capstract:2611.00056v1 [stat.ME] 20 Nov 2026

Dose-Finding for Repeated Presses of a Latched Hall-Call Button: A Continual Reassessment Design with an Efficacy Futility Rule

Mara Lindqvist[1,*] Teodor Brzezinski[1] Anneli Okonkwo-Reyes[2]
[1]Department of Statistics, Oakmere University, Leeds, United Kingdom
[2]Institute for Building Systems Research, Tarn Valley Polytechnic, Zurich, Switzerland

November 20, 2026

Abstract

Presses of an elevator hall-call button that is already latched are discarded by the controller, but the practice is common and its dose has never been bounded. We treat the number of redundant presses as a dose and look for the maximum tolerated dose, with a dose-limiting event defined as an audible remark by a bystander that is addressed to the person pressing or to the button. We describe a continual reassessment design in which a Bayesian futility rule on a continuous efficacy endpoint, the error of the controller's arrival estimate, is added to the toxicity model, and we give the stopping boundary in closed form. In simulations over five toxicity scenarios with 2,000 replicates each, the design selected the correct dose in 28% to 90% of replicates, similar to a Bayesian optimal interval design and above the 3+3 design (30% to 42%). With no efficacy the futility rule stopped 46% to 95% of trials, and with an efficacy of 1.0 s per press it stopped fewer than 1%. We then ran the design over 11 weekday mornings. The futility rule stopped the trial after 33 episodes, when the posterior probability that the slope of arrival-estimate error on dose was steeper than -0.5 s per press fell to 0.018. The estimated maximum tolerated dose was 8 presses (posterior dose-limiting probability 0.27), and the estimated slope was 0.05 s per press (95% CI -0.54 to 0.63). No dose was recommended for further study.

Keywords: phase I design, continual reassessment method, dose-finding, futility stopping, Bayesian adaptive design, maximum tolerated dose

1 Introduction

A passenger who presses an elevator hall-call button that is already lit is, from the controller's point of view, doing nothing. Under collective control the first press creates a queue entry for that floor and direction, and later presses on the same button are discarded (Sorensen and Mbeki, 2021). The practice is nevertheless common: 61% of 2,114 surveyed commuters reported pressing a lit button at least occasionally (Aldous and Ferreira-Nkosi, 2018), and nothing limits the number of presses that one person may apply. We could not find a study that relates the number of presses either to a benefit or to the reaction of the people waiting nearby.

∗Corresponding author: m.lindqvist@stats.oakmere.ac.uk

These are the two questions of a phase I dose-finding study: how large an exposure is tolerated, and whether the exposure does anything. The standard designs address the first. The 3+3 design escalates in cohorts of three according to fixed rules (Storer, 1989), the continual reassessment method (CRM) updates a one-parameter dose-toxicity model after each cohort (O'Quigley et al., 1990), and the Bayesian optimal interval design (BOIN) compares the observed toxicity rate at the current dose with two fixed boundaries (Liu and Yuan, 2015). Designs that use efficacy as well as toxicity exist (Thall and Cook, 2004), but they generally assume that efficacy is present at some dose and ask where the best trade-off lies.

If the exposure has no efficacy at any dose, a toxicity-only design still places most participants near the maximum tolerated dose (MTD), since nothing in the design asks whether the exposure achieves anything. We add to a CRM a futility rule on a continuous efficacy endpoint. The rule is a one-sided posterior probability on the slope of the endpoint against dose, and it has a closed-form boundary (Section 3.2). We evaluate the design by simulation against the 3+3 design and BOIN (Section 4) and report a trial in which the dose was the number of redundant presses (Section 5).

2 Setting and endpoints

2.1 Dose

An *episode* begins when a trained research assistant (the *presser*) stands within 1 m of a hall-call panel at which the call for the presser's direction is latched and at least one other passenger is waiting within 4 m. The presser presses the lit button once, then waits 5 s, then applies d further presses at 1 Hz, and then waits for the car. The dose d is the number of redundant presses. There are $K = 8$ dose levels, $d_1, \dots, d_8 = 1, 2, 3, 4, 6, 8, 12, 16$. Episodes in the same lobby were at least 10 minutes apart and no lobby hosted more than one episode on a given morning.

