Supplement to Up-and-Down "Understanding Research Methods" Article

Assaf P. Oron, Michael J. Souter and Nancy Flournoy

Table of Contents

UDD and Dose-Finding Glossary (alphabetically ordered)

Note: the glossary includes only terms specifically relevant to UDDs and dose-finding. Definitions of other statistical terms can be readily found online.

Adaptive Designs: procedures in which the rule for allocating subjects to treatments changes during the study, and current allocations may depend on prior data. In addition to UDDs and other dose-finding designs, this includes adaptive randomization rules for controlled clinical trials, designs with stopping rules that depend on the data, and those that plan for sample size re-estimation during the course of the study.

Boundary Doses: the highest and lowest dose permitted in a specific experiment.

Dose-Response Curve: for binary outcome variables, this is a curve with the probability of a positive response on the y-axis and the dose magnitude on the x-axis.

Dose-Transition Rules or **Dose-Allocation Rules:** these rules determine the dose allocated to the next patient, given the doses and responses of previous patients.

Monotone Dose-Response Relationship: in a monotone relationship as the dose increases, the probability of positive response never changes direction: it either increases or decreases throughout the entire dose range.

Reversal Points: points in a UDD experiment where the response changed from positive to negative, or vice versa.

Target Dose: the dose that a dose-finding experiment is formally tasked with estimating. Most dose-finding designs aun to concentrate dose-allocations around target.

Using R's 'cir' Package for UDD Target Estimation

R is a free open-source, cross-platform statistical programming language, and also currently the world's most popular statistical language. One of R's strengths is that in addition to its core capabilities bundled with the standard installation, there are numerous, easily-installed add-on packages contributed by researchers and software experts. R is available from the Comprehensive R Archive Network (CRAN: https://cran.r-project.org/). Despite its ease of installation, using R proficiently requires some understanding of programming, because its standard interface is a command line where the user types programming statements, or invokes a file with a sequence of such statements (known in programming as a "script"). Most current R users prefer to interact with the language via the interface provided by another free software known as Rstudio.

The 'cir' package (authored and maintained by Dr. Oron) provides functions that produce centered isotonic regression (CIR) and "plain" isotonic regression estimates with their confidence intervals, as well as basic utilities for handling and plotting dose-response data such as those from UDD experiments.[1] The package incorporates the learnings documented in recent methodological articles, in particular, Oron and Flournoy presenting CIR and its confidence interval[1,2] and Flournoy and Oron describing the bias induced by dose-finding designs.[3]

To install 'cir', either use drop-down menu installation utilities of the R or RStudio user interfaces, or open an R session and type

```
install.packages("cir")
```

(Note: code to be typed is in blue, whereas comments are in black and preceded by a hash '#')

The command above will install the latest stable version the maintainer has uploaded to CRAN. All software undergoes continuous changes and improvements, usually minor. If for any reason, you would like to install the current "live" version Dr. Oron is working on, this version is shared on the GitHub public repository, and can be installed via

```
install.packages("devtools")
# Comment: this will install a package enabling installation from GitHub.
# Just like 'cir' itself, you can install 'devtools' using drop-down menus.
library(devtools)  # this actually loads the 'devtools' package
```

```
install_github("assaforon/cir")
```

Before doing the latter, it is a good idea to contact Dr. Oron (current emails: assaf.oron@gmail.com or assaf@uw.edu) and make sure the live version is not in an unusable *"under construction"* state.

Once 'cir' is installed, you can load it into an active R session via the command

```
library(cir)
```

A tutorial (called "vignette" in R jargon) showing the use of 'cir' with UDD data from Benhamou et al.,[4] is available here. This tutorial is bundled with the package's CRAN version. If you install 'cir' from GitHub you will get the tutorial's latest version (as of spring 2022 there have been no recent changes to it). After loading 'cir', you can see and click into any of its vignettes using the command

```
browseVignettes("cir")
```

Here we only show in brief, how to carry out a UDD target estimate.

The target estimation function is called **`quickInverse()`** [*"Inverse"* because this is an inverse estimate from the dose-response curve, that is, estimating a dose (x) value from the response (y) value]. It accepts many input formats, but the preferable format is an *x-y-n* summary by dose called a **`doseResponse`** object. Here is this object for Benhamou et al.'s ropivacaine arm data:

```
x          y      weight
0.07 0.0000000       3
0.08 0.3750000       8
0.09 0.3846154      13
0.10 0.8000000      10
0.11 0.7500000       4
0.12 1.0000000       1
```

In this case, the 'weight' column is simply the number of observations (n) made at each dose. The vignette shows how to create such an object from the raw data. Assuming we have already created it under the name **`bhamou03ropiDR`**, the recommended code for estimating the target (the ED50 in this illustration) is

```
ropiTargetCIR = quickInverse(bhamou03ropiDR, target=0.5, adaptiveShrink=TRUE)
```

```
# You should get this back as the result:

      target      point             lower90conf upper90conf

1     0.5         0.09383622  0.08090006   0.1060014
```

Notes:

- The target estimate appears under the word `point`: in this case, 0.0938% (after adding units).
- The probability of a positive response, a fraction between 0 and 1, defines the target via the `target` argument. For finding the ED50, the value is 0.5, as above.
- The "`adaptiveShrink`" option performs an empirical correction for bias induced by the adaptive design, as mentioned in the main article.[2] The bias is minimal near the target, but increases rapidly away from it. We perform the correction mainly in order to expand the confidence intervals, since the bias makes dose-response slope estimates too steep.
- Because of this bias, **we strongly recommend to *not* report any dose estimates except for the UDD's designated target**.
- The default confidence interval is 90%, which as explained in the main article, is the maximum confidence level that should be reported unless you use a very large n. To change it use the "`conf`" argument, which is specified as a fraction (i.e., the default is `conf=0.9`).
- **In an experiment that does not target the ED50, but targets another dose such as the ED90, we recommend adding the argument "`adaptiveCurve=TRUE`" to the command above.** This broadens the confidence intervals a bit more to ensure proper coverage, which is more challenging the closer the target is to the edge of the dose-response curve.
- To use the older IR method instead of CIR, add the argument "`estfun=oldPAVA`".

If you would like to obtain the CIR-estimated dose-response curve, then the function **`quickIsotone()`** with the same arguments will give you the y estimates at all doses used. To get data for generating the complete x-y curve, add the argument `full=TRUE`. Note again that

outside the target area, this curve will be biased (upward at doses above the target and vice versa), and the empirical bias correction only partially corrects the bias. See the vignette and other `'cir'` package help pages for more information on manipulating and plotting UDD data.

Comparing UDD Estimation Methods: Performance Simulations

A. Overview

This section demonstrates how design and estimator performance are examined in practice, and also provides some evidence for statements made about estimation performance in the main article.

Most theoretical results about the performance of designs and estimates have to do with long-term behavior as *n* approaches infinity. In the realm of actual sample sizes used in UDD experiments, theoretical results are very rare and one has to resort to comparative simulations. Using simulation to evaluate UDDs has a history almost as long as the design family itself.[7] To run such a simulation we need to decide upon

- The designs and/or estimation methods to compare
- Design parameters such as *n*, number of dose levels, starting dose, etc.
- The form of F(x)

The last decision is all-important. Some choices of F(x) might favor one design or estimate, but the very same design or estimate would do very poorly given another choice. Until recently, researchers usually chose a few specific curves to use in their simulations. These curves, besides falling far short of representing the broad range of conditions encountered in practice, also tended to favor whatever method the researcher performing the simulation had favored. Remarkably, Wetherill himself in his seminal 1960s work decided that a reversal-only estimate is better than using all doses, based on a narrow performance advantage in his simulations on a very specific F(x).[5] The extreme narrowness of this body of evidence was quickly forgotten, or misunderstood, as reversal-only averages became popular, and to this day remain the most popular choice. Ironically, for reversal-only estimates one can obtain *theoretical* results suggesting they are generally inferior to using all doses, so simulations are not as necessary.[6,7]

More recently, in order to make simulations more robust, the dose-finding field has moved towards *random-curve simulations.*[2,8,9] In this approach we let curve parameters, or sometimes the entire curve, be chosen randomly, generating a large and diverse **ensemble** of curves. For each curve, we "run" and estimate a simulated UDD experiment under various choices for the other design elements.

Figure S1 shows some curves generated to examine estimates for the ED50 (left) and the ED_{90} (right) with 10 dose levels. The left panel of curves come from the asymmetric Gamma family; the right panel from the symmetric Logistic family. All curves are constrained to cross their respective targets (y=0.5 and 0.9, respectively) between doses 5 and 6. This enables one to separate the resilience of estimates to curve shapes, from their resilience to starting-dose and boundary effects: we can shift the curves and starting dose left or right to impose different relations between starting dose, target and boundary. For each such shift, the entire ensemble is run and the results tallied.

