due to the most challenging setting, under which experiments start at d_1 and the target is in (d_6, d_7); for 4 of the 6 designs, the performance under this setting now falls below the plot's lower limit of 60%. The Supplementary Material includes a version of Figure 4 where the full performance range is visible. Some designs lose altitude across the board: "CRM Wide" in particular loses 3% average performance even when the most challenging setting is excluded. More can be said about the loss of performance when a cohort structure is imposed, and whether it justifies the actual benefits – but perhaps this is a topic for another article.

Turning to our main business of UDD vs. Aim-for-Target comparison: the UDD variant is Figure 4's clear number 1 in dose-finding performance. Conversely, it is also responsible for the single highest-toxicity ensemble – 32.7% under the setting that starts at d_4 – but this is the only setting in which it exceeds 30% despite having a balance point of $p^* \approx 0.347$, and on average its toxicity rate is $< 25\%$, substantially lower than "CRM High" and similar to "CRM Low'.

For simplicity, we retained the same metrics for the 90th percentile simulations – i.e., establishing a "Best Dose" *(note it is not a "maximum tolerated dose" in this context)* and defining a "desirable window" between the 82.5 and 97.5 percentiles. The rationale for this window is that failure rates nearing 20%, (i.e., double the target rate) would be deemed too high, and conversely near-perfect efficacy rates might suggest that we are choosing doses that are excessive for the vast majority of patients. Thus, the set-up for Figure 5 is very similar to Figure 3, but now the best region in the plot is high and to the right, although still perhaps not too far to the right.[4]

Once again, UDD does very well, and once again, the wide-interval CRM (which here means a half-width of 0.05 vs. 0.025 for the other two variants) shows the worst robustness to changes in setting. BOIN is not shown because the design's official package refuses to calculate design rules for $\Gamma > 0.6$.

3.2.3 Results: Number-Treated-in-Window Metric

In Aim-for-Target editorials and simulation articles, it is commonplace to discuss and examine the metric of how many patients during the experiment were treated at the true MTD, or within the "acceptable window" around it. We call this metric n^* for brevity. It does not strike us as originating from practitioners; there's a good chance that if practitioners were the ones coming up with such a metric, statisticians would have told them it is no less than circular reasoning, because if the true MTD is known then the experiment is not needed, and since it is not known the expectation that most patients be treated at it during a small-sample, binary-Y experiment, is unrealistic.

On a possibly related topic, a hallmark of Aim-for-Target design behavior is the tendency to "settle in" relatively

[4] We reiterate that the anesthesiology field usually prefers continuous target-dose estimates rather than this discrete "Best Dose" approach, although the latter is encountered occasionally. However, even if we presented continuous estimates and metrics, the results would be similar overall.

quickly on the same dose-level for long stretches. This behavior is very widely mis-interpreted as "convergence", even leading to widespread adoption of early-stopping rules designed around it. Oron and Hoff [48] argued that since Aim-for-Target convergence is tied to the convergence of F estimates, and these take place at a root-n rate, *"late-stage convergence"* (loosely speaking, when behavior doesn't change anymore because the estimates have in fact gotten very close to their asymptotic value) is not observable at the rather small phase I sample sizes, barring the occasional lucky individual sample. Instead, the settling behavior is a side-effect of design rules, because the same model is refitted at each step with nearly the same data. Oron and Hoff also provided simulation evidence that aggressive early settling-in is unrelated, or even inversely related, to estimation performance, and therefore stopping rules based on this behavior might be detrimental. Regardless, for our narrow purposes here, we note that this settling-in behavior tends to drive n^* up for Aim-for-Target designs under favorable conditions, surely compared to UDDs and their random walk.

We examine n^*, but instead of ensemble averages we look at the entire ensemble distribution, as Oron and Hoff did in 2013. Figure 6 shows the ensemble distributions of n^* for the single-patient 30th percentile simulations, across 3 settings and 5 designs. Each pane represents a 500-run ensemble, with individual runs (single virtual "experiments") differing both by their $F(x)$ curve and by their set of random response thresholds ("patients"). In contrast, the columns differ only by the location of $F^{-1}(\Gamma)$.

