Exact versus simulated operating characteristics of published adaptive and Bayesian trial designs

Matthew D. Schwartz*
Department of Physics, Harvard University

Draft, September 28, 2026

Abstract

Before a Phase II trial begins, sponsors and regulators judge its design on how often the trial would declare an ineffective drug effective and how often it would miss an effective one. For adaptive and Bayesian designs these operating characteristics are usually estimated from a few thousand simulated trials. Each estimate carries Monte Carlo noise, and one close to the threshold it must clear could fall on the other side in a rerun with a different seed. Morris, White, and Crowther give the standard error of such an estimate and the number of simulated trials a stated precision requires; each follows from the reported numbers alone. Applying both to the tables of nine published adaptive and Bayesian designs, we find that 351 of the 977 reported error rates and powers we classified lie within Monte Carlo noise of their thresholds. Such values were obtained from fewer simulated trials than are needed to resolve them, by a median factor of 18. For most of the reported values no simulation is needed at all, because the quantity estimated is a finite sum or a low-dimensional integral. We compute it exactly wherever the design's decision rule is stated in full, with exact rational arithmetic and cross-checked quadrature, as used for multi-loop Feynman integrals in physics. An exact value also resolves an estimate reported exactly at its threshold, where the reported margin gives no simulation size to plan and a rerun can again land too close to call. Recomputing the values within noise wherever possible, exactly or by adequate simulation, we confirm the reported side in 87 cases and find the other side in 30, about the number expected from Monte Carlo noise alone.

1 Introduction

A Phase II trial decides whether a new treatment proceeds to a confirmatory trial. Its design is judged in advance on two probabilities: that of carrying forward a treatment that does not work and that of carrying forward one that does. For adaptive and Bayesian designs these operating characteristics, the type I error rate and the power, are nearly always estimated by simulation. The authors of such a design simulate it a few thousand times under null and alternative scenarios and report the estimates in a table. They then state the conclusion the table supports, for example that the type I error rate is controlled at ten percent or that the power reaches eighty percent. Sponsors and regulators rely on those conclusions when the design is put into practice.

*Work done with Anthropic, but results and views are not endorsed by Anthropic.

Each entry in a table of simulated operating characteristics is a Monte Carlo estimate, and its standard error is fixed by the reported value and the number of simulated trials. For example, a type I error rate reported as 0.100 from 1,000 simulated trials has a Monte Carlo standard error of about 0.0095, and one of about 0.0042 if reported from 5,000 trials. At two standard errors the first estimate is compatible with a true rate 1.9 percentage points above or below a 10% threshold, and so does not by itself determine on which side of that threshold the true rate lies. The FDA's 2019 guidance on adaptive designs illustrates the precision of a simulated type I error rate with 100,000 simulated trials per scenario and states a confidence-interval condition under which fewer suffice [1]. The draft ICH E20 guideline gives the same 100,000 figure as its example [2]. Both documents ask that the number of simulated trials be reported together with a justification for it.

Reporting standards for simulation studies prescribe how many repetitions to run and how to justify the count [3, 4], and sequential Monte Carlo tests come with error probabilities controlled by construction [5, 6]. Koehler et al. [7] found the Monte Carlo error of published simulation-based analyses rarely reported and, where it could be computed, often material. Morris et al. [3] give the Monte Carlo standard error of each common performance measure and the number of repetitions a target precision requires, and recommend reporting the standard error beside every estimate. Few of the simulation studies published in *Statistics in Medicine* that they reviewed did so. White [8] provides software that computes these standard errors, Boulesteix et al. [9] set out standards for neutral comparison studies, and the RepliSims project re-ran published simulation studies in biostatistics to measure how much the reported results change on replication [10].

For a large class of designs, however, the operating characteristics need not be simulated at all. The type I error rate and power of Simon's two-stage design are finite sums of binomial probabilities and are computed exactly [11], and those of group-sequential designs are obtained from deterministic recursions [12]. Optimal adaptive two-stage designs are found by searching over operating characteristics computed without simulation [13]. In the Bayesian optimal Phase II design BOP2 the posterior stopping rule reduces to integer stopping boundaries on interim response counts [14], so that its operating characteristics are again finite sums of binomial terms. Its published tables were nevertheless obtained by simulation, like those of most adaptive and Bayesian designs. We know of no study that has measured how many of the values in published design tables lie within two Monte Carlo standard errors of their thresholds at the reported number of simulated trials. Nor has any study we know of recomputed such values by the exact methods just described.

In this paper we examine the tables of nine published adaptive and Bayesian trial designs. For every type I error rate and power that they report against a stated threshold, we first compare its distance from that threshold with its Monte Carlo standard error at the reported number of simulated trials. At the papers' own numbers of simulated trials, 351 of the 977 values classified in full lie within two Monte Carlo standard errors of their thresholds, about one in three. Across all nine papers, such values (other than those reported exactly at their threshold) were obtained from fewer simulated trials than would place them two standard errors from their threshold, by a median factor of 17.7. We then recompute, where published materials allow, each value that this comparison leaves unresolved, in the sense that it lies within two standard errors of its threshold.

For most of the nine designs the operating characteristic has the finite-sum form described above, and we evaluate it exactly, as a rational number or as an enclosure, an interval that

provably contains the exact value. For a binary-endpoint design the evaluation is a direct enumeration, or a recursion over the interim looks in exact rational arithmetic. Some of the nine are hierarchical Bayesian designs for basket trials whose sums contain, in each term, a one- or two-dimensional posterior integral, and we evaluate those integrals by quadrature checked by an independent route. An exact result carries no simulation error and resolves even an entry reported exactly at its threshold, for which the reported margin fixes no simulation size and a rerun settles the side only with a probability set by the unknown true margin. Where the value can be obtained only by simulation, we rerun the authors' released code with as many simulated trials as would place the reported value two standard errors from its threshold.

Most of the unresolved values for which the published decision rule permits exact evaluation or re-simulation are confirmed on the reported side of their thresholds. When the recomputed value falls on the other side, it still lies within about two standard errors of the reported value, and such cases occur about as often as expected from those standard errors. For the majority of the unresolved values, however, neither computation is possible, because the calibrated cutoffs, critical values, or comparator analysis rules that would be needed appear neither in the papers nor in their released code.

In Section 2 we review the finite-sum form of these operating characteristics, their Monte Carlo error, and the guidance, and in Section 3 we describe the nine papers, the classification of their table entries, and the computations. In Sections 4–6 we report the precision of the published values, the recomputation of the unresolved ones, and the reproducibility and cost of the simulations, and in Section 7 we discuss the implications for reporting and review.

2 Operating characteristics, Monte Carlo error, and regulatory guidance

There are no new results in this section. We summarize the standard material used in later sections: the finite-sum form of the operating characteristics of common Phase II designs, the Monte Carlo standard error of a simulated operating characteristic, and the FDA and ICH guidance on simulation precision.

2.1 Operating characteristics as finite sums and integrals

Simon's two-stage design (r_1, n_1, r, n) [11] stops for futility if r_1 or fewer of the first n_1 patients respond and otherwise declares the treatment promising if more than r of n respond. Under a true response rate p the probability of that declaration is

$$R(p) \;=\; \sum_{x_1=r_1+1}^{n_1} \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1-x_1} \sum_{x_2=\max(0,\,r+1-x_1)}^{n-n_1} \binom{n-n_1}{x_2} p^{x_2}(1-p)^{n-n_1-x_2}. \qquad (1)$$

The type I error rate is $R(p_0)$ and the power $R(p_1)$, at the null and alternative response rates $p_0 < p_1$. Both are finite sums of binomial terms, exact rational numbers whenever p is rational. Optimal designs are found by evaluating these sums exactly [11, 13]; group-sequential designs have an analogous stagewise recursion [12]. Any design whose decision rule depends only on a finite path of integer counts has likewise

$$\mathrm{OC}(p) \;=\; \sum_{\text{outcome paths } x} \Pr_p[\,x\,]\;\mathbf{1}\big[\text{the event the characteristic counts occurs on } x\big], \qquad (2)$$

with path probabilities that are products of binomial factors. Two conditions must hold before (2) can be evaluated in practice. First, the number of terms, as written or after reduction by the design's structure, must be within computational reach. Second, the decision rule inside the indicator (stopping boundaries, posterior cutoffs, priors, calibration constants) must be fully stated in the paper or its released code.

A Bayesian design changes only the decision rule inside the indicator of (2). BOP2 [14] stops for futility at an interim look when the posterior probability that the treatment is ineffective exceeds a cutoff that depends on the enrollment so far. With a Beta prior, that probability is a decreasing function of the number of responses observed. Thus the futility rule reduces to a table of integer stopping boundaries, from which the operating characteristics follow as sums of the form (2). Posterior cutoffs are usually calibrated: tuning parameters are adjusted until the type I error rate is as close to the nominal level as possible without exceeding it. Table entries produced by such a calibration (calibration entries) lie near their threshold by construction.