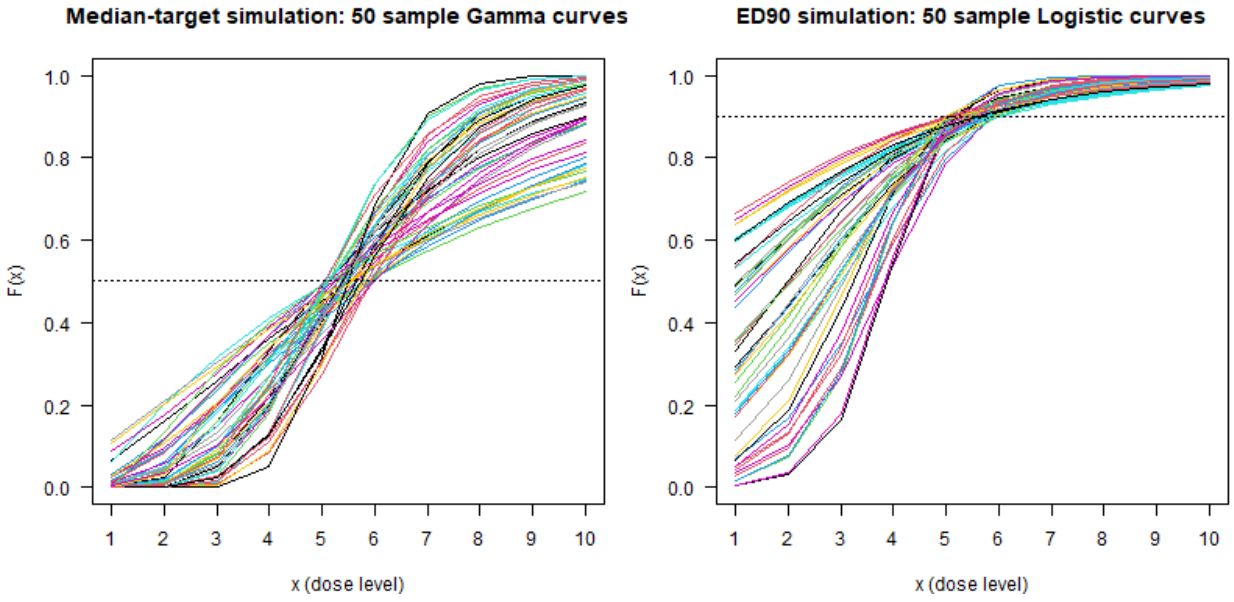

Figure S1: a selection of randomly simulated dose-response curves used in our performance simulations. In each pane, the target response rate is shown by a horizontal dotted line.

B. Comparing Estimation Methods

The most standard estimation performance metrics are **the mean-squared error (MSE)** and its square root, the root-mean-squared error (RMSE). As the name indicates, the MSE is the

average of squared distances between the estimate and the targeted dose. The MSE follows a famous and insightful decomposition formula:

$$MSE = (Bias)^2 + (SD)^2,$$

where SD^2 is the variance of the estimate. Methods that focus only on eliminating bias while letting estimators run very noisily (large SD), or vice versa, will not fare well. To obtain a good estimate, both components need to be controlled, and the formula suggests that they are equally important. Some methods deliberately take on a small amount of bias, in order to achieve a larger reduction in noise. This interplay is known in statistics and machine learning as *the Bias-Variance Tradeoff*.[10] The formula also reveals that for the MSE, the bias direction does not matter - only its magnitude.

Despite the equivalent weighting of bias and variance in the MSE, in many applications and often in medicine, the bias is important on its own. Bias is a systematic error. If our method for estimating the ED90 actually produces, on average, something closer to the ED80, it is important to know that, and will often be less desirable than a noise error component of equal magnitude. Therefore, in most clinical dose-finding contexts the bias can serve as a tie-breaker when selecting between methods with very similar MSE.

In the figures below we show the RMSE and the absolute bias rather than the MSE and squared bias, because the former are in the same units as the doses. We compare seven estimation methods:

- The original 1948 Dixon-Mood estimate ("Dixon-M"),[11] a dose-averaging estimate often cited in anesthesiology as "Dixon-Massey" after the 1950s textbook where it also made an appearance.[12]

- Averaging all doses starting from the first reversal ["Avging (all from R1)"], or from the third ["Avging (all from R3)"]. The latter is the one we recommend, if you want to calculate a secondary averaging estimate in conjunction with CIR.

- Averaging only the doses at reversals, starting with the first ["Avging (reversals only)"]. **This is Wetherill's estimation method;** as mentioned above, it is still likely the most popular across the entirety of UDD studies in various fields.

- An averaging estimator attempting to detect the "right" truncation point, up to which all preceding doses are excluded ('Avging ("Auto-Detect")']. All doses following the

truncation point are included in the average. It has not been published outside Dr. Oron's dissertation.[5]

- CIR and IR as described in the main article, both using the empirical adaptive-design bias correction.

Figure S2 compares RMSEs of these estimates on ED50-targeting experiments with M=10 dose levels and n=30. Dots are color-coded by the approximate distance from the starting dose to the target. Each dot represents an ensemble of 1000 virtual "experiments", each with its own random F(x) under the Gamma family (top panel) and the Logistic family (bottom panel), starting-dose and target-location specifications.

Estimates, on average (horizontal black bars), are slightly less than a single dose-spacing from the target, but not much closer except under very favorable starting choices (the blue dots). In terms of average performance, the "R3" and "auto-detect" averaging estimators seem slightly better than several others, i.e., their average RMSE is smallest - while standard isotonic regression has the largest average RMSE. However, the differences are not overwhelming, with all seven methods within 15-20% of each other.

As often happens, the average does not tell the full story. All estimates do better when things line up nicely together: the target is near the starting dose and there's no boundary effect (blue dots). However, CIR and isotonic regression are far more resilient, or **robust**, to less desirable settings (red dots). Among averaging estimates, "all from R3", and "auto-detect", are more robust than others, while the two estimates most commonly appearing in literature (Dixon-Mood and reversals-only) are the *least* robust.

Figure S3 shows bias magnitudes from the same simulation and estimates. It sheds light on the source of averaging estimates' vulnerability, which indeed is the bias, most often in the direction of the starting dose. When the target is far from the starting dose (red dots), the bias component might contribute to the MSE more than the variance. By contrast, CIR and isotonic regression biases are generally small; most of their estimation error is noise.

Since CIR is competitive with the leading dose-averaging estimates on RMSE, its substantially lower bias provides a "tie-breaker", substantiating our case for using it as the standard UDD estimate; at least until something better is developed.

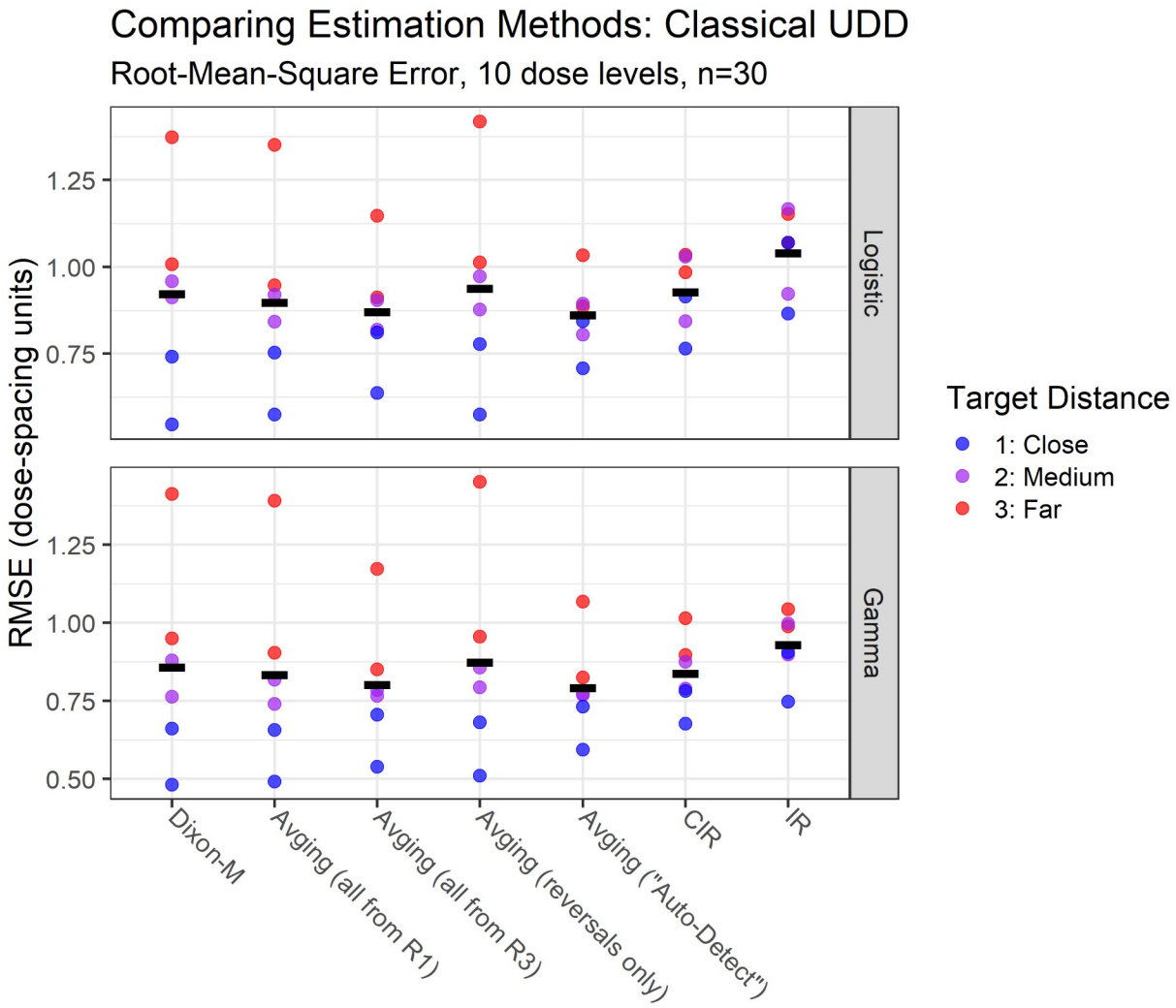

Figure S2: RMSE summaries from a random-curve simulation for ED50-targeting UDDs. Each dot represents an average across 1000 random curves. The horizontal black lines show the average of the dots for each estimate. Dots are color-coded by the approximate distance from the starting dose to the target.