While the UDD histograms (bottom row) show a clear peak with tails, most Aim-For-Target histograms are not too far removed from a uniform distribution. This means that whether the trial will be spent almost entirely within the acceptable window, almost entirely outside of it, or somewhere between those extremes – is anyone's guess. The CRM with a wider "indifference interval" (second row from bottom) is particularly volatile – and in the least favorable setting (rightmost column), the most common single value of n^* for this design is zero. Overall, Aim-for-Target ensemble-average n^* values are $\sim 10-25\%$ higher than k-in-a-row, but their ensemble standard deviations of n^* are $\sim 2x$ higher.

Figure 7 shows analogous distributions from the cohort simulations, counting allocated cohorts of size 3 instead of single patients since all patients in each cohort receive the same dose. Some differences between UDD and Aim-for-Target appear less dramatic here, both because of the strong constraint imposed by the use of cohorts, and because $\text{GUD}_{(3,0,2)}$ allows for the same dose to be repeated

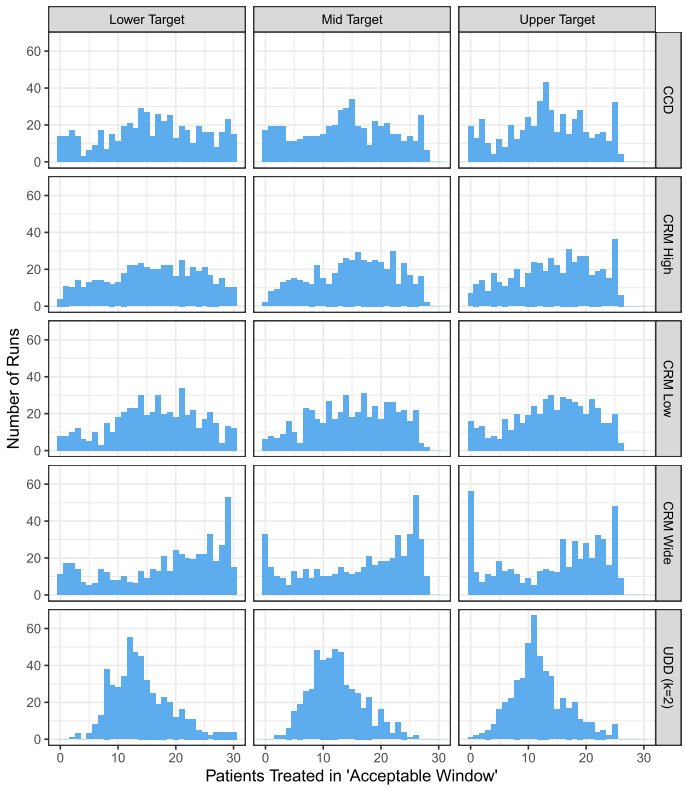

Figure 6: Distributions of the run-specific number of patients treated at acceptable dose-levels, from the 30th percentile single-patient simulations.