Four of the designs examined below (BLAST, BaCIS, CBHM, and Local-MEM [15–18]) are basket designs that borrow information across baskets through a shared prior on their response rates. In the usual hierarchical form the basket-level log-odds of response θ_i share a common $\mathcal{N}(\mu, \sigma^2)$ distribution with priors on its parameters. The treatment is declared promising in basket i when the posterior probability that its response rate p_i exceeds p_0 is above a cutoff. The operating characteristic is still the sum (2), but the indicator now requires, for each outcome vector x, the posterior-tail integral

$$\Pr\big(p_i > p_0 \,\big|\, x\big) \;=\; \int \mathbf{1}\big[\,p_i(\theta) > p_0\,\big]\,\pi(\theta \mid x)\,d\theta \tag{3}$$

over the model's continuous parameters θ. Local-MEM's conjugate model gives a posterior that is a finite mixture over basket partitions with closed-form weights, and the sum (2) can then be enumerated directly. For CBHM, its comparators BHM and BHM-HN, and BaCIS, the integral (3) reduces to nested one- and two-dimensional integrals that can be evaluated by quadrature. BLAST's posterior has no such reduction, and the design's operating characteristics are estimated only by simulation throughout this paper.

Designs with a survival endpoint (here the Weibull and piecewise-exponential scenarios of BayesCTDesign [19] and the two-stage design ScuRMST [20]) have no finite outcome space in general. Under the piecewise-exponential model a sufficient-statistic reduction leaves a sum over interval event counts of an integral over interval exposures, far out of reach at the sizes studied; under a Weibull model of unknown shape there is no such reduction, and the integral runs over every patient's event time (Section 5.4). The rank statistics of the continuous-endpoint ASSISTant designs [21] are enumerable exactly at a single look though not, at feasible cost, at two; Section 5.4 gives the size of the two-look enumeration.

2.2 Monte Carlo error of a reported operating characteristic

The finite sums (2) are exact but sometimes too costly to evaluate, since the number of outcome paths grows exponentially with the number of arms and interim looks, and each term of a basket design's sum requires the integral (3). In contrast, the cost of a simulated trial does not depend on the size of the design's outcome space, and with n simulated trials the error decreases as $1/\sqrt{n}$ for any design. The operating-characteristic tables of adaptive and Bayesian design papers therefore consist largely of simulation estimates. From here on we write n for the reported number of simulated trials behind a table entry (not a sample size).

Among n simulated trials the number declaring the treatment promising is binomial, and a reported proportion $\hat{p}$ has the Monte Carlo standard error [3]

$$\text{MCSE} \;=\; \sqrt{\hat{p}\,(1-\hat{p})/n}\,, \tag{4}$$

which depends only on the reported value and on n. For the Introduction's example of 0.100 reported at $n = 1{,}000$, equation (4) gives about 0.0095. Whether $\hat{p}$ can be said to meet its threshold b (the nominal type I error level or the target power stated in the design paper) depends on the margin $|\hat{p} - b|$ measured in units of MCSE. If the margin is under about two standard errors, a rerun with a different random seed could place the estimate on the other side of b with appreciable probability. The reported n is then too small to establish whether the true value is above or below b.

Solving $|\hat{p} - b| = 2\,\text{MCSE}$ for n gives the number of simulated trials at which a reported margin equals two standard errors,

$$n \;\gtrsim\; \frac{4\,\hat{p}\,(1-\hat{p})}{(\hat{p}-b)^2}, \tag{5}$$

which is the formula of Morris et al. [3] for the number of repetitions, with the required MCSE set to half the margin.[1] This n grows as the inverse square of the margin: halving the distance to the threshold quadruples it. Both (4) and (5) can be evaluated from a reported estimate, its threshold, and its n without further simulation.

2.3 Regulatory guidance on simulation precision

The FDA's 2019 guidance on adaptive designs states [1, p. 19]:

> "It is important to consider the precision of simulated operating characteristics, which depends on the number of simulated trials (iterations). The number of iterations should be sufficient to facilitate an understanding and review of the proposed clinical trial design. Using 100,000 iterations per scenario, for instance, ensures a 95% confidence interval for estimated Type I error probability with a width of approximately ± 0.1%, which would be sufficient in most cases."

and, later in the same passage [1, pp. 19–20]:

> "In some cases, fewer iterations might suffice to evaluate Type I error probability. For example, it might be sufficient to use 10,000 iterations if a particularly fine grid of scenarios is explored and every scenario has an estimated Type I error probability below the desired level. Also, a smaller number of simulations can generally be used if the upper bound of the 95% confidence interval for the Type I error probability estimate is below the desired level."

The draft ICH E20 guideline (Step 2 text, endorsed 25 June 2025) states [2, p. 22]:

[1] Equation (5) is a precision rule: it makes the reported margin two standard errors wide. If the true margin equals the reported one, a fresh run of that size places its estimate outside the 2 MCSE window on the true side only half the time; for an 80% chance of doing so the count is larger by $(2 + z_{0.8})^2/4 \approx 2.02$, and for a 90% chance by $(2 + z_{0.9})^2/4 \approx 2.69$, with z_q the standard normal q-quantile. We use the two-standard-error form throughout.

> "This also includes determining the number of repetitions needed to get sufficient precision in the estimation of important operating characteristics. More precision may be needed for certain operating characteristics or scenarios. For example, it may be important to use 100,000 or more repetitions per scenario to ensure sufficient precision for estimating the Type I error probability, whereas fewer repetitions may suffice for other operating characteristics such as power. Algorithms should be documented and random numbers should be generated in a reproducible way, such as using a documented seed."

On reporting, the FDA asks for "The number of simulated trials (iterations) evaluated for each scenario and a rationale for the adequacy of this number" [1, p. 28], and the ICH draft for "Implementation details, including the data-generating process and the number of repetitions for each scenario, along with justifications for these choices" and "Software package used for simulations and, if custom software was used, the simulation code" [2, p. 23].

In the E20 passage the figure of 100,000 is an example ("it may be important to use") within a requirement of sufficient precision. The example concerns the type I error probability, and fewer repetitions are expressly allowed for power. The E20 draft contains no bright-line repetition mandate and no confidence-interval decision rule anywhere. The confidence-interval condition appears only in the FDA guidance, and it too is permissive ("can generally be used"). In Section 3.2 we apply the FDA's interval condition at each paper's own n and report separately whether that n reaches the 100,000 repetitions of the E20 example.

3 Methods

We classified every table entry of nine design papers as resolved or unresolved at the reported number of simulated trials n, and recomputed the unresolved entries wherever the published materials allowed. The recomputation is exact where the operating characteristic is a finite sum or a low-dimensional integral with a published decision rule, and otherwise a re-simulation with as many trials as the entry's margin requires. The eligibility criteria, screening order, classification rule, and computational limits were specified in two written protocols, one for a feasibility study and one for the full study, each completed before any design paper it governed was retrieved; two later amendments are described below.

3.1 Design papers and table entries

Eligible papers propose an adaptive or Bayesian trial design, release its simulation code, and specify the simulation unambiguously. They also report simulated type I error rates or powers beside a threshold given in the paper, at a stated n [1, p. 28]. Designs whose tables are computed analytically (group-sequential, optimal two-stage) are out of scope. Candidate designs were taken from four sources screened in a fixed order (Appendix A): a feasibility study of three designs, then three public listings that yielded 7 further eligible papers. This yield was far enough below the planned 15 that, under the selection rules, every eligible paper from all four sources was included. Ten designs were thus admitted, nine contributing the 1,076 table entries (Table 1). These nine papers are the whole eligible set rather than a sample: the pooled percentages describe this corpus, and we draw no inference from them about a wider literature.

We also examined public regulatory documents: five FDA complex-innovative-design case studies, the statistical reviews of 153 Drugs@FDA approvals, and 70 European public assessment

reports (Appendix A). None contains a simulated operating characteristic beside a threshold at a stated n. One European report came closest, noting only that "insufficient exploration of its operating characteristics (e.g. effective sample size by mixing weight) prevents robust conclusions" [22]. The sentence is the regulator's and concerns the sufficiency of an operating-characteristic exploration; we make no statement about the underlying product or decision.

A table entry (a cell) is a reported simulated type I or familywise error rate or power for which the paper itself states a threshold, read from the sentence that assigns it to the entry. No regulatory default is substituted. Where that reading required judgment, the cell is marked, and all counts can be reproduced with such cells removed. Each entry is a per-trial proportion, and the product $\hat{p}\,n$ is an integer in all 1,076. A stated type I error level applies to every null scenario of its paper. A power target is attached to a power entry when the paper states it for that entry's scenario, or reads that entry's table against it in its own words; 290 of the 371 power entries carry a target of this kind. BayesCTDesign [19] states 0.80 power for one hazard or odds ratio in each of its examples and, in its text, also reads its hazard-ratio-0.6 table against 0.80 [19, pp. 30–31]; for one entry of that table (`T02_n100_a0` in Table 7) the reported and recomputed values lie on opposite sides of 0.80. The remaining 81 power entries of that paper, 21 piecewise-exponential and 60 binomial, sit at ratios the paper never compares with 0.80 in its text, and for them 0.80 is a benchmark rather than a claim the paper makes; one of the 81 lies within Monte Carlo noise of it and none is a reversal. A stricter reading, which treats the 25 hazard-ratio-0.6 entries as benchmark comparisons too (106 entries set aside), is reported beside the main counts as a sensitivity bound.