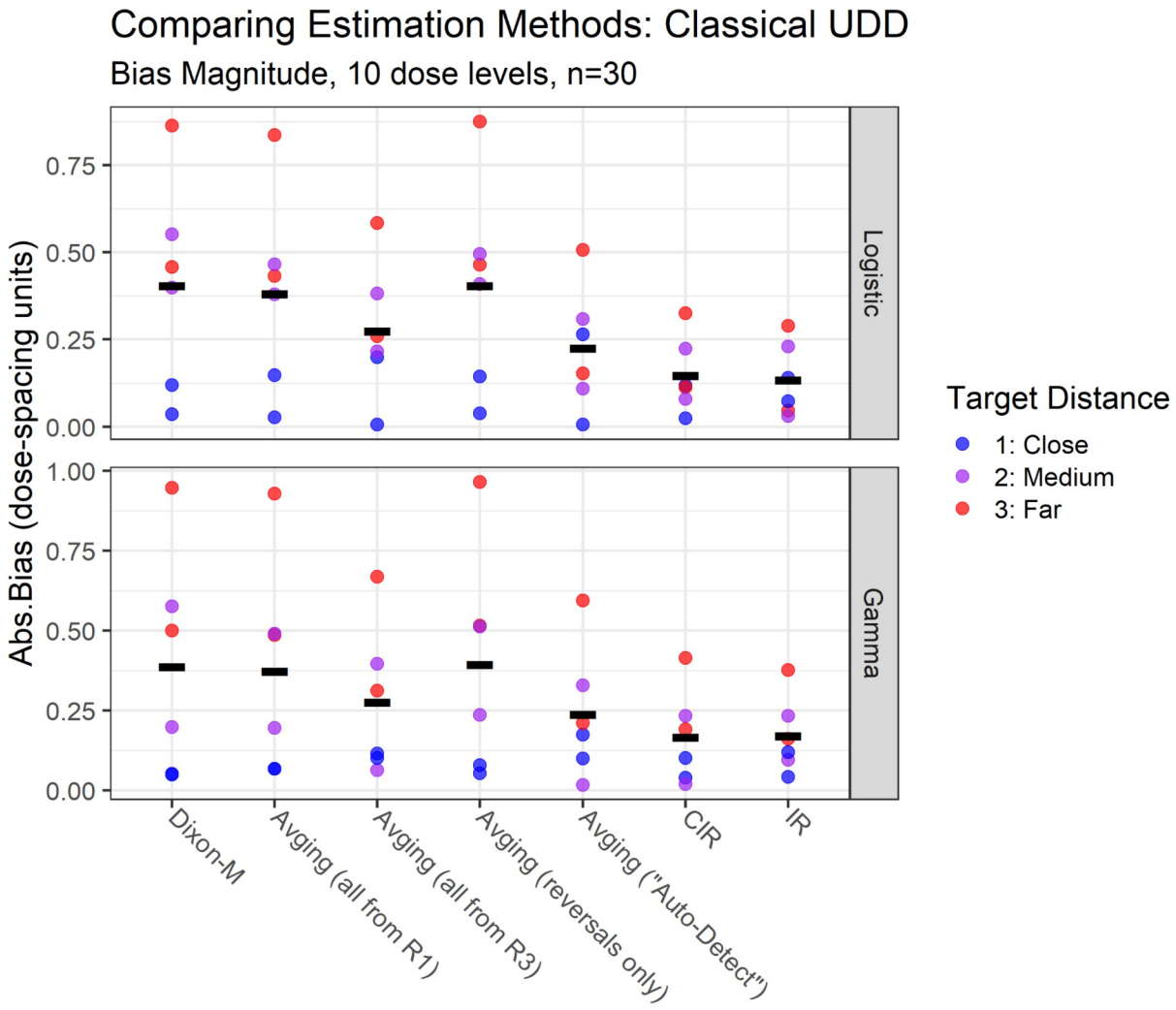

Figure S3: Summaries of absolute biases from the random-curve simulation for ED50-targeting UDDs. Each dot represents an average across 1000 random curves. The horizontal black lines show the average of the dots for each estimate. Dots are color-coded by the approximate distance from the between starting dose to the target.

Figures S4-S5 show analogous summaries with ED90-targeting KRDs (k=6) and n=50. Two estimation methods were clearly tailored for ED50-finding, and fare very poorly when used off-target here: It is not surprising that one is the Dixon-Mood estimator, but the other is the reversals-only average which unfortunately is used very often with KRD in sensory studies, despite being rather inadequate for the task.

While the other averaging estimators exhibit fairly good RMSE and bias performance for the settings used in our simulations, at present we cannot offer any reasonable confidence interval to accompany them for estimating non-ED50 targets. Some more details are provided further below.

By contrast, in these simulations CIR's 90% CIs had 85-90% coverage for both targets, after adding the **`"adaptiveCurve=TRUE"`** argument.

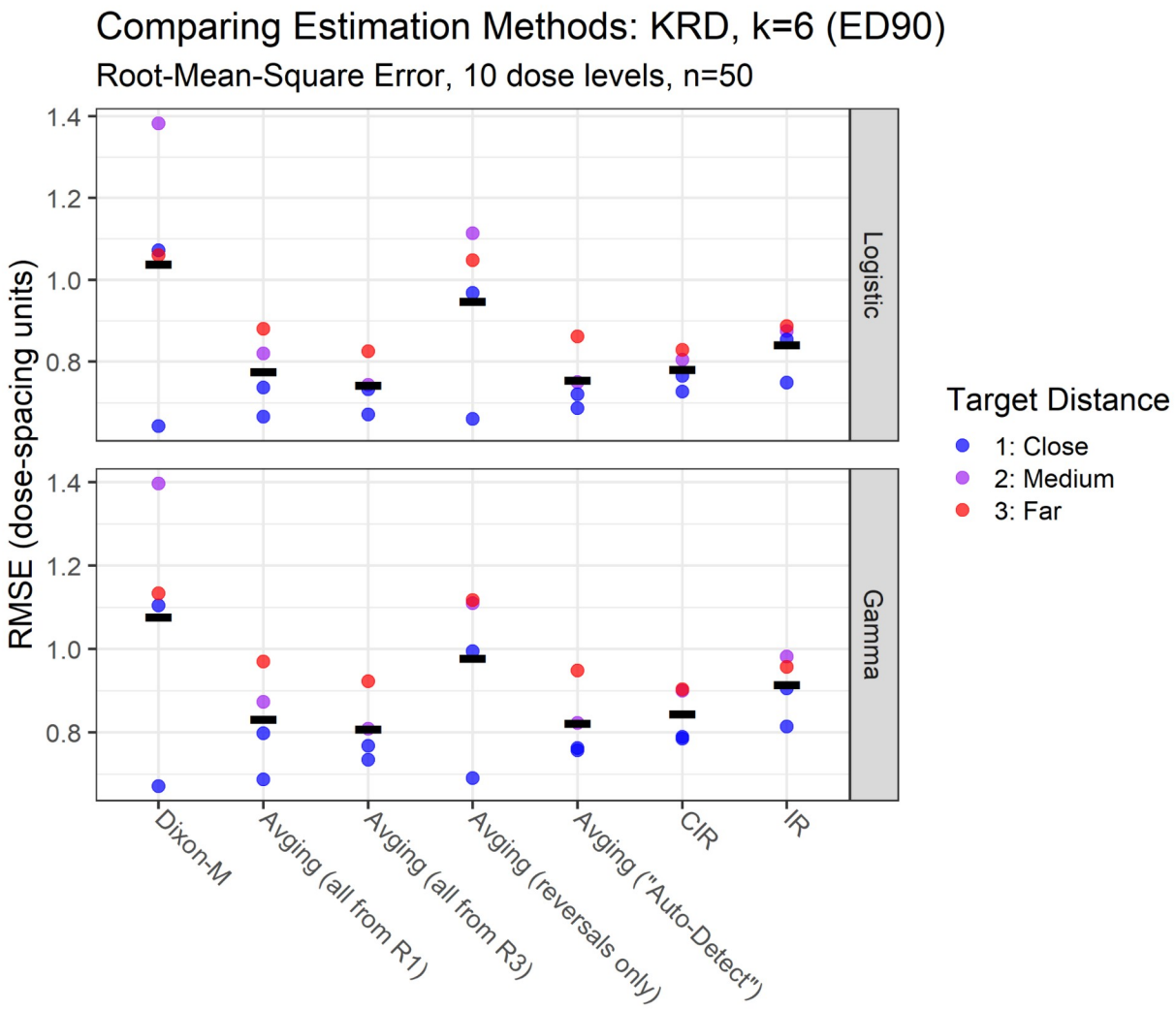

Figure S4: RMSE summaries from the random-curve simulation from KRDs with k=6, targeting the ED90. Each dot represents an average across 1000 random curves. The horizontal black lines show the average of the dots for each estimate. Dots are color-coded by the approximate distance from the starting dose to the target.