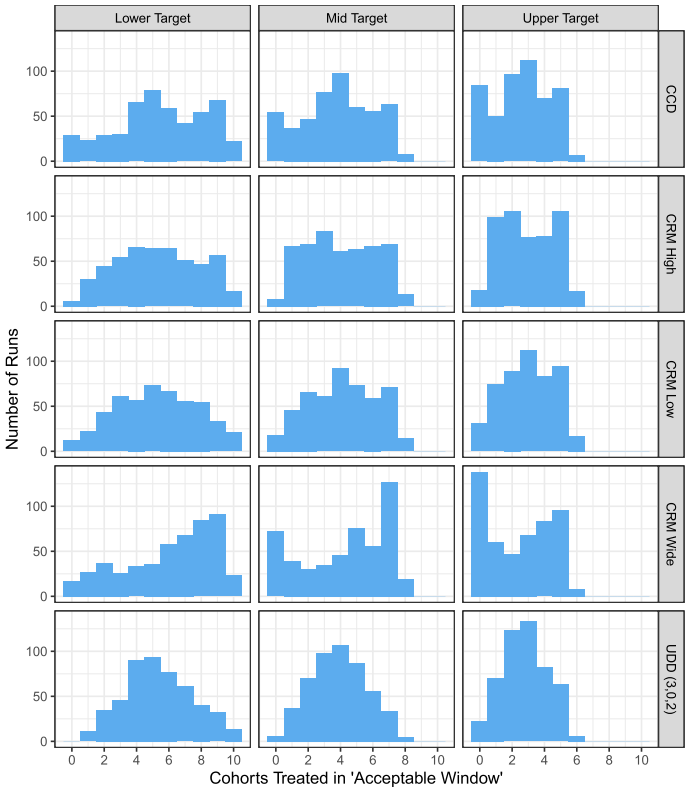

Figure 7: Distributions of the run-specific number of cohorts treated at acceptable dose-levels, from the 30th percentile cohort simulations.

over more consecutive observations than the single-patient UDD. Indeed, average n^* values for $\text{GUD}_{(3,0,2)}$ are similar to those of the other designs. It does still have the lowest standard deviation of n^* in each setting – although not by a factor of 2 as in the single-patient simulations. As seen in other simulation results, "CRM Wide" once again is the least robust, this time showing marked sensitivity both between settings and between individual runs. Under the least favorable setting, $> 25\%$ of this design's runs never made it into the "acceptable window" during the experiment.

4. DISCUSSION

UDDs are widely used with a long track record of reliability. They are high-performing, flexible and modifiable. We recommend strongly to pertinent application fields where UDDs are currently not part of mainstream discussion, to consider them again. For phase I and similar clinical toxicity trials in particular, UDDs have the dual advantage of being simpler and more tractable than '3+3', which might appeal to practitioners and regulators, yet performing at least as well as the best novel "Aim-for-Target" designs, which should appeal to everyone.

In that context, UDD's random walk is often faulted because it allows a dose experiencing multiple prior toxicities to be visited again, more readily than most estimation-based designs. This valid concern is not limited to UDDs, and a simple, generic solution is known and is applicable to UDDs as well: incorporate a dose-exclusion rule based on current information. Such rules have been proposed for UDDs at least once [47]. We note that regardless of design, many of these rules ignore the bias in $\mathbf{R}$ and hence tend to be too aggressive. Also regardless of design, there is some loss of performance in exchange for reducing the risk of visiting high doses. We plan to examine new ideas for UDD-specific dose-exclusion rules, which in contrast to the above will be cognizant of the bias, and might end up improving the design's dose-finding performance.

There are many opportunities for further extensions and improvements to UDDs. For example, in anesthesiology a key adverse-response endpoint is change in blood pressure. Given the volatility of blood pressure readings it seems more sensible to discretize this continuous measure as ordered-ternary Y (decrease, inconsequential change, increase) rather than to dichotomize it. Fortunately, a UDD extension to accommodate ordered-ternary Y will probably be simple and straightforward, like the UDD extensions mentioned in Section 2.2. A more sophisticated potential extension using the full range of ordinal toxicity-grade data was explored briefly 20 years ago in the context of phase I designs, and can also be followed upon [55]. Another potential extension is related to $\text{GUD}_{(3,0,2)}$ which fared well in Section 3.2's cohort-based comparative simulation. As mentioned earlier its balance point is $p^* \approx 0.347$, a tad high if the target is ≤ 0.3. One could propose a modification whereby in case 1 toxicity out of 3 is observed, a biased coin is tossed to determine whether to repeat the dose or de-escalate. Such a variant could target, e.g., the 30th or the 25th percentile, depending upon coin probability. Baldi Antognini *et al.* examined GUD with biased coins some time ago, but their exploration was generic rather than focusing on concrete experimental applications and their specific properties [4]. Last but not least, in sensory studies it is common to run a UDD experiment on a single participant, who repeatedly reports whether they notice a stimulus as its intensity varies up and down. This introduces additional dependence to the observations, as well as "drifts" in response due to fatigue, etc. While the sensory-studies field has been cognizant of these issues, we feel that their impact upon UDD properties and the potential implications for design and estimation have not been studied thoroughly.