2.2 Dose-limiting event

Bystander reactions were graded on a four-point scale adapted from Vasquez and Linde (2020). Grade 1 is a non-verbal reaction such as a look or an audible exhalation. Grade 2 is a remark to a companion. Grade 3 is a remark addressed to the presser or to the button. Grade 4 is a bystander touching or covering the button, or leaving the lobby by the stairs while remarking on the presses. A dose-limiting toxicity (DLT) is a reaction of grade 3 or higher by any bystander during the episode. Two raters coded audio transcripts from which the press sounds had been removed. On 120 transcripts coded by both, agreement on the DLT indicator was 94.2% (Cohen's $\kappa = 0.88$).

2.3 Efficacy endpoint

The efficacy hypothesis is that redundant presses bring the car sooner. The endpoint is the *ETA error*, the time at which a car door opened at the lobby minus the arrival time that the controller displayed when the presser pressed for the first time. Negative values mean that the car came earlier than predicted. The displayed estimate is computed from the dispatch queue and so does not respond to discarded presses. The planning value for the standard deviation of the ETA error, from the previous month's controller logs, was $\sigma = 4.0$ s.

2.4 Conduct and ethics

The trial ran on 11 weekday mornings in January and February 2026 in the lobbies of 11 office towers, with one cohort of three episodes per morning in a different tower. The posterior was updated and the next dose assigned after each cohort. The study was approved by the Oakmere University Research Ethics Committee (OU-REC-2026-0147) as observation and staged behaviour in a public place. Building managers consented, and signs at the entrances stated that research on lobby behaviour was in progress. Bystanders were not identifiable from the audio transcripts, and the presser could end any episode at any time. No episode was ended early.

3 Design

3.1 Toxicity model and allocation

Let p_k be the probability of a DLT at level k and $\Gamma = 0.25$ the target. The MTD is the highest level with $p_k \le \Gamma$. We use the one-parameter power model

$$p_k(a) = s_k^{\exp(a)}, \qquad a \sim \mathrm{N}(0, 1.34), \tag{1}$$

with skeleton $(s_1, \dots, s_8) = (0.06, 0.12, 0.18, 0.25, 0.34, 0.44, 0.56, 0.68)$. The prior variance of 1.34 follows Cheung (2011). After c cohorts, let n_k and t_k be the number of episodes and DLTs at level k and let $\hat{p}_k = \mathbb{E}\{p_k(a) \mid n, t\}$ be the posterior mean. The next cohort is treated at the level k that minimises $|\hat{p}_k - \Gamma|$, subject to not exceeding the current level by more than one (Goodman et al., 1995). Cohorts have three episodes and the maximum sample size is 60, or 20 cohorts. All posterior calculations use a grid of 201 values of a on $[-3, 3]$.

3.2 Futility rule

Let d_i and e_i be the dose and ETA error of episode i, $i = 1, \dots, n$. We assume

$$e_i = \beta_0 + \beta\, d_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathrm{N}(0, \sigma^2), \tag{2}$$

with σ known, a flat prior on β_0 and $\beta \sim \mathrm{N}(0, \tau^2)$ with $\tau = 2$ s per press. Efficacy corresponds to $\beta < 0$. Let $S_{xx} = \sum_i (d_i - \bar{d})^2$, $I = S_{xx}/\sigma^2$, and let b be the least-squares slope. The posterior is $\beta \mid \text{data} \sim \mathrm{N}(m, v)$ with

$$v = \left(\tau^{-2} + I\right)^{-1}, \qquad m = v\, I\, b. \tag{3}$$

Let $\delta = 0.5$ s per press be the smallest slope of interest. After each cohort from the eighth onward, enrolment stops for futility if $\mathbb{P}(\beta < -\delta \mid \text{data}) < \alpha_F$, with $\alpha_F = 0.05$. Before the eighth cohort the rule is not applied, because the dose spread is small and the posterior probability can fall below α_F by chance.