Reprints (66 of the 1,076), which repeat another table block digit for digit, and calibration entries, which the paper states were calibrated to the threshold, are marked. Every count is given with and without them. The first of the two later amendments added two Local-MEM [18] entries, found when the threshold-attachment rule was applied uniformly across that paper's tables. The second extended the calibration mark to all 150 entries of ScuRMST [20], whose released selection code enforces each entry's threshold. The direction of the change is against this paper's own headline, because it removes 150 entries, all within two standard errors of the threshold, from the non-calibration counts, lowering their unresolved share.

3.2 Classification at the reported number of simulations

Let $\hat{p}$ be the reported value as a proportion, n the reported number of simulated trials, b the paper's stated threshold (its claim boundary), and MCSE as in equation (4). Each entry is classified by the FDA's confidence-interval condition [1, pp. 19–20] applied at the paper's own n: an estimate whose 95% interval at that n contains its threshold does not determine on which side of the threshold the true value lies. The FDA sentence is written one-sided, for below-threshold claims; its extension to above-boundary claims is ours. Thus an entry is undetermined (unresolved) at its reported n when

$$|\hat{p} - b| \;\le\; k\,\mathrm{MCSE}, \qquad \text{that is,} \qquad (\hat{p} - b)^2 \;\le\; k^2\,\frac{\hat{p}\,(1-\hat{p})}{n}, \tag{6}$$

and determined otherwise, where $k = 2$ is primary (rule P), $k = 1.96$ gives the 95% Wald interval (rule F), and $k = 3$ gives a secondary tally. The right-hand form is evaluated in exact rational arithmetic, $\hat{p}$ being the rational given by its printed digits and $1.96 = 49/25$.

The Wald interval is used because it reproduces the arithmetic of the guidance's own example [1, p. 19], in which 100,000 iterations give a half-width of about ±0.1%, the Wald 95% half-

design	endpoint; decision rule	n	entries	calib.	exact method here	CPU time per trial
BOP2 [14]	binary; posterior-probability rule, integer stopping boundaries	10^4	28	28	recursion (8)	not timed[c]
BLAST[†] [15]	basket; hierarchical model (MCMC)	1,000	99	36	–	0.33–0.58 h (floor)
BaCIS [16]	basket; hierarchical model (MCMC)	5,000	93	5	quadrature (9)	≥ 5.3
CBHM [17]	basket; calibrated hierarchical model (MCMC)	5,000	159	28	recursion (8) with quadrature (9)	17.7
adaptr [23]	three-arm adaptive; conjugate beta-binomial	10^5 / 10^4	16	2	–	0.157[d]
ASSISTant [21]	continuous; rank statistics (Designs A–C)	5,000	6	4	exact rank-statistic distribution (Design C)	0.004–0.017
BayesCTDesign [19]	binomial, Weibull or piecewise-exponential outcome; Bayesian test	100 – 10^4	308	0	enumeration (binomial outcome)	0.003–0.048[a]
Local-MEM [18]	basket; conjugate multisource exchangeability model	5,000	217	28	enumeration	0.031–0.063[b]
ScuRMST [20]	survival; two-stage restricted mean survival time design	5,000	150	150	–	0.0065–0.0072

The type I error levels stated in these papers are 2.5, 5, and 10 percent and the power targets 80 and 90 percent. The one other admitted design, a platform-trial package, contributed no entries, because none of the publications cited for it was eligible (Appendix A). [†] For BLAST's 99 entries, extracted in the feasibility study before the rules of Section 3.2 were written, we report only how the margin from the threshold compares with 2 MCSE and 3 MCSE; the full classification covers the 977 entries of the other eight designs. BLAST's CPU time (floor) is a lower bound from a timing of its sampler alone. [a] The piecewise-exponential outcome (35 of the 308 entries) took 0.69 CPU-seconds per trial, and, as with the MCMC-cost designs, no 10^6-trial reference rerun was made for those 35 entries. [b] Four-basket protocols, on which the design's class rests; the six-basket protocol took 0.248 CPU-seconds per trial, the overhead of an interpreted loop rather than of a sampler. [c] Computed here with no simulated trials. [d] The cost of a conjugate model in interpreted code (10,000 posterior draws for three arms at up to 39 looks), timed on a heavily loaded machine; at this cost adaptr falls in the same class as the sampler-fitted designs under the 0.1 CPU-second rule, and we did not re-time it on an idle machine, where a lower measured cost might have changed the class.

Table 1: The nine design papers examined, the reported number of simulated trials n behind their table entries, and the exact method, if any, that replaces simulation for each design in this paper. Columns: design and primary reference; endpoint and decision rule; n per entry; entries with a paper-stated threshold; calib., how many of those are calibration entries (stated by the paper as calibrated to that threshold, or forced to it by released selection code); the exact method applied here (–, none); and the measured CPU time per simulated trial with the released code, in seconds except where marked h (hours). Each timed design's cost class follows from that cost alone: a design under 0.1 CPU-seconds per simulated trial is a cheap-simulation design, and one at or above it an MCMC-cost design. BOP2, not timed because we evaluate it without simulation, is classed as exactly enumerable. Reference values, exact or from 10^6-trial reruns, were computed only for exactly enumerable and cheap-simulation designs.

width at $\hat{p}$ near 0.025, namely $1.96\sqrt{0.025 \cdot 0.975/100{,}000} = 0.00097 \approx 0.1\%$. Wilson's interval is numerically indistinguishable from Wald at that n and $\hat{p}$; the FDA example therefore does not determine the interval family. We make no Wilson or Clopper–Pearson substitution in the rules themselves. A Clopper–Pearson recount is reported as a sensitivity analysis. As a separate indicator we note whether a type I error entry meets the ICH E20 example's precision ($n \geq 100{,}000$) and its 95% interval excludes b (rule E). For power, where E20 gives no number, rule E is the interval condition alone. Because the E20 figure is an example rather than a requirement, the rule-E counts indicate only how far reported practice is from it.

Each rule assigns one of four outcomes: undetermined, determined but unverified (no reference value), determined and confirmed (a reference value on the same side of b), or flipped (a reference value strictly on the other side). Undetermined takes precedence: a cell inside its rule's window whose reference value lies on the opposite side is classified as undetermined and not as flipped, because the paper never determined it. A reference value is either an exact value or an estimate V from a re-simulation with the authors' code at $n_{\text{ref}} = 10^6$ trials. An exact value places the entry above or below b when its enclosure excludes b; the estimate V does so only if

$$|V - b| \;>\; 4\,\text{MCSE}_{\text{ref}}, \qquad \text{MCSE}_{\text{ref}} = \sqrt{V(1-V)/n_{\text{ref}}}\,. \tag{7}$$

An estimate inside that window is rerun once at 10^7 trials and is not used as a reference value if still inside. A reported $\hat{p}$ of exactly 0 or 1 has MCSE = 0 and is taken as reported.

Repeating all 3,228 comparisons of (6) (three values of k for each of the 1,076 entries) in double precision reproduces the exact-arithmetic classification in every case. The exact arithmetic therefore changes no result here; we use it so that each classification follows from the printed digits alone and does not depend on floating-point rounding. The classification is descriptive and no hypothesis test is performed, in part because the entries cluster within papers (ScuRMST alone contributes 14% of the 1,076).

No design, table, verdict, or approval is said to be wrong anywhere in this paper, in any sense. We assess only whether printed cells are determined at the papers' own simulation counts, that is, whether the printed side of a claim boundary is fixed at the published precision, or could change under a different random seed with appreciable probability. Sentences of the form "this design should have been built differently" or "this trial should not have been approved" do not appear here and are not implied.