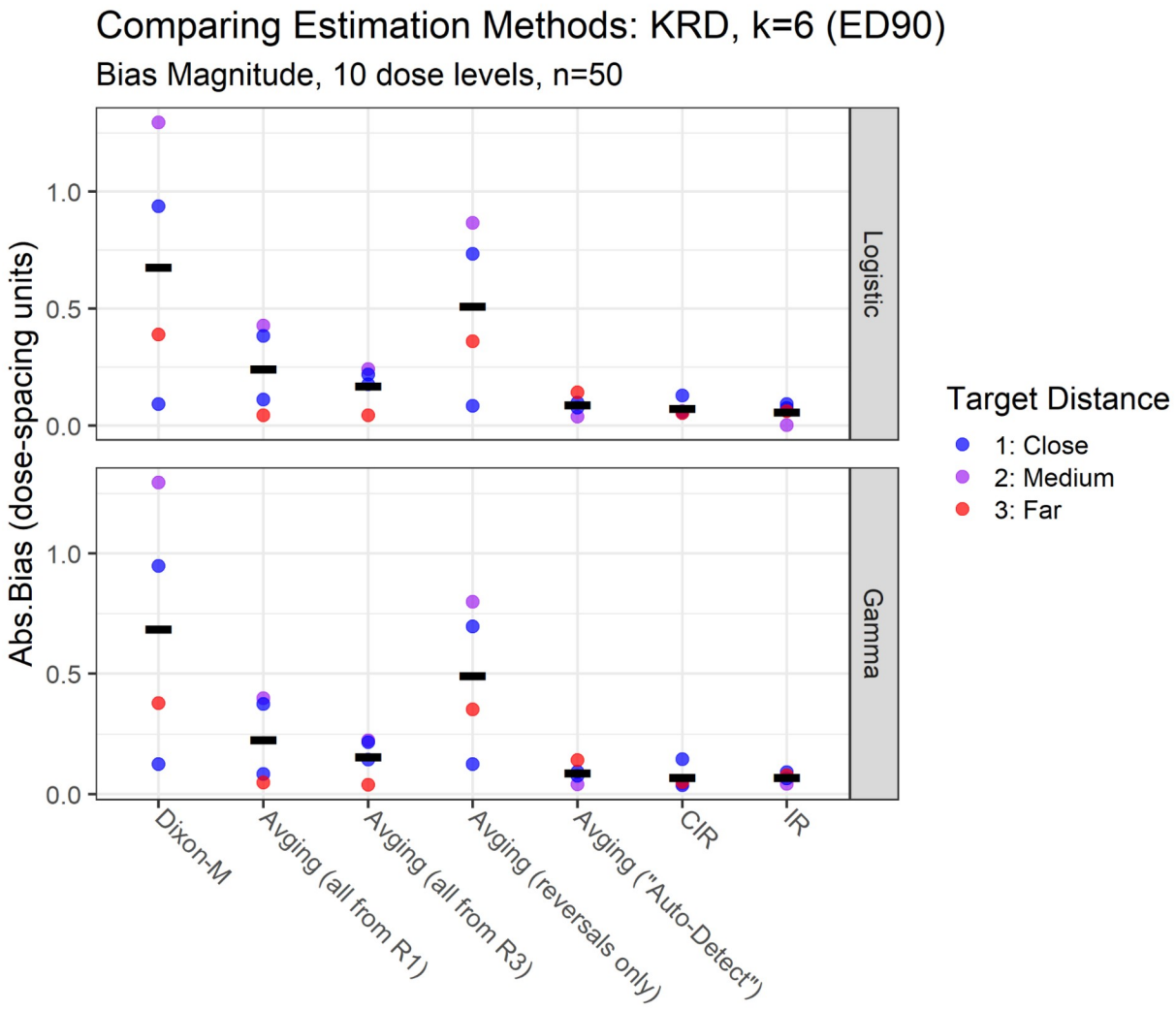

Figure S5: Absolute bias summaries from the random-curve simulation from KRDs with k=6, targeting the ED90. Each dot represents an average across 1000 random curves. The horizontal black lines show the average of the dots for each estimate. Dots are color-coded by the approximate distance from the starting dose to the target.

C. Comparing Design Choices and Decisions

We briefly summarize some simulation insights about additional design aspects, without showing additional figures.

If the target is far from the starting dose, a number of initial patients may be "used" to get the dose-allocation sequence near to the target. One possible procedure for getting allocations close to a non-ED50 target with fewer patients is to start with the classical UDD and then switch to an appropriate KRD. This idea for accelerating initial dose-allocations toward the target is mentioned in the main article's Box 2, and is discussed in more detail below.

To evaluate this two-stage procedure for targeting the ED90, we compared the performance of a standard k=6 KRD, with a procedure that starts with the k=1 KRD (equivalent to the Classical UDD) and switches to k=6 after the first negative response. Indeed, when starting from the top of the dose range there is a substantial performance improvement. Starting from the top with k=6 wastes an enormous number of observations before getting near the target. RMSEs with CIR were ~20% smaller when using the k=1 startup stage. In addition, in 10% of runs without a startup stage no CIR estimate was possible, because all the dose levels visited had observed response rates above 90%. When instituting the k=1 startup rule, only ~1% of runs suffered this fate.

However, when the starting dose is not high, there is a price to be paid for starting with a k=1 rule. We found that when the target is near or somewhat below the starting dose, a k=1 startup rule may actually incur a small increase to CIR's RMSE.

We also ran a large battery of ED95 targeting experiments. Confidence-interval coverage was 5-10% lower: 80% or less for the 90% CI. The results indicate that such an extreme target will require a larger *n* (approaching n=100 or more) to obtain acceptable coverage, and to allow for some movement between doses.

As to design comparison: we had compared KRD and BCD in 2009 targeting the 30th percentile, inspired by Phase I cancer trials.[13] The designs for that target would be the mirror-image of designs targeting ED70. Specifically, we compared KRD with k=2 to the BCD with a coin probability of 3/7. In that comparison, KRD clearly had the upper hand, with consistently smaller RMSEs.

To our surprise, when targeting ED90 in simulations for this Supplement, the two designs appear equivalent, with CIR RMSEs within a few percent of each other in either direction, and no clear winner. To cross-check our prior work, we ran a refreshed version of the ED30 simulations using the current setup and code; KRD still showed the advantage we previously found, although it was more modest than before (now KRD had ~5-8% smaller RMSEs, versus 15%-20% smaller in older studies). It may be that the various improvements to CIR since our 2009 study have helped close the gap. While KRD has the operational advantage of no randomization and a fixed, generally shorter maximum wait before decreasing a dose (6 responses with ED90, versus 9 for BCD on average), we can report no difference between the two in terms of ED90 estimation performance.

Please see additional simulation-based insights below, when discussing CIs for averaging estimators.

Preferred Methods for Dose-Averaging ED50 Estimates

A. **Target estimate, with R functions**

Dose-averaging estimates are fairly straightforward, but as a courtesy we provide an annotated version of the R functions we have written and used to calculate them for these simulations. As the simulation results above suggest, dose-averaging should be discouraged when the target is not the ED50. A function for confidence intervals appears further below.

```
### First, a small utility that identifies points with reversals
# The function returns indices, i.e., the locations of reversals

reversals <- function(y) which(diff(y)!=0)+1

# ------- MAIN FUNCTION ---------
# Reversal-anchored averaging estimators: enables both reversal-only
#        and all doses Starting from a reversal
# Default is our recommendation: all doses, starting from reversal 3

reversmean <- function(x, y, rstart=3 ,all=TRUE, before=0, full=FALSE)

# Arguments:
# x - the sequence of doses given
# y - the sequence of 0/1 responses
# rstart - which reversal to begin with (integer, default 3)
# all - logical, whether to include all doses (TRUE, default), or
#            only reversals
# before - If set to 1, will start 1 data point before the reversal
```

```
#             (default 0)
# full - logical, whether to return a fuller report or
#            only the estimate (FALSE, default)
{
### Validation checks
n = length(x)
if(!(length(y) %in% c(n-1,n))) stop('X vector must be equal-length or
1 longer than Y.\n')
if(!(before %in% 0:1)) stop('argument before can only be 0 or 1.\n')

# Locating the reversals using the reversals() utility
revpts = reversals(y)

### Exception handling
if(length(revpts) == 0) { # fully degenerate, no reversals
     if(full) return(data.frame(est=mean(x[-1]),cutoff=1))
     return(mean(x[-1]))
}
# part-degenerate: fewer reversals than needed to start
if(rstart>length(revpts)) rstart = length(revpts)

### After all this, the estimate itself is anti-climactic:
est = ifelse(all, mean(x[(revpts[rstart]-before):n]),
         mean(x[revpts[rstart:length(revpts)]]))
if(!full) return(est)
data.frame(est=est, cutoff=revpts[rstart]-before)
}
```

B. CI estimate, with R function `avgHalfCI`

Confidence intervals for dose-averaging UDD estimates are obtained via an estimate of the standard error of the mean (SEM):

$$SEM = \frac{SD}{\sqrt{n_{eff}}},$$

where SD is (an estimate of) the standard deviation of the dose-allocation distributions, and n_{eff} estimates the effective sample size. In practice estimating the SEM from UDD data is a very challenging task because

- The sequence of assigned doses is positively auto-correlated, which complicates the estimation of both numerator and denominator in the formula.
- The doses are assigned from a limited discrete set of dose-levels, and therefore estimated quantities such as the observed SD and the auto-correlation are coarse in that they can only take a discrete set of possible values.