Proposition 1 (Futility boundary)**.** *Let $z = \Phi^{-1}(1 - \alpha_F)$. The rule stops enrolment if and only if*

$$b > \left(1 + \frac{1}{\tau^2 I}\right)\left(-\delta + z\sqrt{v}\right). \tag{4}$$

As $I \to \infty$ the right-hand side tends to $-\delta$, and if the true slope is $-\delta$ and $b \sim \mathrm{N}(-\delta, I^{-1})$, the probability of stopping at a single look tends to α_F.

Proof. By Eq. (3), $\mathbb{P}(\beta < -\delta \mid \text{data}) = \Phi\{(-\delta - m)/\sqrt{v}\}$, which is below α_F if and only if $-\delta - m < -z\sqrt{v}$, that is, $m > -\delta + z\sqrt{v}$. Substituting $m = vIb$ and noting that $1/(vI) = 1 + 1/(\tau^2 I)$ gives Eq. (4). As $I \to \infty$, $v \sim 1/I$ and the factor in front tends to one, so the boundary is $-\delta + z/\sqrt{I} + o(I^{-1/2})$. If $b \sim \mathrm{N}(-\delta, I^{-1})$ then $\mathbb{P}(b > -\delta + z/\sqrt{I}) = 1 - \Phi(z) = \alpha_F$. ∎

The boundary depends on the doses only through I, so that the rule needs a spread of doses before it can stop. With $S_{xx} = 100$ and $\sigma = 4$ we have $I = 6.25$, and enrolment stops if $b > 0.15$ s per press. With $S_{xx} = 233$, which is the value at the end of the trial in Section 5, $I = 14.6$ and the boundary is -0.07 s per press. The probability of stopping when the true slope is exactly $-\delta$ exceeds α_F in practice because the rule is applied at up to thirteen looks (Table 3).

3.3 Final selection

When enrolment ends, by futility or at 60 episodes, the MTD estimate is the level k that minimises $|\hat{p}_k - \Gamma|$. If enrolment stopped for futility, no dose is recommended for further study whatever the MTD estimate.

4 Operating characteristics

4.1 Scenarios and comparators

We used five true dose-toxicity curves (Table 1). In each, the true MTD is the highest level with $p_k \leq 0.25$. The comparators were the 3+3 design in its standard form, in which a dose with 2 or more DLTs among 3 to 6 episodes ends escalation and the MTD is the next lower dose, and BOIN with target 0.25, boundaries $\lambda_e = 0.197$ and $\lambda_d = 0.298$, cohorts of three, a maximum of 60 episodes, and the dose chosen by isotonic regression of the observed rates (Liu and Yuan, 2015). The CRM column uses the design of Section 3 without the futility rule. Each entry is based on 2,000 replicates. The 3+3 design has no fixed sample size, and its mean is reported.

Table 1: True DLT probabilities by dose level in the five scenarios. The MTD is the highest level with probability at most 0.25.

Scenario	1	2	3	4	5	6	7	8	MTD
S1	0.02	0.05	0.10	0.25	0.40	0.55	0.70	0.80	4
S2	0.05	0.12	0.25	0.38	0.50	0.62	0.74	0.85	3
S3	0.01	0.02	0.04	0.08	0.15	0.25	0.38	0.52	6
S4	0.25	0.40	0.52	0.64	0.74	0.82	0.88	0.93	1
S5	0.02	0.03	0.05	0.07	0.10	0.14	0.20	0.25	8

4.2 Toxicity-only results

The CRM and BOIN selected the correct dose in similar proportions in S1 to S4, and both were well above the 3+3 design (Table 2). The CRM was ahead in S3 (66.7% against 52.5%) and BOIN was ahead in S5, where the MTD is the top level (38.9% against 27.8%). BOIN treated fewer episodes above the MTD in S1 to S3, and the CRM treated more at the MTD. The 3+3

design used fewer episodes, a mean of 7.7 to 24.7, but at the price of selecting the correct dose in 30% to 42% of replicates. In S4 the 3+3 design treated 35.0% of its episodes above the MTD.