3.3 Exact computation

Exact evaluation of two-stage binomial designs by enumeration is classical [11–13]. For the enumerations and quadratures below we use the rigorous enclosure techniques of multi-loop integral computation in mathematical physics, namely dynamic programming with exact rational weights, threshold comparisons in ball (interval) arithmetic, and quadrature checked by an independent route. For a binary-endpoint design the sum (2) is evaluated directly with exact rational binomial weights, or by recursion over interim looks [12]. With n_k the cumulative sample size at look $k = 1, \ldots, K$, s the cumulative responses, and $P_k(s)$ the probability of reaching look k with s responses, $P_1(s) = \text{Bin}(s;\, n_1,\, p)$ and

$$\begin{aligned} P_{k+1}(s') \;&=\; \sum_{s} P_k(s)\, \mathbf{1}\big[\text{continue at } (k,s)\big]\, \text{Bin}\big(s' - s;\; n_{k+1} - n_k,\; p\big), \\ \text{OC}(p) \;&=\; \sum_{k=1}^{K} \sum_{s} P_k(s)\, \mathbf{1}\big[\text{declare at } (k,s)\big], \end{aligned} \tag{8}$$

where the indicators are read from the published stopping-boundary table or cutoff rule. With several arms, s is a vector and the weights are products of binomials. This recursion was used for BOP2 [14], whose integer stopping boundaries are either published or fixed by its released code, and for CBHM. Local-MEM's conjugate mixture over 15 basket partitions was enumerated in full over its 20^4 outcome space, as were that paper's Simon two-stage comparator columns and BayesCTDesign's binomial-outcome entries.

For the hierarchical basket designs the declaration indicator in (2) depends on the posterior tail probability (3). When the between-basket variance σ^2 is fixed by the data, as in CBHM, the basket log-odds are conditionally independent given the hyper-mean μ. The tail (3) is then a ratio of nested one-dimensional integrals,

$$\Pr\big(p_i > p_0 \mid x\big) \;=\; \frac{\int \pi(\mu)\, T_i(\mu) \prod_{j \neq i} L_j(\mu)\, d\mu}{\int \pi(\mu) \prod_j L_j(\mu)\, d\mu}, \quad L_j(\mu) \;=\; \int \mathrm{Bin}\big(x_j;\, n_j,\, p(\theta)\big)\, \mathcal{N}(\theta; \mu, \sigma^2)\, d\theta,$$
$$T_i(\mu) \;=\; \int \mathbf{1}\big[\, p(\theta) > p_0 \,\big]\, \mathrm{Bin}\big(x_i;\, n_i,\, p(\theta)\big)\, \mathcal{N}(\theta; \mu, \sigma^2)\, d\theta, \tag{9}$$

with $\pi(\mu)$ the prior on μ, $p(\theta)$ the response rate at log-odds θ, and x_j the number of responses among the n_j patients of basket j. A prior on σ, as in CBHM's BHM comparator, adds an outer dimension, and BaCIS's posterior tails reduce to nested integrals of the same dimensions. We evaluate these integrals by quadrature checked by an independent route, the two agreeing near the cutoffs to a few parts in 10^6. A state whose computed tail lies within 10^{-4} of its cutoff is carried as unresolved and its probability is added to the width of the resulting interval; an entry is placed above or below its threshold only when that interval excludes the threshold. For CBHM's sensitivity configurations, whose cutoffs are not published, the cutoffs were recomputed exactly from the design's own calibration rule. That rule's calibration convention can be read in two ways, and results are reported under both. Under the first, the at-most reading, the cutoff is calibrated to give an exact type I error rate of at most the 10% target. Under the second it is the next achievable cutoff, whose rate lies just above the target.

A third exact method applies to ASSISTant's single-look Design C [21]: because its rank statistic is distribution-free, the integral over the data, of dimension about 500, reduces to a finite sum over rank configurations. One entry computed this way lies 2.4×10^{-5} from its threshold, a margin that would take about 1.3×10^9 simulated trials to resolve. No exact method is available for the Weibull and piecewise-exponential scenarios of BayesCTDesign or for ASSISTant's multi-look Designs A and B (Section 5.4), and those entries were re-simulated; BLAST has none either, and its entries were left at their reported precision. Each exact evaluator above agrees with an independent brute-force or second-quadrature calculation and, where the design's authors released a sampler or simulation code, with its output within Monte Carlo error (Appendix C, Table C1).

3.4 Reference simulation and the expected number of reversals

No reference simulation was run for BLAST, BaCIS, CBHM [15–17], or adaptr [23], because their released code takes 0.1 CPU-seconds or more per simulated trial (Table 1). For the other designs, wherever no exact method exists, we re-simulated each unresolved entry with the released code at the n its margin requires by equation (5), up to a maximum of two million repetitions. A cell whose margin is still inside its noise at that maximum is reported as knife-edge and is not forced to a side. These entries are listed in Appendix D, together with two entries decided by runs

beyond that maximum. Fourteen of the unresolved BayesCTDesign entries are reprints of other entries of that paper, each of which we recomputed. They were given those entries' outcomes (a decision for thirteen, unresolved for one) and were not counted as computed.

A recomputed value on the other side of the threshold from the reported value is a reversal. As an illustrative model, we treat the reported value $\hat{p}_i = X_i/n_i$ of entry i as the proportion of successes in n_i binomial draws and place a uniform prior on the true value p_i, so that $p_i \mid X_i \sim \text{Beta}(X_i + 1,\, n_i - X_i + 1)$; q_i is the mass of this distribution on the other side of b_i from $\hat{p}_i$. The number of reversals R among decided entries is then modeled as

$$R \;=\; \sum_i B_i, \qquad B_i \sim \text{Bernoulli}(q_i) \text{ independently}, \qquad E[R] \;=\; \sum_i q_i\,, \tag{10}$$

with a 90% interval from the distribution of R. The model has no fitted parameter and was specified before any entry was recomputed. It is illustrative rather than exact in three respects: entries whose cutoff was calibrated on the same simulations are not simple binomial observations; entries estimated from a common batch of simulated trials are not independent, which leaves $E[R]$ unchanged but makes the 90% intervals too narrow, by an amount we illustrate in Appendix E under a stated block structure; and the comparison with observed counts is restricted to entries that could be decided. Replacing the uniform prior by the Jeffreys prior changes each expected count by less than 3%, and replacing the beta tail by a normal tail gives nearly the same expected number (Appendix E).

Data and code availability. The supplementary material gives each entry's reported value, threshold sentence, reported n, reprint and calibration marks, classification and reference value, and each recomputation in detail. It also holds the required n behind Table 4, the study protocols with amendments, and the analysis code that regenerates every count and Figure 1.

4 Reported precision in nine design papers

Whether a table entry is resolved at its own Monte Carlo precision depends on three published numbers: the reported value $\hat{p}$, the reported number of simulated trials n, and the threshold b that the paper states for it. From them, equations (4) and (6) classify each of the 1,076 entries as determined or undetermined at its reported n, by exact arithmetic on the reported digits, and we refer to this classification as the census. Equation (5) then gives the n that the entry's margin would have required. Every count in the census depends on one quantity, the margin $|\hat{p} - b|$ in units of the entry's MCSE. The ICH-derived indicator (rule E) adds a condition on n itself.

4.1 Entries within Monte Carlo error of their thresholds

The nine papers contain 1,076 entries with a paper-stated threshold and a reported n (Table 2). At the papers' own n, 422 of these entries (39%) lie within 2 MCSE of their threshold and 505 (46%) within 3 MCSE (Figure 1). Such percentages are rounded down throughout (Appendix B). When we set aside the 281 calibration entries, which were tuned to lie near the threshold, 795 entries remain, of which 167 (21%) lie within 2 MCSE and 239 (30%) within 3 MCSE. In particular, all 28 entries of BOP2 [14] are calibration entries and therefore do not enter the without-calibration counts. Dropping the 66 reprints (12 of them BLAST's) leaves 1,010 distinct

entries, with 382 within 2 MCSE and 454 within 3 MCSE. The lower block of Table 3 omits reprints in the same way.

Eight of the nine designs, with 977 entries, were classified in full (Table 3). For BLAST [15], whose entries were extracted in the feasibility study (Appendix A) before rules P, F, and E were written, only the margin comparisons of Tables 2 and 4 were made. Of the 977 entries, 561 are determined at their reported n and carry no reference value. A further 65 are determined and confirmed by a reference value on the reported side of the threshold. No determined entry has a reference value on the other side, and the flipped class is accordingly empty. Because no entry in these papers has a margin between 1.96 MCSE (the half-width of the FDA's 95% interval [1]) and 2 MCSE, rules P and F classify every entry identically. Under either window the other 351 of the 977 entries (about one reported claim in three, calibration entries included) are undetermined at their reported n. Without the 81 BayesCTDesign power entries for which 0.80 is a benchmark (Section 3.1), 350 of 896 entries (39%) are undetermined; under the stricter reading that also sets aside that paper's 25 hazard-ratio-0.6 entries, 346 of 871 (40%).

ScuRMST [20] selects its critical values on the same simulated data sets from which it reports type I error and power. Its search applies filters that force the reported type I error into $(0.95\,\alpha_0, \alpha_0)$ for a nominal level α_0. All 150 of its entries are therefore calibration entries and lie within 2 MCSE by construction. Of these, 53 equal the threshold exactly, and the largest margin in its tables is 1.21 MCSE. In CBHM [17], simulated at $n = 5{,}000$, four entries reported as 0.109 have $|\hat{p} - b| = 0.0090$ against a 2 MCSE window of 0.008815 and are thus determined, though they clear the window by only 0.0002. BLAST calibrates to a 10% type I error rate and reports its operating characteristics at $n = 1{,}000$, where 2 MCSE at $\hat{p} = 0.10$ is ± 0.019, a window that contains 35 of its 63 non-calibration entries. Notes on the other six designs are in Appendix B.