- UDD samples are usually small, and for dose-averaging early results are excluded to remove the starting dose effect.

To our knowledge, the last theoretical attempt to develop a dose-averaging SEM cognizant of these challenges was by Choi in 1990, taking account of the autocorrelation as well.[14] We examined it with our random-curve simulations, and it fails to capture the true SD variations between curves. Therefore it produces CIs with very poor coverage.

The method in the R function `avgHalfCI` below relies upon a simple robust and conservative alternative to the SD: half the difference between the 90th and 10th percentile of allocated doses. The user can vary this choice of percentiles chosen via trial and error, but the formula above was found to produce sufficient Cis for ED50 estimates.

To obtain the SEM, one must still estimate the denominator n_{eff}. It will be smaller than the nominal sample size, both because of excluding the early observations and even more so, because of the autocorrelations. Like Choi, we rely upon standard random-walk theory to meet this challenge: if one splits the sequence of assigned doses by visits to the same dose-level, each of the sub-sequences between these visits is independent of the other. This suggests that the number of visits to the most-visited dose level, less 1, is a reasonable approximation of n_{eff}.

Warning: the `avgHalfCI` function provided below is to be used only for ED50-finding. In the simulation setups presented earlier for ED50-finding, 90% CIs using these functions had empirical 85-90% coverage for the "all from R3" estimate. **However, for the ED90 the same approach's 90% CIs barely exceeded 50% coverage.**

Here is the `avgHalfCI` function with some documentation. As default it returns the CI's half-width:

```
avgHalfCI <- function(x, conf=0.9, refq=c(.1,.9), full=FALSE)

# Returns half the dose-averaging confidence-interval.
# Uses visits to most-visited-level as proxy for n_eff
#    and quantiles of assigned doses as proxy for the SD
# The CI will be symmetric, obtained by adding and subtracting from
#    the point estimate.
# We assume the t distribution.
# WARNING: TO BE USED ONLY FOR ED50-FINDING!
#   DOSE-AVERAGING IS ILL-SUITED FOR OTHER TARGET DOSES.
```

```
# Arguments:
# x: the dose sequence to be included. You should exclude the early
part that is not participating in the dose-average.
# conf: confidence level as a fraction between 0 and 1.
# refq: the reference quantiles (also between 0 and 1) between which 2
times SD is estimated. It is recommended to have them symmetric around
0.5.
# full: a logical flag alternating between
#       reporting only the half-CI width (full=FALSE) and
#       reporting the underlying estimates of n_eff and SD (full=TRUE)
{

neff = max(table(x))-1 # effective n via number of visits

sdeff = diff(quantile(x, probs=refq, type=6))/2 # type=6 is unbiased

if(!full) return( qt(0.5+conf/2, df=neff-1) * sdeff/sqrt(neff) )

return( data.frame(neff=neff, sdeff=sdeff) )
}
```

The Impact of Boundaries on Dose-Allocation Distributions and Estimates

The main article states that hard dose boundaries near the target break the approximate symmetry of the allocated dose distribution. This has a devastating effect on dose-averaging estimates which become substantially biased, but also upon CIR estimates because information is lacking about the dose-response curve beyond the boundary.

Figure S6 illustrates this effect. Its four panels depict the dose distribution after n=40 for a hypothetical 8-dose ED50-targeting UDD with hard boundaries. The first 8 patients were excluded to mitigate the starting-dose effect. In all panels F(x) remains the same, but the doses and their boundaries shift successively one level to the left. When the upper and lower boundaries are sufficiently far from target (top), the expected average dose assignment (light-blue vertical line) falls right on top of target (red line, not visible in top panel). As the upper boundary moves closer to target, the gap between target and average increases, reaching almost a full dose level when the boundary is right next to the target (bottom panel).

As a result, dose-averaging estimates incur a bias from the magnitude of the distance between the red and blue lines, on top of any starting-dose bias (if not mitigated properly).

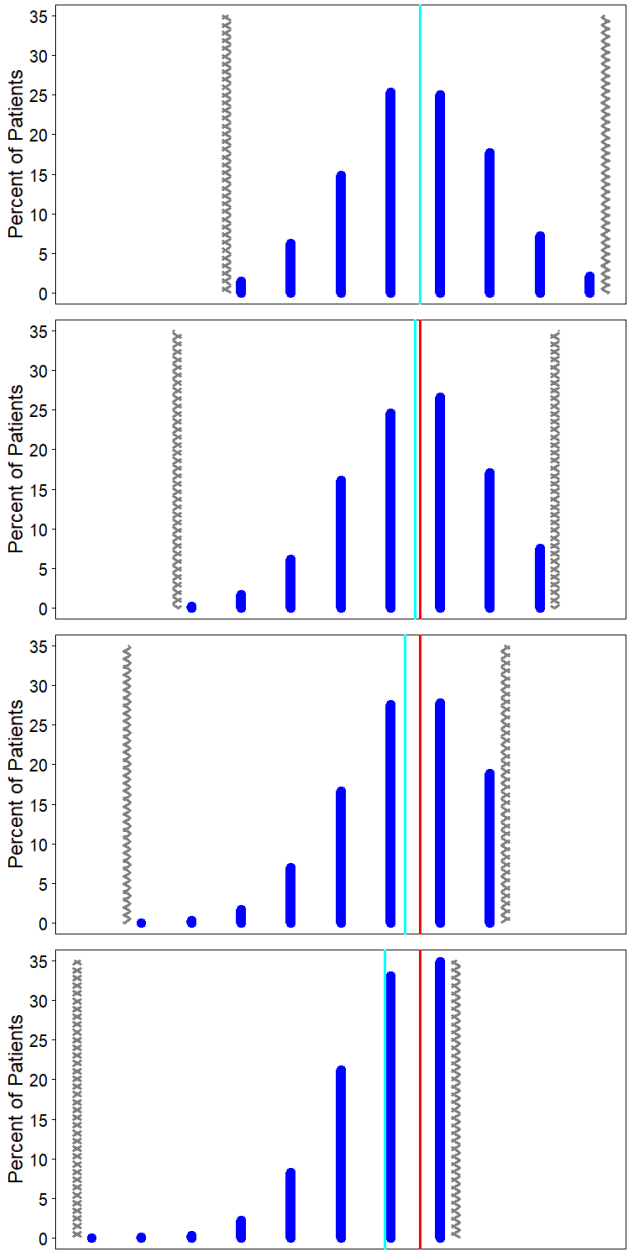

Figure S6: calculated average dose-allocation distributions (blue bars) from patients 9-40 of an n=40 ED50-targeting design. The same F(x) was used in all panels, but the boundaries (vertical black curls) shift one dose-level to the left in each panel from top to bottom. The expected average of allocation from patients 9-40, as shown as a light blue vertical line, is seen to shift away from target (red).

Important UDD Variations Not Discussed in the Main Article

A. Cohort or Group UDD

Sometimes researchers would like to treat several patients simultaneously, or in quick succession without waiting for the most recent patient's result at each step. In this context, a **group UDD (GUDD)**[27,28] comes in handy:

- Specify the group size *g,* and **lower and upper transition thresholds** l and u ($l<u$)
- At each step, treat and evaluate a group of size *g* at the same dose;
- Count the number y of effective responses in the current group.
- If $y<=l$ increase the dose;
- If $y>=u$ decrease the dose;
- If $u<y<l$ the next group receives the same dose.

We denote a specific GUDD as $GUDD_{g,l,u}$. For example, $GUDD_{1,0,1}$ is another name for the Classical UDD. Like KRD, GUDDs cannot target any arbitrary percent response. Unlike KRD, most GUDDs lack a closed-form formula for the target, and require numerical calculation. One exception is $GUDD_{g,g-1,g}$ which **shares the same targets as KRD with *k=g*.** The formula for *p*, the target response rate of these designs (not provided in the main article) is

$$p=\frac{1}{2}^{1/k}.$$

For example, as implied in the main article: with k=6, *p* = (0.5)^(1/6), or approximately 0.89.