Given the paucity of person-hours devoted to UDD methodology in recent decades, even better opportunities surely await the intrepid researcher. A sense of how the methodology has progressed due to the efforts of the few, can be attained by comparing the 2007 Anesthesiology UDD tutorial by Pace and Stylianou [53], the chapter written by us for a 2015 experimental-design handbook [16], and the 2022 Anesthesiology tutorial written by us in collaboration with a senior anesthesiologist [52]. We are thrilled to be in the final stages of completing the first-ever book solely dedicated to UDDs, which contains further developments. We would love to see younger researchers taking up the challenges presented there.

We end on a philosophical note. While it is likely that UDDs had sprung out of common-sense and intuition rather than deep theoretical introspection, they seem to have hit a sweet spot with respect to the handling of uncertainty in a highly constrained, low-information problem. UDDs do not attempt to control the dose-allocation process too tightly; instead, their rules leverage uncertainty to generate random walks with reasonable behavior and good data-collection properties. By contrast, '3+3' and similar escalation designs place very tight constraints on the number of DLTs in the trial. They generally succeed in stopping experiments quickly with few DLTs, but the price is very poor estimation performance, defeating the phase I trial's entire purpose. At the opposite end, Aim-for-Target designs introduce considerable complexity in the attempt to tame uncertainty via repeated estimation, which in practice plays out as declaring a "best dose" early on based on minimal evidence, and sticking with it until proven otherwise. In case this early bet was wrong, correcting it might require longer than the entire experiment's duration. Therefore, as some have suggested in a more general context, it may be possible that letting go just a little bit rather than try to control randomness forcibly, is the winning approach all things considered [62].

ACKNOWLEDGEMENTS

We thank the anonymous reviewers, whose insightful and inquisitive comments have helped to improve the manuscript substantially.

Accepted 5 November 2024

REFERENCES

[1] Anscombe, F. J. (1956). On Estimating Binomial Response Relations. *Biometrika* **43** 461–464. https://doi.org/10.1093/biomet/43.3-4.461. MR0081598

[2] ASTM (1991). Standard test method for estimating acute oral toxicity in rats. *American Society for Testing and Materials*. 1163–90.

[3] Babb, J., Rogatko, A., Rogatko, A. and Zacks, S. (1998). Cancer Phase I Clinical Trials: Efficient Dose Escalation with Overdose Control. *Stat. Med.* **17** 1103–1120.

[4] Baldi Antognini, A., Bortot, P. and Giovagnoli, A. (2008). Randomized group up and down experiments. *Annals of the Institute of Statistical Mathematics* **60** 45–59. https://doi.org/10.1007/s10463-006-0081-5. MR2400060

[5] Brownlee, K. A., Hodges Jr., J. L. and Rosenblatt, M. (1953). The up-and-down method with small samples. *JASA* **48** 262–277. MR0055644

[6] Carter, S. K. (1973). Study design principles in the clinical evaluation of new drugs as developed by the chemotherapy programme of the National Cancer Institute. In *The Design of Clinical Trials in Cancer Therapy* (M. J. Staquet, ed.) 242–289. Editions Scientific Europe, Brussels.

[7] Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2022) *Introduction to algorithms*, 4th Edition. MIT press. MR2572804

[8] Derman, C. (1957). Non-parametric up-and-down experimentation. *Ann. Math. Stat.* **28** 795–798. https://doi.org/10.1214/aoms/1177706895. MR0090956

[9] Diaconis, P. and Stroock, D. (1991). Geometric Bounds for Eigenvalues of Markov Chains. *Ann. App. Prob.* **1** 36–61. MR1097463

[10] Dixon, W. J. and Mood, A. (1948). A method for obtaining and analyzing sensitivity data. *JASA* **13** 109–126.

[11] DOD (2001). *MIL-STD-1751A – Safety and Performance Tests for the Qualification of Explosives (high explosives, propellants, and pyrotechnics)*. United States Department of Defense.