Table 2: Operating characteristics of three designs without a futility rule, 2,000 replicates per scenario. Correct is the percentage of replicates in which the selected dose was the true MTD. At MTD and above MTD are percentages of episodes. DLT is the percentage of episodes with a DLT.

Scenario	Design	Correct (%)	At MTD (%)	Above MTD (%)	DLT (%)	Mean episodes
S1	3+3	35.8	21.3	14.9	17.1	16.9
	BOIN	73.5	42.8	19.0	21.7	60.0
	CRM	**73.8**	**49.0**	23.3	24.1	60.0
S2	3+3	34.3	25.1	18.1	21.2	13.8
	BOIN	67.3	43.0	21.6	22.9	60.0
	CRM	**69.5**	**46.2**	25.6	24.9	60.0
S3	3+3	29.9	13.0	9.2	13.1	22.3
	BOIN	52.5	29.3	15.4	18.5	60.0
	CRM	**66.7**	**40.7**	15.8	20.4	60.0
S4	3+3	42.0	65.0	35.0	37.8	7.7
	BOIN	**90.6**	75.5	24.5	29.1	60.0
	CRM	89.8	**79.8**	20.2	28.7	60.0
S5	3+3	31.4	7.4	0.0	10.4	24.7
	BOIN	**38.9**	**21.6**	0.0	14.4	60.0
	CRM	27.8	14.7	0.0	14.8	60.0

4.3 Futility rule

Table 3 gives the performance of the CRM with the futility rule for three of the scenarios and three true slopes. At $\beta = 0$ the rule stopped 45.7% of trials in S1, 85.0% in S3 and 94.8% in S5. The proportion depends on the doses reached, as the boundary in Eq. (4) does through I: when the MTD is low the trial explores a narrow range of doses and has less information about the slope in 60 episodes. At $\beta = -1.0$ s per press the rule stopped at most 0.8% of trials, and the proportion of correct selections was within 0.2 percentage points of the CRM without the rule in S1 (73.6% against 73.8%). At $\beta = -\delta = -0.5$ the stopping probability was 10.7% to 15.1%, above the single-look value of 5% because of repeated looks.

Table 3: CRM with the futility rule, 2,000 replicates per row. Stopped is the percentage of replicates in which enrolment ended for futility. Correct is the percentage in which the selected dose was the true MTD, counted over all replicates, including those that stopped.

Scenario	True slope (s/press)	Stopped (%)	Mean episodes	Correct (%)
S1	0	45.7	46.1	62.5
	−0.5	10.7	56.8	73.5
	−1.0	0.8	59.7	73.6
S3	0	85.0	32.6	47.6
	−0.5	10.9	56.9	64.0
	−1.0	0.1	60.0	67.7
S5	0	94.8	29.5	20.5
	−0.5	15.1	56.0	28.0
	−1.0	0.1	60.0	28.9

5 Application

Enrolment began at level 1 (one redundant press). The first four cohorts, at levels 1 to 4, had no DLTs (Table 4 and Figure 1). The fifth cohort, at level 5 (6 presses), had one. The model then moved to level 6 (8 presses), where it stayed for the remaining six cohorts. Among the 18 episodes at 8 presses there were 5 DLTs, 27.8%. There were 6 DLTs in 33 episodes overall (18.2%), and none was of grade 4. The mean number of bystanders per episode was 2.6.

The futility rule was first applied after the eighth cohort. The posterior probability that the slope was steeper than $-\mathbf{0.5}$ s per press was 0.170 at cohort 8, 0.090 at cohort 9, 0.059 at cohort 10 and 0.018 at cohort 11, and enrolment stopped (Figure 1). At that point $\boldsymbol{S_{xx} = 232.9}$, the least-squares slope was 0.05 s per press, and a 95% interval using the residual standard deviation of 4.55 s was $-\mathbf{0.54}$ to $\mathbf{0.63}$. The boundary in Eq. (4) at this information was $-\mathbf{0.07}$ s per press, and the observed slope exceeded it.