One scenario in the nine papers was simulated at the 100,000 repetitions that the draft ICH E20 guideline gives as an example [2], namely the null scenario of adaptr [23], whose type I error is reported twice. Every other type I error entry was published at 10^4 repetitions or fewer. When the E20 example count is added as a condition on type I error entries (rule E), 729 of the 977 entries are undetermined (Table 3). They comprise 604 type I error entries below the example count and 125 power entries that fail the interval condition. The 248 entries determined under this rule are 246 power entries, to which only the interval condition applies, and the 2 entries of the adaptr null scenario. The count of 729 reflects the distance between the reported n and the E20 example and is not a rate of noncompliance with any mandate.

4.2 Required versus reported numbers of simulations

For each of the 422 entries within 2 MCSE of its threshold we computed M, the smallest n at which the reported value would clear that window. This M follows from equation (5) applied to the reported digits in exact arithmetic, and Table 4 summarizes it by design beside the reported n. Among the 345 entries with a finite M, the median M is 45,697 and the range 134 to 4,731,997, whereas their reported n ranges from 100 to 10,000. Taken entry by entry, the ratio M/n of required to reported simulations has median 17.7 and quartiles 4.3 and 71.4. By equation (5) this ratio is four for an entry reported one MCSE from its threshold and four hundred for one reported a tenth of an MCSE away. The largest ratio, 946, belongs to a ScuRMST type I error entry reported as 0.0498 against a threshold of 0.05 at $n = 5{,}000$, for which $M = 4{,}731{,}997$. For 103 entries M exceeds 100,000, more than any n reported in the nine papers. A further 77 entries equal their threshold exactly (53 in ScuRMST, six in CBHM, and 18 across six of the

		all entries				without calibration	
design	n per entry	entries	≤ 2 MCSE	≤ 3 MCSE	calibration	≤ 2 MCSE	≤ 3 MCSE
BOP2	10^4	28	13	21	28	0	0
BLAST†	1,000	99	71	76	36	35	40
BaCIS	5,000	93	14	22	5	9	17
CBHM	5,000	159	51	59	28	23	31
adaptr	10^5 / 10^4	16	0	2	2	0	0
ASSISTant	5,000	6	6	6	4	2	2
BayesCTDesign	100 – 10^4	308	72	110	0	72	110
Local-MEM	5,000	217	45	59	28	26	39
ScuRMST	5,000	150	150	150	150	0	0
total		1,076	422	505	281	167	239

Table 2: Across all nine designs (1,076 entries, BLAST included), about two in five reported values lie within 2 MCSE of their own threshold, and about one in five once the calibration entries are set aside. Columns: reported number of simulated trials n; entries (reported values $\hat{p}$ with a paper-stated threshold b); entries within 2 MCSE and 3 MCSE of b, where MCSE = $\sqrt{\hat{p}(1-\hat{p})/n}$; calibration entries (stated as calibrated to b, or forced to b by released selection code); the two counts without them. Of the 281 calibration entries 255 lie within 2 MCSE; the other 26 lie conservatively below b (nine in Local-MEM; the adaptr null, 4.8% against 5% at n = 100,000, reported twice; fifteen in BOP2, by its rule "closest to and not higher than the nominal value"). †Not among the 977 entries of Table 3.

rule	undetermined	determined, unverified	determined, confirmed	flipped
all 977 entries				
P (paper's own precision)	351	561	65	0
F (FDA-derived)	351	561	65	0
E (ICH-derived)	729	248	0	0
deduplicated: each printed value counted once, 923 entries				
P (paper's own precision)	323	535	65	0
F (FDA-derived)	323	535	65	0
E (ICH-derived)	696	227	0	0

Table 3: Of the 977 entries classified under all three rules (all designs of Table 2 but BLAST), 351 are undetermined at their reported n under rules P and F, which agree entry for entry, and none is flipped under any rule. Undetermined: $|\hat{p} - b| \leq k$ MCSE, with $\hat{p}$ the reported value, b the paper-stated threshold, MCSE = $\sqrt{\hat{p}(1-\hat{p})/n}$, and $k = 2$ (P) or $k = 1.96$ (F, the FDA 95% interval). Determined entries are confirmed or flipped when a reference value (exact, or a 10^6-repetition rerun) lies on the reported or the opposite side of b, and unverified otherwise. Rule E adds $n \geq 100{,}000$ for type I error entries; the draft ICH E20 guideline gives that figure as an example, and the counts in row E therefore reflect the distance of the reported n from that example. Lower block: each printed value counted once, dropping the 54 digit-for-digit reprints.

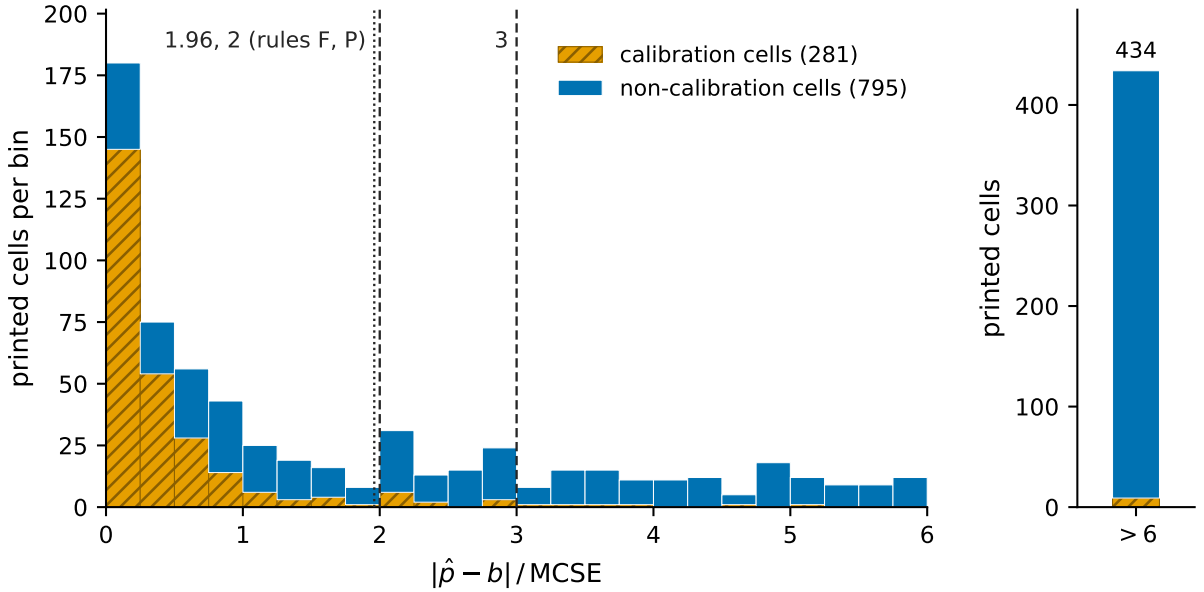

Figure 1: Reported margins concentrate within two Monte Carlo standard errors of the threshold; 422 of the 1,076 entries (39%) lie at or below 2 MCSE and 505 (46%) at or below 3 MCSE. Left: histogram of $|\hat{p} - b|/\mathrm{MCSE}$ ($\hat{p}$ the reported value, b the paper-stated threshold, $\mathrm{MCSE} = \sqrt{\hat{p}(1-\hat{p})/n}$ at the reported number of simulated trials n; bin width 0.25; the axis labels say "cells" for entries). Calibration entries (hatched; calibrated to b by the paper's statement or by its released code) concentrate below 1 by construction; 167 of the 795 non-calibration entries still lie at or below 2. Dotted line, 1.96 (the FDA 95% interval); dashed lines, 2 and 3. Right: the 434 entries beyond 6, including 9 calibration entries and the 9 entries reported as exactly 0 or 1 (zero Wald MCSE, taken as reported).

other designs), and for them equation (5) gives no finite M: the reported value alone sets no simulation size, a fresh run places such an entry off its threshold with a probability governed by the unknown true margin, and an exact value settles the side outright or shows the equality to be exact.