[12] Durham, S. D. and Flournoy, N. (1995). Up-and-down Designs I: Stationary Treatment Distributions. In *Adaptive Designs* (N. Flournoy and W. F. Rosenberger, eds.) 139–157. Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215451483. MR1477678

[13] Durham, S. D., Flournoy, N. and Rosenberger, W. F. (1997). A Random Walk Rule for Phase I Clinical Trials. *Biometrics* **53** 745–760.

[14] Durham, S. D. and Flournoy, N. (1994). Random Walks for Quantile Estimation. In *Statistical Decision Theory and Related Topics, V* (S. S. Gupta and J. O. Berger, eds.) 467–476. Springer-Verlag Inc. MR1286322

[15] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1) 27–38. https://doi.org/10.1093/biomet/80.1.27. MR1225212

[16] Flournoy, N. and Oron, A. P. (2015). Up-and-down designs for dose-finding. In *Handbook of Design and Analysis of Experiments* (D. Bingham, A. M. Dean, M. Morris and J. Stufken, eds.) 24, 862–898. CRC Press, Chapman Hall. MR3699370

[17] Flournoy, N. and Oron, A. P. (2020). Bias Induced by Adaptive dose-finding designs. *Journal of Applied Statistics* **47**(13-15) 2431–2442. https://doi.org/10.1080/02664763.2019.1649375. https://doi.org/10.1080/02664763.2019.1649375. MR4149564

[18] Flournoy, N., Durham, S. D. and Rosenberger, W. F. (1995). Toxicity in Sequential Dose-response Experiments. *Sequential Analysis* **14** 217–227. https://doi.org/10.1080/07474949508836333. MR1365660

[19] Flournoy, N., Moler, J. and Plo, F. (2020). Performance measures in dose-finding experiments. *International Statistical Review* **88**(3) 728–751. https://doi.org/10.1111/insr.12363. MR4180676

[20] Garcìa-Perez, M. A. (1998). Forced-Choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Res.* **38** 1861–1881.

[21] George, R. B., McKeen, D., Columb, M. O. and Habib, A. S. (2010). Up-down determination of the 90% effective dose of phenylephrine for the treatment of spinal anesthesia-induced hypotension in parturients undergoing cesarean delivery. *Anesthesia & Analgesia* **110**(1) 154–158.

[22] Gezmu, M. The geometric up-and-down design for allocating dosage levels (1996). PhD thesis, American University, Washington, DC. MR2695534

[23] Gezmu, M. and Flournoy, N. (2006). Group up-and-down designs for dose-finding. *J Stat. Plan. Inf.* **136**(6) 1749–1764. https://doi.org/10.1016/j.jspi.2005.08.002. MR2255594

[24] Giovagnoli, A. and Pintacuda, N. (1998). Properties of Frequency Distributions Induced by General 'up-and-down' Methods for Estimating Quantiles. *J Stat. Plan. Inf.* **74** 51–63. https://doi.org/10.1016/S0378-3758(98)00076-7. MR1665120

[25] Gorla, C., Rosa, F., Conrado, E. and Concli, F. (2017). Bending Fatigue Strength of Case Carburized and Nitrided Gear Steels for Aeronautical Applications. *International Journal of Applied Engineering Research* **12**.

[26] Heijmans, R. (1999). When does the expectation of a ratio equal the ratio of expectations? *Statistical Papers* **40** 107–115. https://doi.org/10.1007/BF02927114. MR1668879

[27] Hughes, B. D. (1995) *Random Walks and Random Environments. Vol. 1. Oxford Science Publications*. The Clarendon Press Oxford University Press, New York. Random walks. MR1420619

[28] Iasonos, A. and O'Quigley, J. (2014). Adaptive dose-finding studies: a review of model-guided phase I clinical trials. *Journal of Clinical Oncology* **32**(23) 2505–2511.

[29] ISO (2012). *International Organization of Standardization*. 12107 Metallic materials–Fatigue testing–Statistical planning and analysis of data. Geneva.

[30] ISO (2016). *International Organization of Standardization*. 14801 Dentistry–Implants–Dynamic loading test for endosseous dental implants. Geneva.