GUDDs with $l+u=g$ are symmetric, all of them targeting the ED50. Gezmu and Flournoy provide an extensive table of $\{g,l,u\}$ sets and their targets, focusing on targets below the median.[28]

The R function `gudtarg` below calculates GUDD's target response rate given $\{g,l,u\}$:

```
gudtarg<-function(g,l,u)
{
# Transition balance equation
tbalance <- function(x, gee, ell, you)  {
     pbinom(q=ell, size=gee, prob=x) + (pbinom(q=you-1, size=gee,
          prob=x) - 1)  }

# Target rate: where the balance is zero
uniroot(f=tbalance, interval=0:1, gee=g, you=u, ell=l)$root
}
```

B. Parallel UDDs to Test for Differences Between Groups

The practice of splitting the sample into groups (e.g., by patient properties or by different treatments), and running separate but similarly-designed UDDs to compare their target doses is currently popular in anesthesiology. Pace and Stylianou's 2007 article revisited two such experiments,[15] and similar studies continue to appear.

In the case of two groups, the hypothesis-testing method used in many studies is to calculate 83% confidence intervals; if the intervals don't overlap, the null hypothesis of no difference is rejected at $p<0.05$, and vice versa. One could extend the concept to a larger number of groups with multiple-testing adjustments. Theoretical examinations suggest that with UDD data this method's probability of falsely rejecting the null hypothesis (the Type I error) is robust. However, its power to detect a difference when one exists is unimpressive.[16]

When using UDD with an interval-overlap test, a conservative CI would be prudent. As a case in point, we revisited the Benhamou et al. study mentioned above in the 'cir' package section,[4] which was previously revisited by Pace and Stylianou. The goal was to compare the EC50s of levobupivacaine and ropivacaine for analgesia during labor (Figure S7). The original researchers used the Dixon-Mood UDD estimate for each agent. The method used for group comparisons was not specified, but since the single-group estimators were dose averaging, we presume that the Dixon-Mood SEMs were used assuming independence between the two groups. The ED50 estimates were 0.092% (95% CI,0.082–0.102%) for ropivacaine and 0.077% (0.058–0.096%) for levobupivacaine and, with the difference 0.015% (-0.008 – +0.037%) deemed not significant.

Pace and Stylianou revisited this study using isotonic regression, bootstrap CIs, and the 83%-CI overlap test. Their estimates were 0.093% (0.080–0.100%) and 0.068% (0.058–0.095%) for ropivacaine and levobupivacaine, respectively. The 83% CIs were (0.087–0.097%) and (0.059–0.081%), respectively, indicating a significant difference.

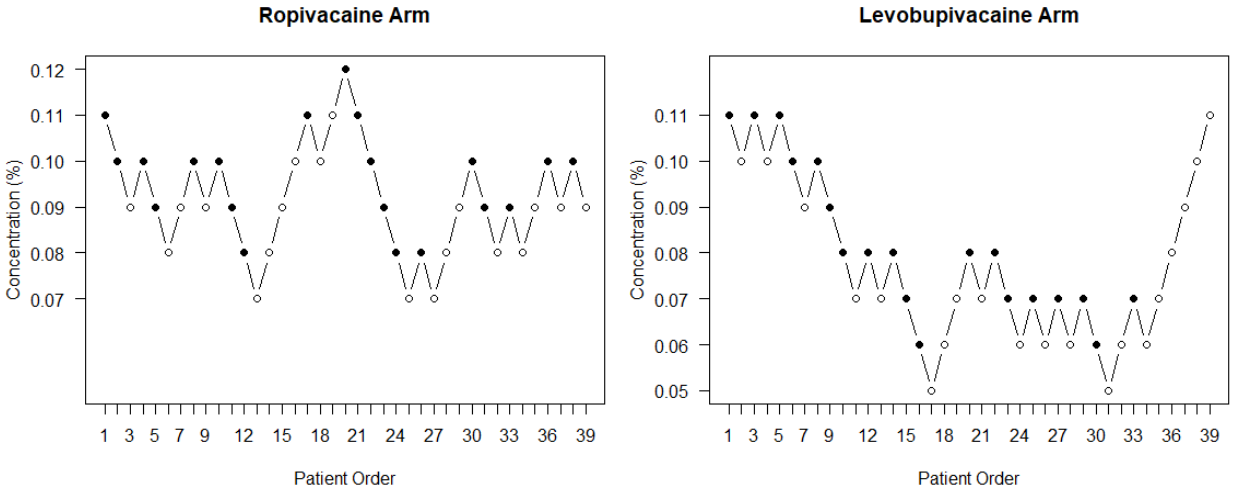

Figure S7: Benhamou et al.'s data with filled and empty circles denoting effective and ineffective responses, respectively. Each arm's last patient was omitted due to unknown response.

Our own re-analysis using CIR and its analytically-derived CI yields 0.094% and 0.068% for the point estimates, very similar to Pace and Stylianou; however, our 83% CIs were 0.083–0.104% and 0.056–0.082%, respectively. Both are wider, with our ropivacaine interval more than twice as wide as the Pace-Stylianou bootstrap CI. While the ropivacaine and levobupivacaine intervals still do not overlap in our version, they are nearly touching indicating very borderline evidence for different EC50s at the α=0.05 level. We note that neither re-analysis could use the 40th and last observation from each arm, because the original study's figures did not distinguish positive and negative responses; there was no tabular data summary and access to the original data had since been lost (D. Benhamou, personal communication). If the last response on the ropivacaine arm was ineffective, while the point estimate wouldn't change appreciably the 83% CI would expand to 0.082–0.106%, placing further doubt on the evidence for a difference between EC50s.

Looking at the estimated dose-response curves (Figure S8), the bootstrap CIs indeed seem too narrow. In particular, the 95% bootstrap CI for ropivacaine barely touches on the 0.08% dose, despite the observed efficacy rate at that dose being nearly 0.4, nearly identical to the rate at dose 0.09% which is the closest one to the target estimate (3 of 8 vs. 5 of 13).

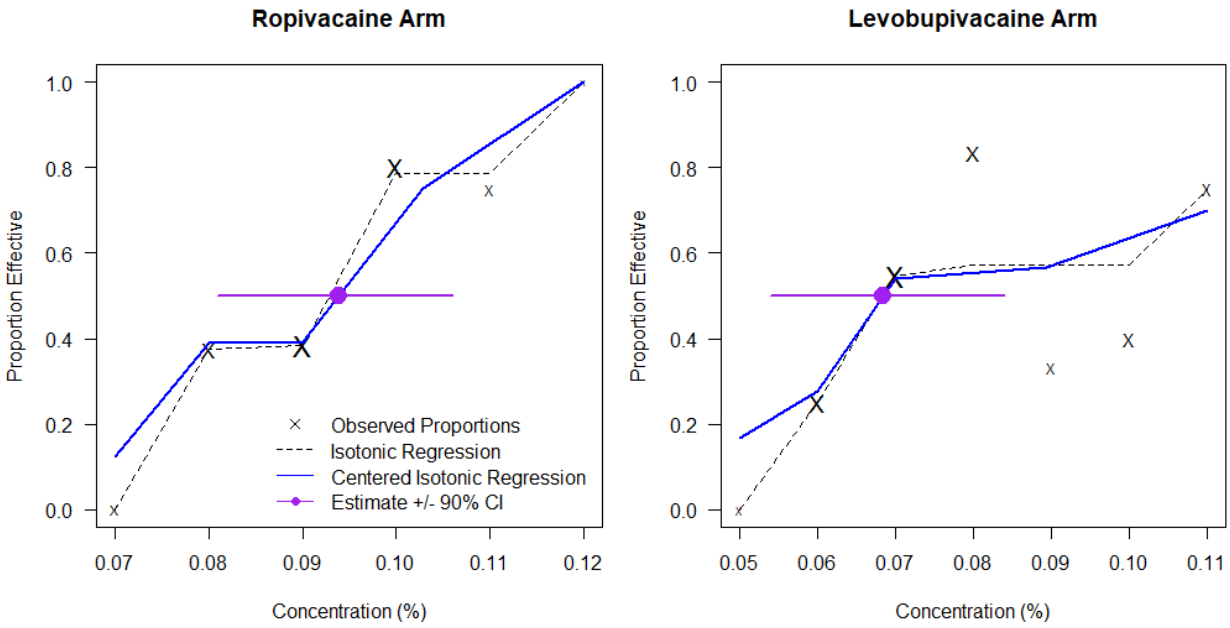

Figure S8: Benhamou et al.'s data as dose-response summaries, with the Pace-Stylianou isotonic regression and our CIR estimates curves. The purple marks show the CIR target estimates with 90% CIs.

C. Quick-Start Stage for Non-ED50 UDDs

Asymmetric UDDs allow for targeting any percentile. In addition, they tend to generate sharper dose-allocation peaks around the target, due to the rules that facilitate repeating the same dose. For example, near the target of an ED90-targeting UDD, after each observation, the same dose would be repeated with ~80% probability, while dose increase/decrease occur only with ~10% probability each.

However, with these designs if the experiment begins far away from target, it may take an inordinately large number of observations to traverse the distance. Due to the asymmetry, this number will depend on which direction the experiment needs to go to reach the target. With the high-percentile targets common in anesthesiology, dose decreases are slower. Therefore, we recommend beginning such experiments somewhat *below* the presumed location of the target, or somewhat below the middle of the effective range (see the main article regarding effective range).