4.3 Reference values and sensitivity to the interval convention

Reference values (exact values, or estimates from reruns at 10^6 repetitions that satisfy equation (7)) exist for 76 entries in three designs, and 74 of the 76 are exact. Local-MEM [18] has 64, all by exact enumeration of its conjugate model, 50 of them conditional on decision cutoffs recovered from the paper's released code at the seeds it sets. BOP2 has 10, by the exact recursion of equation (8) over its integer stopping boundaries, 2 of them conditional on designs recovered by rerunning its released code. BayesCTDesign [19] has 2, for its two determined entries nearest their threshold, both reported as 0.0458 against 0.05; reruns at 10^6 repetitions gave 0.047922 and 0.048074. These reruns were made in order of increasing margin until the computational limit for that design was reached (Appendix A). The other six designs have none. Four (the MCMC-cost class) have per-trial simulation costs that rule out a reference run and were classified at their reported precision only; the other two, ASSISTant [21] and ScuRMST,

				M required ($\delta > 0$)				
design	n	entries	≤ 2 MCSE	median	min	max	$M > 10^5$	$\delta = 0$
BOP2	10^4	28	13	21,697	13,757	356,797	1	5
BLAST	1,000	99	71	88,397	1,054	363,197	21	7
BaCIS	5,000	93	14	21,697	5,222	363,197	2	1
CBHM	5,000	159	51	88,397	6,022	363,197	16	6
adaptr	10^5 / 10^4	16	0	–	–	–	0	0
ASSISTant	5,000	6	6	45,697	45,697	49,297	0	1
BayesCTDesign	100 – 10^4	308	72	3,415	134	161,197	2	1
Local-MEM	5,000	217	45	39,397	6,886	356,797	5	3
ScuRMST	5,000	150	150	158,797	13,757	4,731,997	56	53
total		1,076	422	45,697	134	4,731,997	103	77

design	entry	$\hat{p}$	b	n	M	M/n
ScuRMST	Table 1, minimax, log-rank test, type I error α_E (1/1.5, 0.05–0.8)	0.0498	0.05	5,000	4,731,997	946
ScuRMST	eleven power entries reported as 0.801 (Tables 1–3)	0.801	0.80	5,000	637,597	128
BLAST	Table 1, arm A1, independent, type I error	10.1%	10%	1,000	363,197	363
BayesCTDesign	power scenario `T05_n175_a0`	0.798	0.80	500	161,197	322
CBHM	Table 1, scenario 1, BHM, subgroup 1 (one of 77 $\delta = 0$)	0.100	0.10	5,000	∞	–

Table 4: The reported numbers of simulated trials fall short of what the reported margins require by a median factor of 17.7. Top, by design: reported n; entries; entries with margin $\delta = |\hat{p} - b| \leq 2$ MCSE ($\hat{p}$ the reported value, b the paper-stated threshold, MCSE $= \sqrt{\hat{p}(1-\hat{p})/n}$); median, minimum, and maximum, over those with $\delta > 0$, of M, the smallest number of simulated trials at which $\hat{p}$ would clear the 2 MCSE window by equation (5); entries with $M > 10^5$; entries equal to the threshold itself ($\delta = 0$), for which equation (5) gives no finite M. With the 1.96 MCSE window the pooled median M shifts by less than 4% (to 43,887) and no conclusion changes. Bottom: the most extreme entries, located by the design paper's own table, design, and column or scenario labels, with reported $\hat{p}$, b, and n, the required M, and the ratio M/n. Every M follows by arithmetic from the reported digits and the reported n; no new simulations were run.

have no determined entries.

Of the 76 reference values, 65 fall on determined entries, 56 in Local-MEM, 2 in BayesCTDesign, and 7 in BOP2. Each of these 65 reference values lies on the reported side of the threshold, and the corresponding entries are the confirmed entries of Table 3; no determined entry is contradicted. Because the 561 other determined entries carry no reference value, that statement rests on those 65 cells alone. All three designs with reference values belong to the exactly enumerable and cheap-simulation classes, and the statement does not extend to the MCMC-cost class, which has no reference values.

The other 11 reference values fall on undetermined entries, 8 in Local-MEM and 3 in BOP2. We find ten of them on the reported side, the nearest 0.00063 from its threshold. The eleventh belongs to the null entry of scenario 3 in BOP2's Table II, reported as 10.0%, exactly the 0.10 threshold. The exact type I error of that design is 0.10121, above the nominal level and 0.40 MCSE from the reported value. BOP2 selects the cutoff whose type I error is "closest to and not higher than the nominal value." The estimate from the paper's simulation at the seed set in its released code meets that convention and the exact value does not. The stopping boundaries of scenario 3 are not printed, and we recovered them by rerunning the paper's released code at that seed; the exact value is therefore conditional on a logged recovery of the design's code. Three of the eight Local-MEM values on undetermined entries belong to the paper's calibrated global-null entries, reported as 0.097, 0.098, and 0.098. Their exact values are 0.0958, 0.0979, and 0.0994, all within 2 MCSE of the 0.10 target, as one might expect of calibrated entries.

We ran three exploratory sensitivity analyses of how far the counts in Tables 2 and 3 depend on the conventions adopted; none of the three is used to alter a reported tally. First, we perturbed every reported value up and down by half a unit in its last printed digit and recomputed the margin and MCSE in exact arithmetic. The perturbation changes the 2 MCSE status of 23 of the 1,076 entries and the 3 MCSE status of 15 (35 distinct entries, 3.3%). Since every entry was perturbed at once, these changes are upper bounds on how far rounding could move any count. Second, the nine entries reported as exactly 0 or 1 have zero Wald MCSE, and we take them as reported; their exact Clopper–Pearson intervals exclude the threshold (Appendix B), and no determination depends on the zero-MCSE convention.

Third, we replaced the Wald window by the exact two-sided 95% Clopper–Pearson interval at the reported n, which changes 9 of the 1,076 entries (0.8%), all from determined to undetermined. The number of entries whose interval contains the threshold would move from 422 to 431, and the undetermined count from 351 to 359 of 977. Eight of the nine changed entries belong to the 977 classified in full, and the ninth is a BLAST entry. Four of the eight are among the 65 confirmed entries and would become undetermined, although their reference values lie on the reported side. Of the 351 undetermined entries, the eleven with a reference value stay undetermined under the precedence rule, and we recompute all 351 in Section 5, the second stage of the study. No count in this section is revised on the basis of the second stage.

5 Recomputation of the unresolved entries

Exact computation applies to most of the entries examined. For 597 of the 977 (61 percent), those of BOP2, BaCIS, CBHM, Local-MEM, and BayesCTDesign's binomial-outcome scenarios, the operating characteristic is a finite sum of the form (2) or a low-dimensional integral of the form (3). The figure of 597 counts cells by design class, whether or not a value was computed for them here. It includes comparator columns whose decision rules are unpublished and which

remain uncomputed. Wherever the decision rule is published we computed the value exactly; the values so obtained include the 74 exact reference values of Section 4.3 and give the 67 exact decisions of this section. Where the value can be obtained only by simulation, we re-simulated at the n that the entry's own margin requires by equation (5).

5.1 Outcome of the recomputation

Table 5 gives the outcome for each of the 351 entries unresolved at the reported n. For 152 of them the published materials were sufficient to begin a recomputation, and for the other 199 they were not. Of the 152, we decided 126. In 87 the recomputed value confirms the reported side, and in 30 (the reversals) it falls on the other side of the threshold. One of the confirmations and none of the reversals is a BayesCTDesign entry whose threshold is a benchmark (Section 3.1). Under the stricter reading that also sets aside that paper's 25 hazard-ratio-0.6 entries, three further confirmations and one reversal are set aside; that reversal, `T02_n100_a0`, is an entry whose reported 0.776 the BayesCTDesign paper itself reads as short of 0.80 [19, p. 30]. The remaining 9 were reported exactly at the threshold, and only their recomputed values place them above or below it: eight exact values and, for an ASSISTant entry whose exact law is too large to evaluate, a reference simulation of 200,000 trials, which places it 4.8 MCSE below (Appendix D). In each of the 30 reversed cells the computed value lies within 2.2 standard errors of the printed one, so none is an error by the paper concerned. The reference computation is not of the same kind in all thirty. In twenty it evaluates the decision rule and cutoffs of the paper's own instance, exactly or by re-simulating the released code; in five (two CBHM and three BHM entries) it evaluates the cutoffs distributed with the code, which the paper does not print and which need not equal its calibrated values; and in five it evaluates cutoffs we re-derived from the paper's calibration rule. Only the first twenty compare two evaluations of one quantity; the other ten compare the printed estimate with the exact value of the design as released or as recalibrated (Table 7). Among the confirmations, 2 are of the second kind and 9 of the third. For five of them the side obtained depends on which of two readings of a calibration convention is adopted (Table 7). By method, 67 of the 126 were decided exactly (Table 5, lower block); thirteen are reprints, which take the outcome of the entry they repeat.

Of the 152, a further 11 remain within 2 MCSE of their threshold at the largest number of simulated trials run for each and are left unresolved. Another 15 ended otherwise: 12 BaCIS comparator entries, for which a needed input proved to be unpublished, and three BOP2 entries. The BOP2 entries are the three that already carry exact reference values in the census while undetermined. Two of those values lie on the reported side; the third, for the scenario-3 entry of BOP2's Table II, reported exactly at its threshold, lies above it. The three entries keep that status and are not counted as decided here. In all, 140 of the 351 entries provide enough information to reproduce and adjudicate. The other 211 (60%) cannot be recomputed from published materials by any method.