[31] Ivanova, A., Flournoy, N. and Chung, Y. (2007). Cumulative cohort design for dose-finding. *J Stat. Plan. Inf.* **137** 2316–2317. https://doi.org/10.1016/j.jspi.2006.07.009. MR2325437

[32] Ivanova, A., Montazer-Haghighi, A., Mohanty, S. G. and Durham, S. D. (2003). Improved Up-and-down Designs for Phase I Trials. *Stat. Med.* **22**(1) 69–82.

[33] Ji, Y., Liu, P., Li, Y. and Nebiyou Bekele, B. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical Trials* **7**(6) 653–663.

[34] JSME (1981). *Standard method of statistical fatigue testing*. Japan Society of Mechanical Engineers, Japan. JSME S 002.

[35] Langlie, H. J. (1962). A Reliability Test Method for "One-Shot" Items. Technical Report No. U-1792, Ford Motor Company, Ford Motor Company Aeronautics Division. https://apps.dtic.mil/sti/citations/tr/ADP014612.

[36] Lee, S. M. and Cheung, Y. K. (2009). Model Calibration in the continual reassessment method. *Clinical Trials* **6** 227–238.

[37] Lee, S. M. and Cheung, Y. K. (2011). Calibration of prior variance in the Bayesian continual reassessment method. *Stat. Med.* **30** 2081–2089. https://doi.org/10.1002/sim.4139. MR2829158

[38] Liu, S. and Yuan, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C: Applied Statistics* 507–523. https://doi.org/10.1111/rssc.12089. MR3325461

[39] Maeda, A., Villela-Franyutti, D., Lumbreras-Marquez,

M. I., Murthy, A., Fields, K. G., Justice, S. and Tsen, L. C. (2023). Labor analgesia initiation with Dural puncture Epidural Versus Conventional Epidural techniques: a Randomized biased-Coin Sequential Allocation Trial to determine the effective dose for 90% of patients of Bupivacaine. *Anesthesia & Analgesia* 10–1213.

[40] Morris, M. D. (1988). Small-Sample Confidence Limits for Parameters under Inequality Constraints with Application to Quantal Bioassay. *Biometrics* **44** 1083–1092. https://doi.org/10.2307/2531737. MR0981001

[41] Narayana, T. V. Sequential procedures in probit analysis (1953). PhD thesis, University of North Carolina. MR2938682

[42] NATO (1999) *STANAG 4489 – Explosives, impact sensititity test.* North Atlantic Treaty Organization.

[43] NIEHS (2001). The revised up-and-down procedure: A Test method for Determining the Acute Oral Toxicity of Chemicals. Technical Report No. 2-4501, Washington D.C.

[44] Novik, G. P. and Christensen, D. (2024). Increased impact sensitivity in ageing high explosives; analysis of Amatol extracted from explosive remnants of war. *Royal Society open science* **11**(3) 231344.

[45] OECD (2022) *Test No. 425: Acute Oral Toxicity: Up-and-Down Procedure.* https://www.oecd-ilibrary.org/content/publication/9789264071049-en.

[46] O'Quigley, J., Pepe, M. and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 33–48. https://doi.org/10.2307/2531628. MR1059105

[47] Oron, A. P. (2017). Up-and-Down Designs Enhanced with SPRT Rules for Phase I Cancer Trials. In *Society for Clinical Trials Annual Meeting, Liverpool*. SCT.

[48] Oron, A. P. and Hoff, P. D. (2013). Small–sample behavior of novel Phase I cancer trial designs. *Clinical Trials* **10**(1) 63–80.

[49] Oron, A. P. and Flournoy, N. (2017). Centered isotonic regression: point and interval estimation for dose-response studies. *Journal of Biopharmaceutical Statistics*.