The posterior mean DLT probabilities at the stop were 0.016, 0.040, 0.071, 0.115, 0.182, 0.270, 0.393 and 0.535 for levels 1 to 8. The estimated MTD was therefore level 6, 8 presses, with a posterior mean DLT probability of 0.27. Because the trial stopped for futility, no dose was recommended for further study.

Table 4: Results of the trial at the stop, after 33 episodes in 11 cohorts. ETA error is the mean and standard deviation of the controller-estimate error in seconds. The posterior DLT probability is the posterior mean under Eq. (1).

Level	Presses	Episodes	DLTs	Posterior DLT prob.	ETA error (s)
1	1	3	0	0.016	−1.4 (2.2)
2	2	3	0	0.040	4.9 (7.2)
3	3	3	0	0.071	−0.5 (4.6)
4	4	3	0	0.115	1.8 (5.0)
5	6	3	1	0.182	4.7 (3.2)
6	8	18	5	0.270	1.3 (4.3)
7	12	0	0	0.393	
8	16	0	0	0.535	

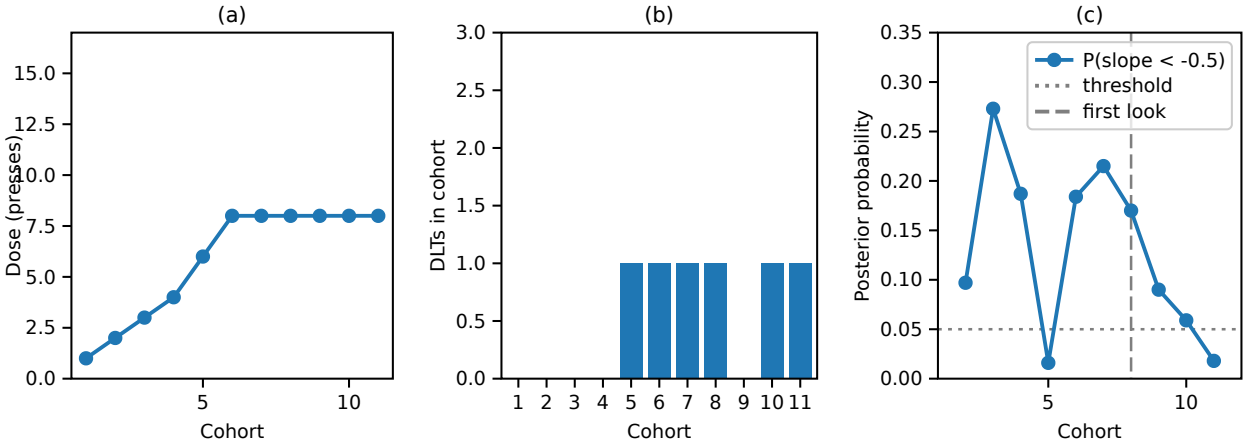


Figure 1: Course of the trial. (a) Dose assigned to each cohort. (b) Number of DLTs among the three episodes of each cohort. (c) Posterior probability that the slope of ETA error on dose is steeper than -0.5 s per press; the dotted line is the futility threshold of 0.05 and the dashed line marks the first cohort at which the rule applies. The probability is undefined at cohort 1.

6 Discussion

The futility rule did what it was designed to do. Of the 33 episodes, 15 were spent on escalation through levels 1 to 5 and 18 at 8 presses, the level preferred by the toxicity model. The slope of 0.05 s per press was far enough from $-\delta$ that the posterior probability of an efficacy of at least δ fell below the threshold at cohort 11. The 95% interval for the slope, -0.54 to 0.63, still contains $-\delta$, so the data do not exclude an efficacy as small as 0.5 s per press. The rule stopped because such an efficacy is unlikely given the prior and the data, and not because it was ruled out. The probability depends modestly on the prior and on σ. A flat prior on β gives 0.019 at cohort 11 instead of 0.018, and using the residual standard deviation of 4.55 s in place of 4.0 gives 0.032. The rule would have stopped in both cases.