The middle block of Table 5 lists by cause the 211 entries that cannot be recomputed. ScuRMST [20] accounts for 150 of them: the critical values of the designs in its tables are not archived, and the code version that generated those tables is not in the release (Section 6). CBHM's Independent comparator [17] accounts for 25: its analysis rule is documented neither in the paper nor in the release. Local-MEM [18] contributes 14: 11 comparator entries for which no simulator is released and no cutoff is stated in the paper, and three sensitivity configurations of the two-stage version of the proposed design, whose cutoffs are likewise unstated and for which the release has no calibration routine. BOP2 [14] contributes 10: for eight comparator

entries the calibrated cutoffs of the two-stage comparators are not given and no comparator code is archived, and two designs of an appendix table have unstated stopping boundaries. These boundaries cannot be regenerated from the archived seed, because the released code run at that seed does not reproduce the one design of that table whose boundaries are stated.

BaCIS's traditional hierarchical comparator [16] accounts for the last 12, the only entries among the 211 for which a computation could be started. The released package implements the proposed design only, and the paper gives neither the comparator's hyperparameters nor its thresholds. The supplement that might contain them could not be obtained through any channel we tried, and we did not attempt to reconstruct the missing parameters by guessing. BLAST's release [15] also contains no comparator code, and CBHM's covers the BHM comparator but not its other two (Table 8). Where a comparator's decision rule is unpublished we computed no value, because the census rules do not permit reimplementing a missing or ambiguous protocol element. Although the six ASSISTant entries [21] were costly to recompute, each was treated in full, and none is among the 211. These blockages are not a consequence of the study's compute budget. Rather, each of the 211 lacks a specific published input: critical values, a cutoff or threshold, or a comparator analysis rule. This figure describes the public record as acquired on the study's dates.

5.2 Three recomputations in full

The simplest of the three recomputations concerns the Simon two-stage design used as a comparator in Table 3 of the Local-MEM paper [18]. Its power and type I error rate are given by equation (1) and are single rational numbers, $0.800947\ldots$ and $0.023403\ldots$. This design was selected under feasibility constraints (type I error rate at most 0.025, power at least 0.80), and its exact power lies above 0.80 by construction. The paper reports this power in ten entries as 0.792 to 0.806 at $n = 5{,}000$, where MCSE ≈ 0.006. The five entries below the target, 0.792 to 0.798, are Monte Carlo draws of a quantity whose exact value is 0.80095, and each is a reversal. In general, for a design selected or calibrated to meet a constraint, any reported estimate on the other side of that constraint is necessarily a reversal, a mechanism we refer to as constraint calibration. The two type I error entries of the same design are reported as 0.025, exactly at their threshold; the exact value, 0.0234, lies below it.

The second recomputation concerns the calibration scenario of CBHM [17], whose released code is distributed with calibrated posterior cutoffs. The global-null row of its Table 1 is the calibration scenario itself. That row reports type I error rates of 0.099, 0.098, 0.098, and 0.100 in the four subgroups at $n = 5{,}000$, against the advertised 10% target. We evaluated the recursion (8) at the distributed cutoffs, with each posterior tail (3) computed by the quadrature of Section 3.3. The exact type I error rate lies within $[0.103073, 0.105752]$ in subgroups 1 and 3, $[0.108407, 0.108718]$ in subgroup 4, and $[0.086851, 0.096344]$ in subgroup 2 (Table 6). Thus in three of the four subgroups the enclosure lies above the 10% target.

The printed values are consistent Monte Carlo at the paper's count of 5,000, all within about two standard errors of the exact values; nothing about the paper's table is in error. The paper's own cutoff values, produced by a parallel-seed calibration path, are unarchived and not recoverable from the public record, so the printed table describes the paper's calibrated instance while the exact enclosures describe the design as distributed, with the decision rule applied to the exact posterior. The released code applies that rule instead to a posterior tail estimated from 30,000 retained sampler draws, and run at that chain length on every dataset near the cutoff it reproduces the printed value: it decides differently from the exact posterior on datasets

Outcome for the 351 entries unresolved at the reported n	
Computation could be started from published materials	152
decided	126
reported side confirmed	87
recomputed value on the other side of the threshold (reversal)	30
reported exactly at the threshold; side assigned by the computation	9
within 2 MCSE of the threshold at the 2×10^6 maximum; no side assigned	11
ended otherwise: three BOP2 entries treated as reference values, and 12 BaCIS comparator entries for which a needed input was shown to be unpublished	15
No computation could be started	199
Enough published information to reproduce the entry (the 152 less the 12 BaCIS comparator entries)	140
Not recomputable from published materials by any method (the 199 and the 12 BaCIS comparator entries)	211
The 211 entries not recomputable, by cause	
ScuRMST: critical values of the reported designs not archived; generating code version not in the release	150
CBHM, Independent comparator: analysis rule documented neither in the paper nor in the release	25
Local-MEM: comparator simulator not released and comparator cutoffs unprinted (11); sensitivity configurations with unprinted cutoffs (3)	14
BOP2: two-stage comparators' calibrated cutoffs unprinted and no comparator code (8); appendix designs with unprinted stopping boundaries (2)	10
BaCIS, traditional hierarchical comparator: no public implementation; hyperparameters and thresholds unprinted; supplement unobtainable	12
The 126 decided entries, by method	
Exact computation (rational value or enclosure)	67
Re-simulation at the n the reported margin requires, up to 2×10^6	43
Simulation continued past 2×10^6	2
Reference simulation of an entry reported exactly at its threshold (2×10^5 trials)	1
Reprints of another entry of the same paper, taking that entry's outcome	13

Of the 77 entries in the nine papers reported exactly at their threshold, nine are placed above or below it in this section; Appendix D accounts for the other 68 by design.

Table 5: Of the 351 table entries unresolved at the reported number of simulated trials n (those whose reported estimate $\hat{p}$ lies within 2 MCSE of its threshold, MCSE $= \sqrt{\hat{p}(1-\hat{p})/n}$), 140 can be recomputed from published materials and 126 are decided, while 211 (60%) cannot be recomputed by any method (upper block). The middle block gives the 211 by design and cause, each cause an input (critical values, a cutoff or threshold, or an analysis rule) absent from the publications as they stood when we obtained them; the lower block gives the method behind each decision.

design	subgroup	reported	exact type I error rate	side of 0.10
CBHM	1	0.099	$[0.103073, 0.105752]$	above (reversal)
	2	0.098	$[0.086851, 0.096344]$	below (confirmed)
	3	0.098	$[0.103073, 0.105752]$	above (reversal)
	4	0.100	$[0.108407, 0.108718]$	above (side assigned)
BHM	1	0.100	$[0.10091, 0.10167]$	above (side assigned)
	2	0.098	$[0.10091, 0.10167]$	above (reversal)
	3	0.098	$[0.10328, 0.10407]$	above (reversal)
	4	0.098	$[0.10328, 0.10407]$	above (reversal)

Table 6: At the cutoffs distributed with its code, CBHM's exact type I error rate misses the advertised 10% target in three of the four subgroups of its calibration scenario, while every value reported there [17] (global-null row of that paper's Table 1; $n = 5{,}000$ simulated trials; MCSE ≈ 0.0042 near 0.10) lies within about two MCSE of its enclosure. The paper's own calibrated cutoffs are not archived; the enclosures describe the design as distributed. Lower block: the BHM comparator of the same row. Entries reported as 0.100 are placed above or below 0.10 only by their enclosures; subgroups 1 and 3 (upper block) share a cutoff and hence an enclosure.

carrying 1.6 percent of the probability and gives a subgroup-1 type I error rate of 0.098, against 0.104 under the exact posterior and 0.099 as printed. The difference between the printed row and the enclosures is therefore of the size produced by the finite chain, not a feature that more simulated trials would remove, and it does not require cutoffs different from the released ones. No design, table, or approval is said to be wrong.

How far CBHM's type I error rate lies from its target depends on the fourth decimal of the cutoff. When a common cutoff is swept from 0.885 to 0.895, the exact type I error rate falls from 12.7 to 8.2 percent. About two of those percentage points lie within the single step of one thousandth that contains the distributed cutoffs, and a calibration with 5,000 simulated trials cannot locate the cutoff within that step. With the same computation we obtained enclosures for all thirty proposed-design entries of the paper's main table. For the 25 of these entries that were determined at the reported n but had no reference value, the enclosures confirm the reported side. These second-stage confirmations are not added to the confirmed column of Table 3. In the scenarios with heavy borrowing the reported values lie three to five standard errors from the exact ones, consistent with the paper's tables having been produced with cutoffs or a sampler that differ from the release's.