[50] Oron, A. P. and Hoff, P. D. (2009). The k-in-a-row up-and-down design, revisited. *Stat. Med.* **28**(13) 1805–1820. https://doi.org/10.1002/sim.3590. https://doi.org/10.1002/sim.3590. MR2751599

[51] Oron, A. P., Azriel, D. and Hoff, P. D. (2011). Dose–finding designs: The role of convergence properties. *Int. J Biostat.* **7**(1) 39. https://doi.org/10.2202/1557-4679.1298. MR2873999

[52] Oron, A. P., Souter, M. J. and Flournoy, N. (2022). Understanding research methods: Up-and-down designs for dose-finding. *Anesthesiology* **137**(2) 137–150.

[53] Pace, N. L. and Stylianou, M. P. (2007). Advances in and Limitations of Up-and-down Methodology: A Précis of Clinical Use, Study Design, and Dose Estimation in Anesthesia Research. *Anesthesiology* **107**(1) 144–152.

[54] Parasuraman, S. (2011). Toxicological screening. *Journal of Pharmacology and Pharmacotherapeutics* **2**(2) 74–79. https://doi.org/10.4103/0976-500X.81895.

[55] Paul, R. K., Rosenberger, W. F. and Flournoy, N. (2004). Quantile estimation following non-parametric phase I clinical trials with ordinal response. *Statistics in medicine* **23**(16) 2483–2495.

[56] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statistics* **22** 400–407. MR0042668 (13,144j).

[57] Shi, L., Khalij, L., Gautrelet, C., Shi, C. and Benasciutti, D. (2024). Two-phase optimized experimental design for fatigue limit testing. *Probabilistic Engineering Mechanics* **75** 103551. https://doi.org/10.1016/j.probengmech.2023.103551.

[58] Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)* 310–313. MR0637943

[59] Sørensen, C. B., Adams, T. B., Pedersen, E. R., Nielsen, J. and Schmidt, J. H. (2023). AMTAS^{TM} and user-operated smartphone research application audiometry—An evaluation study. *Plos one* **18**(9) 0291412.

[60] Stylianou, M. and Flournoy, N. (2002). Dose Finding Using the Biased Coin Up-and-down Design and Isotonic Regression. *Biometrics* **58**(1) 171–177. https://doi.org/10.1111/j.0006-341X.2002.00171.x. MR1891376

[61] Takano, T., Yoshinari, M., Sakurai, K. and Ueda, T. (2024). Cyclic Fatigue Properties of Titanium Alloys for Application in Dental Implants. *The Bulletin of Tokyo Dental College* 2023-0025.

[62] Taleb, N. N. (2001) *Fooled by Randomness.* Random House, New York.

[63] Treutwein, B. (1995). Minireview: adaptive psychophysical procedures. *Vision Res.* **35** 2503–2522.

[64] Tsutakawa, R. K. (1967). Asymptotic Properties of the Block Up-and-down Method in Bio-assay. *Ann. Math. Stat.* **38** 1822–1828. https://doi.org/10.1214/aoms/1177698615. MR0217951

[65] von Békésy, G. (1947). A new audiometer. *Acta Oto.Laryn.* **35** 411–422.

[66] Watson, A. B. and Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception & Psychophysics* **33** 113–120.

[67] Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *J Roy. Stat. Soc. B* **25** 1–48.

[68] Wetherill, G. B. and Levitt, H. (1965). Sequential estimation of on a psychometric function. *Brit. J Math. Stat. Psych.* **18** 1–10.

[69] Woolf, B. (1955). On Estimating the Relation between blood group and disease. *Annals of Human Genetics* **19** 251–253.

[70] Wu, C. F. J. (1985). Efficient Sequential Designs with Binary Data. *Journal of the American Statistical Association* **80**(392) 974–984. MR0819603

[71] Zhao, H., Li, X., Tang, N., Jiang, X., Guo, Z. and Lin, H. (2018). Dielectric properties of fluoronitriles/CO2 and SF6/N2 mixtures as a possible SF6-substitute gas. *IEEE Transactions on Dielectrics and Electrical Insulation* **25**(4) 1332–1339. https://doi.org/10.1109/TDEI.2018.007139.

Assaf P. Oron. Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA. E-mail address: assaf@uw.edu

Nancy Flournoy. University of Missouri System, Columbia, MO, USA. E-mail address: flournoyn@umsystem.edu