The MTD of 8 presses rests on 18 episodes at that level and on the skeleton for the other levels. The posterior DLT probabilities at levels 6 and 7 differ by 0.12, and the observed rate of 5

DLTs in 18 episodes has an exact 95% interval of 9.7% to 53.5%. The simulations show that the correct dose is selected in about two thirds of trials of this size when the MTD is in the middle of the range, and in about 28% of trials when it is the top level.

There are several limitations. The pressers were trained research assistants who applied presses at a fixed rate, and untrained passengers press in irregular bursts. The DLT definition depends on the grading of remarks, and a lexicon that is accepted in one city may not be in another. The ETA error is a sensitive efficacy endpoint only if the controller's estimate is accurate. A design with two efficacy endpoints, or with the perceived wait rather than the measured one, would need a different rule. The trial was conducted in one cohort per morning, so that tower and day are confounded with dose assignment. We did not model this. We also did not study whether the toxicity depends on the number of bystanders, the time of day or the presence of a companion.

The proposed rule stops a trial when the data cannot support a slope of at least δ and gives no information about benefits at other levels of the endpoint. It could be used for any exposure whose tolerability is dose-limited and whose benefit is doubtful. In the trial reported here the toxicity model identified a dose and the efficacy model found no reason to use it.

Data availability

Episode-level doses, DLT indicators and ETA errors and the simulation code are available from the corresponding author. Audio recordings were not retained.

A Cohort log

Table 5 lists the allocation and the futility statistics after each cohort. The least-squares slope and posterior probability are not defined at the first cohort because all three episodes were at one dose.

Table 5: Cohort log of the trial. Slope is the least-squares slope of ETA error on dose after the cohort, in seconds per press. Probability is $\mathbb{P}(\beta < -0.5 \mid \text{data})$ under Eq. (3) with $\sigma = 4.0$. The rule applies from cohort 8.

Cohort	Presses	DLTs	Slope	Probability
1	1	0		
2	2	0	6.28	0.097
3	3	0	0.44	0.273
4	4	0	0.40	0.187
5	6	1	0.79	0.016
6	8	1	−0.16	0.184
7	8	1	−0.25	0.215
8	8	1	−0.22	0.170
9	8	0	−0.13	0.090
10	8	1	−0.08	0.059
11	8	1	0.05	0.018

References

Nils Aldous and Thandiwe Ferreira-Nkosi. Rider Beliefs about Hall-Call Buttons: A Survey of 2,114 Commuters. *Journal of Building Operations Research*, 7(2):101–118, 2018.

Ying Kuen Cheung. *Dose Finding by the Continual Reassessment Method*. Chapman and Hall/CRC, Boca Raton, FL, 2011.

Steven N. Goodman, Marianna L. Zahurak, and Steven Piantadosi. Some Practical Improvements in the Continual Reassessment Method for Phase I Studies. *Statistics in Medicine*, 14(11):1149–1161, 1995.

Suyu Liu and Ying Yuan. Bayesian Optimal Interval Designs for Phase I Clinical Trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(3):507–523, 2015.

John O'Quigley, Margaret Pepe, and Lloyd Fisher. Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer. *Biometrics*, 46(1):33–48, 1990.

Kjeld Sorensen and Lwazi Mbeki. Hall-Call Latching and Queue Discipline in Collective-Control Dispatch: An Analysis of Controller Event Logs. *Journal of Building Operations Research*, 10(1):33–52, 2021.

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

Peter F. Thall and John D. Cook. Dose-Finding Based on Efficacy–Toxicity Trade-Offs. *Biometrics*, 60(3):684–693, 2004.

Renata Vasquez and Pontus Linde. A Grading Scale for Bystander Remarks in Vertical-Transport Lobbies. *Journal of Building Operations Research*, 9(3):211–224, 2020.