If that is not possible, one may consider quick-starting the experiment with a Classical-UDD stage. That is, for a prespecified initial duration the Classical UDD rules would be applied

mandating a dose transition after every observation, before switching to the main design that targets, say, the ED90. In the main article we cautioned against adaptive design changes early in the experiment. However, we explained that the issue is more of quantity than principle. When targeting an extreme percentile and facing a dose-transition rate that is at least 6x slower in one direction, the benefit-risk assessment for beginning with a faster albeit “wrong-target” stage may be more appealing.

That said, the risks should still be considered. Since the start-up stage targets the ED50, the most substantial risk is overshooting the eventual target. Even though recovery towards the true target region after switching designs should be relatively quick, there may be ethical concerns regarding spending too much time in a suboptimal region of the dose range.

For example, Dosani et al. investigated slower administration of propofol to children undergoing endoscopy, in order to reduce apnea frequency to ~5%; in other words, their study targeted the ED95 of apnea prevention.[17] They began with a relatively slow rate (equivalent to top of the range for dose-efficacy studies) and a Classical UDD stage ending after the third observed apnea; then they switched to a BCD targeting the 95th percentile of apnea prevention. This is a relatively late switching point; most UDD studies employing a quick-start stage switch after the first reversal.[18] During the experiment, the switch to BCD took place only after 17 patients out of a total of 50, by which time the propofol administration rate was 3-4 spacing units faster than the eventual target estimate. Very likely, the target was overshot and it took additional subjects for the sequence to backtrack to the target region. There were 7 apneas observed among the 50 patients, about 3 times the experiment’s designated 5% target rate.

Besides the risk of overshooting, dose-finding designs induce an observed-rate bias at doses flaring away from the target, as mentioned in the main article. Therefore, the initial ED50-targeting stage has a different bias pattern from the main non-ED50 targeting stage. This is another reason to keep the quick-start stage short, so that it accounts for a small fraction of the final sample size.

In summary, the recommended practice for non-ED50 UDDs is to begin them somewhat off-center towards the side having faster expected transition whenever feasible, obviating the need for a separate quick-start stage. For above-median targets this starting point would be a somewhat lower dose than otherwise planned. A quick-start stage may be considered, particularly when forced to start at a higher dose. But it should be kept very short, e.g., until the first undesirable response (negative response for high-target designs and vice versa).

More on Long-Memory Dose-Finding Designs

As described in brief in the main article's discussion section, dose-finding designs using Bayesian methods, such as the Continual Reassessment Method (CRM),[19] Escalation With Overdose Control (EWOC),[20] the modified probability toxicity interval (mTPI)[21] and Bayesian Optimal Interval Design (BOIN)[22] have become popular in the Phase I cancer trial design literature, and are making inroads into anesthesiology as well. We wrote that on balance, for the straightforward task of dose-finding with small to moderate samples, UDDs still seem preferable. Here we expand the discussion and explain our recommendation, in view of the interest these other designs have generated and their increasing use in anesthesiology.

Oron and Hoff included these designs in a class they called "long-memory" designs,[23] in contrast to the "short memory" of UDDs. Long-memory designs share two main traits:

1. Dose allocations rely upon estimates of F(x) that may use all prior data starting from the beginning of the experiment.
2. The allocations follow a design-specific optimization criterion, intended to provide the "best" or at least a highly desirable dose to the next patient.

Because of the second trait, Fedorov and Leonov call this class "best-intention designs".[24] Their discussion of the topic also includes designs with more complex optimization criteria lacking the second trait; we do not discuss the latter, as they are not in prevalent use.

Of special note is an important subtype of long-memory designs that is often not perceived as such in the literature. These are **interval designs**, which typically repeat the same dose if the cumulative response rate there is within a tolerance interval around the target rate. For example, an ED50-targeting interval design might repeat the same dose if the cumulative proportion of positive response at the current dose is between 40% and 60%, increase the dose if <40%, and decrease it if >60%. Interval designs' "optimization criterion" is often so simple as to be overlooked, and some of them are completely model-free or were inspired by UDDs, e.g., the "Narayana design" and the Cumulative Cohort Design.[25–28] The former antedates all other long-memory designs by almost 40 years, having been developed by the late T.V. Narayana in his 1953 dissertation. At the time only the Classical UDD existed, and the design does not

appear to have been put into use until its rediscovery and modification in the 21st Century.[45] Even though interval designs seem very distinct from designs such as the CRM, they belong to the long-memory family due to their use of cumulative response-rate estimates, and the principle of repeating the same dose when deemed too good to be replaced by another.

The two main arguments for long-memory designs are (1) that **when deciding the next allocation it is better to use all data** rather than only a few observations, and (2) that **long-memory dose allocations eventually converge to deliver only the best dose** (or best two doses in some cases) to all subsequent patients. Usually, that would be the dose closest to target. Regarding the first argument, as explained in the main article, UDDs' particular short-memory rules generate target-centered random walks, with robust and beneficial properties for dose-finding. Therefore, the short memory is not necessarily a liability for dose allocation. In addition, during target estimation after a UDD experiment, all (or nearly all) observations are used so efficiency is not compromised, and indeed UDDs find the target with efficiency similar to leading long-memory designs.[23,29–31]

The second argument, regarding long-memory convergence, requires greater scrutiny. Focusing on CRM, by far the most commonly used long-memory design: its one-parameter model is too unrealistic to capture the broad range of forms F(x) might take. Hence, it is by definition a misspecified model. This is acknowledged by CRM developers who call it a "working model". Therefore, it is of interest to examine its behavior under model misspecification. The original 1990 CRM article provided no convergence proof for the design. The first such proof arrived six years later, but with conditions on F(x) so restrictive as to be practically irrelevant.[9,32,33] Eventually in 2009, Lee and Cheung reported that CRM converges to within an interval around the target rate, rather than to the dose closest to target. Interval width is determined indirectly by design parameters, and Lee and Cheung provide software to help control it.[34] Interval convergence makes CRM's asymptotic behavior nearly identical to interval designs, which are usually far simpler to run. Interval-design convergence was first proven by Oron et al. in 2011.[9] To our knowledge, EWOC still lacks a convergence proof when its two-parameter model is misspecified.

Bringing dose allocations exclusively to within an interval around the target is decent asymptotic behavior. But how long does it take to observe such behavior with long-memory designs? UDD random walks converge to their long-term behavior at a geometric rate, and therefore in a typical experiment one can observe the characteristic meandering around (the presumed) target

within a few dozen observations at the latest.[13,35] By contrast, long-memory designs rely upon the convergence of observed response rates. This occurs far more slowly, at a root-n pace. Our simulations suggest that it takes hundreds of observations for long-memory designs to settle reliably at near-asymptotic behavior.

Numerous simulation studies about long-memory designs report measures of performance averaged over a large number of simulated experiments. While these averages may look good, individual long-memory experimental trajectories famously tend to settle or "stick" relatively early with a single dose, allocating it repeatedly.[23] This is often preceded by volatile dose transitions during the first few allocations, and therefore the subsequent settling has been misunderstood as genuine convergence behavior, even by CRM design experts.[30,36] Simply put, that is impossible; due to basic probability calculations with binary outcomes, near-asymptotic behavior requires hundreds of patients. **Rather than convergence, the early sticking behavior is a side-effect of repeatedly fitting the same model using almost exactly the same data.** These early "bets" on a single dose can be right when all stars align; however, they are often wrong, leading to unwieldy and seemingly inexplicable experimental behavior. There is little relationship between a long-memory experiment settling early on a dose, and that dose being the one truly closest to target.[23,24] Results are worse for Bayesian designs utilizing a prior distribution: they prefer to settle on the dose favored by that prior.[23] The Narayana design has unique behavior in that respect: its long-memory element is equivalent to an interval design with zero interval width. Rather than stick to a single dose, a zero-width interval promotes zigzagging transitions between two adjacent doses, both asymptotically and with small samples.However, the original Narayana design also included a (Classical) UDD rule, and mandated that unless both rules point in the same direction, the current dose is to be repeated.[25] This double restriction further exacerbates the early "sticking" to a single dose, thus making the original design rather impractical, despite being conceptually decades ahead of its time. Recent anesthesiology studies using a modified Narayana design did not include a UDD rule, and indeed the expected zigzagging was observed.[27,28]

Regardless of subtype, long memory designs lend the first few observations and the prior distribution (if there is one) a disproportionately large influence upon subsequent behavior, because they participate in every subsequent dose allocation estimate. As a result, while the *average* number of patients treated closest to target tends to be higher with CRM than with analogous UDDs, the *variability* between experiments, as well as the variability in response to variations in F(x), are far greater with CRM. In simulations, it is common to see long-memory

runs with few, and even zero, patients treated at the “best” dose. By contrast, UDDs treat roughly the same number of patients at that dose, give or take a few. Thus, UDDs are more robust to variations and surprises encountered in typical experimental settings.[23]

Furthermore, the more complex long-memory designs, and in particular the CRM, are prone to misguided implementation practices that increase the chances of off-target “sticking” and other undesirable and unexpected side effects.[37–39] There has been some acknowledgment of these issues in the long-memory design community. Cheung provides planning tools to prevent the worst of erratic CRM behavior,[40] and others have attempted to mitigate the early “sticking” via randomization.[41,42] However, the randomization modifications are associated with increased sample sizes and have not gained broad adoption, and quite a few CRM studies, including all the studies we’ve encountered in anesthesiology, do not make use of Cheung’s safer design aids.

Many of these problems can be mitigated with larger sample sizes, but long-memory designs are usually implemented under the misguided paradigm that they “achieve convergence” with smaller samples and aggressive stopping rules. The most common sample size we found in anesthesiology CRM studies is 24 patients, with the target usually the ED80 or a more extreme percentile. As we suggest in the main article, this is woefully insufficient. Similarly-targeted anesthesiology UDD studies rarely use less than 40 patients, and we recommend at least 50-60. Furthermore, in the anesthesiology CRM studies the stopping rules included stopping when the estimated probability of dose transition during the next several observations becomes too small.[43,44] Hence, the very same problematic “sticking” behavior is still viewed and utilized as an asset and as a sign of convergence, and probability calculations from a model known to be misspecified are used at face value for the major decision of ending the experiment.

We conclude that for straightforward dose-finding UDDs are, at least at present, the more appropriate choice, particularly with recent improvements in estimation methods. Even when more complex study goals call for more complex designs than UDD, UDDs may still provide useful start-up stages that avoid early “sticking” behavior and other complications that are seen with long-memory designs.

References

1. Oron AP. cir*: Centered Isotonic Regression and Dose-Response Utilities*.; 2021. Accessed February 10, 2022. https://CRAN.R-project.org/package=cir
2. Oron AP, Flournoy N. Centered Isotonic Regression: Point and Interval Estimation for Dose-

Response Studies. *Stat Biopharm Res*. 2017;9(3):258-267.
3. Flournoy N, Oron AP. Bias induced by adaptive dose-finding designs. *J Appl Stat*. 2020;47(13-15):2431-2442. doi:10.1080/02664763.2019.1649375
4. Benhamou D, Ghosh C, Mercier FJ. A Randomized Sequential Allocation Study to Determine the Minimum Effective Analgesic Concentration of Levobupivacaine and Ropivacaine in Patients Receiving Epidural Analgesia for Labor. *Anesthesiology*. 2003;99(6):1383-1386. doi:10.1097/00000542-200312000-00022
5. Wetherill GB, Chen H, Vasudeva RB. Sequential estimation of quantal response curves: A new method of estimation. *Biometrika*. 1966;53:439-454.
6. Kershaw CD. Asymptotic properties of W, the estimator of the ED50 suggested for use in up-and-down experiments in bioassay. *Ann Stat*. 1985;13(2):85-94.
7. Oron AP. Up-and-Down and the Percentile-Finding Problem. Ph.D. dissertation, University of Washington; 2007. https://arxiv.org/abs/0808.3004
8. Paoletti X, O'Quigley J, Maccario J. Design efficiency in dose finding studies. *Comput Stat Data Anal*. 2004;45:197-214.
9. Oron AP, Azriel D, Hoff PD. Dose -Finding Designs: the role of convergence properties. *Int J Biostat*. 2011;7(1):Article 39.
10. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed. Springer-Verlag; 2009.
11. Dixon WJ, Mood AM. A method for obtaining and analyzing sensitivity data. *J Am Stat Assoc*. 1948;43:109-126.
12. Dixon WJ. *Introduction to Statistical Analysis: By Wilfrid J. Dixon and Frank J. Massey, Jr*. McGraw-Hill; 1951.
13. Oron AP, Hoff PD. The k-in-a-row up-and-down design, revisited. *Stat Med*. 2009;28:1805-1820.
14. Choi SC. Interval estimation of the LD50 based on an up-and-down experiment. *Biom J Biom Soc*. 1990;46(2):485-492.
15. 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):144-152.
16. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *J Insect Sci*. 2003;3(1). doi:10.1093/jis/3.1.34
17. Dosani M, McCORMACK J, Reimer E, et al. Slower administration of propofol preserves adequate respiration in children. *Pediatr Anesth*. 2010;20(11):1001-1008. doi:10.1111/j.1460-9592.2010.03398.x
18. Storer BE. Design and analysis of phase I clinical trials. *Biom J Biom Soc*. 1989;45(3):925-937.
19. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for Phase I clinical trials in cancer. *Biom J Biom Soc*. 1990;46(1):33-48.
20. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat Med*. 1998;17:1103-1120.
21. Ji Y, Liu P, Li Y, Nebiyou Bekele B. A modified toxicity probability interval method for dose-finding trials. *Clin Trials*. 2010;7(6):653-663.
22. Liu S, Yuan Y. Bayesian optimal interval designs for phase I clinical trials. *J R Stat Soc Ser C Appl Stat*. 2015;64(3):507-523. doi:10.1111/rssc.12089
23. Oron AP, Hoff PD. Small-Sample Behavior of Novel Phase I Cancer Trial Designs. *Clin Trials*. 2013;10(1):63-80.
24. Fedorov VV, Leonov SL. Other Applications of Optimal Designs. In: *Optimal Design for Nonlinear Response Models*. CRC Press; 2013.
25. Narayana TV. Sequential procedures in the probit analysis. Ph.D. Dissertation, University of

North Carolina; 1953.
26. Ivanova A, Flournoy N, Chung Y. Cumulative cohort design for dose-finding. *J Stat Plan Inference*. 2007;137:2316-2327.
27. Tanaka M, Balki M, Parkes RK, Carvalho JCA. ED95 of phenylephrine to prevent spinal-induced hypotension and/or nausea at elective cesarean delivery. *Int J Obstet Anesth*. 2009;18(2):125-130. doi:10.1016/j.ijoa.2008.09.008
28. Mohta M, Dubey M, Malhotra RK, Tyagi A. Comparison of the potency of phenylephrine and norepinephrine bolus doses used to treat post-spinal hypotension during elective caesarean section. *Int J Obstet Anesth*. 2019;38:25-31. doi:10.1016/j.ijoa.2018.12.002
29. Durham SD, Flournoy N, Rosenberger WF. A random walk rule for phase I clinical trials. *Biom J Biom Soc*. 1997;53(2):745-760.
30. Oron AP, Hoff PD. Small-Sample Behavior of Novel Phase I Designs: Rejoinder. *Clin Trials*. 2013;10(1):88-92.
31. Flournoy N, Moler J, Plo F. Performance Measures in Dose‐Finding Experiments. *Int Stat Rev*. 2020;88(3):728-751.
32. Shen LZ, O'Quigley J. Consistency of continual reassessment method under model misspecification. *Biometrika*. 1996;83(2):395-405. doi:10.1093/biomet/83.2.395
33. Cheung YK, Chappell R. A simple technique to evaluate model sensitivity in the continual reassessment method. *Biom J Biom Soc*. 2002;58(3):671-674.
34. Lee SM, Cheung YK. Model Calibration in the continual reassessment method. *Clin Trials*. 2009;6:227-238.
35. Diaconis P, Stroock D. Geometric bounds for eigenvalues of Markov chains. *Ann Appl Probab*. 1991;1(1):36-61.
36. Cheung YK. Commentary on 'Small-sample behavior of novel phase I cancer trial designs.' *Clin Trials*. 2013;10(1):86-87. doi:10.1177/1740774512470221
37. Mathew P, Thall PF, Jones D, Perez C, Bucana C, Troncoso P, Kim S-J, Fidler IJ, Logothetis C. 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):3323-3329.
38. Neuenschwander B, Branson M, Gsponer T. Critical aspects of the Bayesian approach to phase I cancer trials. *Stat Med*. 2008;27:2420-2439.
39. 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.
40. Cheung YK. *Dose Finding by the Continual Reassessment Method*. Chapman and Hall / CRC; 2011.
41. Thall PF, Nguyen HQ, Zohar S, Maton P. Optimizing Sedative Dose in Preterm Infants Undergoing Treatment for Respiratory Distress Syndrome. *J Am Stat Assoc*. 2014;109(507):931-943. doi:10.1080/01621459.2014.904789
42. Koopmeiners JS, Wey A. The Randomized CRM: An Approach to Overcoming the Long-Memory Property of the CRM. *J Biopharm Stat*. 2017;27(6):1028-1042. doi:10.1080/10543406.2017.1293076
43. Beloeil H, Eurin M, Thévenin A, Benhamou D, Mazoit JX. Effective dose of nefopam in 80% of patients (ED80): a study using the continual reassessment method. *Br J Clin Pharmacol*. 2007;64(5):686-693. doi:10.1111/j.0306-5251.2007.02960.x
44. Le Gouez A, Bonnet MP, Leclerc T, Mazoit JX, Benhamou D, Mercier FJ. Effective concentration of levobupivacaine and ropivacaine in 80% of patients receiving epidural analgesia (EC80) in the first stage of labour: A study using the Continual Reassessment Method. *Anaesth Crit Care Pain Med*. 2018;37(5):429-434. doi:10.1016/j.accpm.2017.12.009
45. Ivanova A, Haghighi AA, Mohanty SG, Durham SD. Improved up-and-down designs for Phase I trials. Stat Med. 2003;22:69-82.