We computed the calibration row of the CBHM paper's BHM comparator by the same method (Table 6, lower block). The subgroup-1 entry, reported exactly at 0.100, has an enclosure above 0.10, and the other three entries are reversals, each reported value lying within 1.35 standard errors of its enclosure. For CBHM's remaining comparator, BHM-HN, we re-derived the cutoff exactly from the design's calibration rule under the at-most reading, and its four calibration entries are then confirmed on their reported side. Their numerical enclosures extend across the boundary, so the decision rests on the convention, and under the opposite reading of it their sides would invert, like those of the five recalibrated reversals.

The reported values of BaCIS [16] also trace to one parameter setting while a published source specifies another, in this case to a default of the released code and not to the parameter its paper states. The paper gives the threshold parameter $\theta_c = 0.5$, whereas the released function

by default computes an adaptive θ_c from the data. We computed the operating characteristics exactly under both choices. The reported values lie 0.1 to 3.5 MCSE from the exact values under the adaptive default and 1.9 to 43.6 MCSE from those under the stated $\theta_c = 0.5$. The two cells decided here are confirmed on their printed side under both parameter choices, so no classification in this paper depends on which one is used. For BLAST [15] the reported values likewise trace to the sampler configuration of its release (Section 6). In each case the printed values are consistent with an ordinary Monte Carlo draw under the parameter choice they trace to, as shown by exact computation or by running the authors' own code at large simulation counts.

The third recomputation concerns the tests used in BayesCTDesign [19] and ASSISTant [21], both with nominal level 0.05. For a binomial endpoint the size of BayesCTDesign's CLT-based test is an exact finite sum. Over the recomputed binomial-endpoint entries it oscillates between 0.0488 and 0.0516 as the sample size changes. For example, one binomial scenario with 480 patients, reported as 0.0481, has exact size 0.05110. One Weibull scenario analyzed with the same test, reported as 0.044, has estimated size 0.05100 at 2×10^6 simulated trials. Both entries are reversals, listed in Table 7 under test discreteness.

Under the piecewise-exponential model the BayesCTDesign paper reports sizes of 0.048 to 0.068 across its sample sizes, which could be read as a drift of the type I error rate with sample size. Re-simulated at the n their margins require, the six size entries give 0.0494 to 0.0503, that is, Monte Carlo noise around 0.0500 ± 0.001 at every sample size. At this resolution the test holds its nominal level. The six entries nevertheless remain within 2 MCSE of 0.05 and are among the eleven left unresolved (Appendix D). For two Weibull entries we continued the simulation past the 2×10^6 maximum, and both confirm their reported side; for one of them the reported value of 0.058 is not itself reproduced (0.0505 ± 0.0001). ASSISTant's single-look Design C has an exact combinatorial law for its rank statistic, which gives 0.05002 where 0.048 is reported, a difference of 0.66 standard errors.

5.3 Reversals and their expected number

Table 7 lists the thirty reversals among the 126 decided entries by the three mechanisms that produce them. Ten arise from constraint calibration: the five Local-MEM Simon power entries of the first recomputation above and five entries of CBHM's sensitivity table. For the latter the cutoff re-derived under the at-most reading achieves an exact size of 0.0998 to 0.1000, at or below the 10% target, and the reported values of 0.101 and 0.102 therefore sit on the opposite side. These five decisions would invert under the opposite reading of that design's calibration convention. Under the opposite reading the cutoff is the next achievable value and the exact sizes are 0.1000 to 0.1003; Table 7 marks the five as recalibrated. Two reversals, the binomial and Weibull entries of the third recomputation, arise from test discreteness.

The remaining eighteen reversals are Monte Carlo side errors: reported values that are ordinary Monte Carlo noise at the reported n and whose recomputed value falls across the threshold. Seven are Weibull or piecewise-exponential entries of BayesCTDesign, and three are reprints of decided entries of that paper. Six of the eighteen, two of them reprints, are power entries reported as 0.804 to 0.812 whose re-simulated values are 0.7699 to 0.7746, the largest differences among the thirty. The two CBHM calibration entries and the three BHM comparator entries of the second recomputation are of this kind as well; the BHM entries are reported as 0.098 against enclosures of 0.1009 to 0.1041. The final three are ASSISTant entries, two decided by Design C's exact law and one by simulation at the n set by its margin.

For an entry reported exactly at its threshold the condition of equation (6) holds at the reported n and equation (5) gives no count for a rerun: a fresh simulation places such an entry off its threshold with a probability set by the unknown true margin, whereas an exact value, which has no Monte Carlo error, settles the side outright. Of the 77 such entries in the nine papers, nine were placed above or below their thresholds in this section, eight by exact values and one by a reference simulation. Four of the eight are the two Simon type I error entries reported as 0.025 and the two entries reported as 0.100 in Table 6, treated in the first two recomputations above. A fifth is a binomial-endpoint entry reported as 0.0500 whose exact size is 0.04877. The last three are entries of CBHM's sensitivity table reported as 0.100. Their re-derived cutoffs achieve sizes below the target (0.0998 to 0.1000 when rounded to four decimals), with the same dependence on the calibration convention as the five recalibrated reversals. The ninth is an ASSISTant Design A entry reported as 5%, whose exact law is too large to evaluate (Section 5.4); a reference simulation of 200,000 trials with the released code gives 0.0477, or 4.8 MCSE below the threshold. The other 68 threshold-valued entries are accounted for in Appendix D.

The number of reversals can be compared with the number expected from the Monte Carlo standard errors alone under the illustrative binomial model of equation (10), whose intervals assume independent entries and are therefore lower bounds on the spread (Appendix E). The recomputation was carried out in two parts, and for the 219 entries judged computable at the outset, the model gives 58.6 expected reversals, with 90% interval $[49, 69]$. Restricted to the 84 of them that were decided in the first part, the same model gives 18.8 ($[13, 25]$). Over the 94 that were decided or left unresolved it gives 21.3 ($[15, 28]$). The sixteen reversals that occurred among them lie inside both intervals. The gap between 58.6 and 16 reflects the composition of the 219-entry pool and is not evidence about the model's calibration, because that pool was drawn up under three expectations that were not met. That pool included BLAST's entries, whose reference simulations were not run (Table 8), entries that needed the exact evaluators for BaCIS and for the re-derived CBHM cutoffs, which were built only later, and the comparator entries, which proved not to be computable (Appendix E).

A second comparison is out of sample: the same model, applied to the 32 further entries of the second part before any of them was recomputed, gives 8.52 expected reversals ($[5, 13]$). Over the nineteen of them that were decided it gives 5.9 ($[3, 9]$ if the entries were independent; the nineteen come from five simulation scenarios, nine of them from one, and treating each scenario's entries as perfectly dependent widens this to $[0, 14]$); eight reversals occurred, five of them the convention-dependent CBHM decisions of that one scenario. The other reversals of Table 7, the three reprints and the three BHM comparator entries, lie outside both comparisons: their outcomes were fixed by other entries or had already been computed when the second forecast was specified (Appendix E).

5.4 Limits of exact computation

We could not replace simulation by an exact computation for the survival-endpoint statistics of BayesCTDesign or for the rank statistics of ASSISTant's Designs A and B. In the released BayesCTDesign code, which fits six hazard intervals and censors every patient administratively at one fixed time, the piecewise-exponential decision depends on the data only through a 19-dimensional sufficient statistic: the six interval event counts pooled over arms, the number of events in the treated arm, and the six interval exposures (total time at risk) of each arm, except on an event of probability below 10^{-6} in which some interval holds fewer than two events in

design	entry	quantity	$\hat{p}$	b	recomputed	method
constraint calibration: the feasible Simon design's exact power is 0.80095						
Local-MEM	Table 3, Simon, 2-succ., C	power	0.794	0.80	0.80095	exact
Local-MEM	Table 3, Simon, 2-succ., D	power	0.792	0.80	0.80095	exact
Local-MEM	Table 3, Simon, 3-succ., B	power	0.797	0.80	0.80095	exact
Local-MEM	Table 3, Simon, 4-succ., B	power	0.798	0.80	0.80095	exact
Local-MEM	Table 3, Simon, 4-succ., D	power	0.797	0.80	0.80095	exact
constraint calibration: the recalibrated cutoff's achieved size is at most the target (CBHM's Table 3, scen. 1)						
CBHM	config. (0.5, 80), sg. 1	type I	0.102	0.10	0.09981	exact, recal.
CBHM	config. (1.0, 50), sg. 2	type I	0.101	0.10	0.09990	exact, recal.
CBHM	config. (1.5, 80), sg. 2	type I	0.101	0.10	0.09997	exact, recal.
CBHM	config. (1.5, 80), sg. 3	type I	0.101	0.10	0.09997	exact, recal.
CBHM	config. (1.5, 80), sg. 4	type I	0.101	0.10	0.09997	exact, recal.
test discreteness: the exact size of the test oscillates around the nominal level as the sample size changes						
BayesCTDesign	`T11_n480_e1`	type I	0.0481	0.05	0.05110	exact
BayesCTDesign	`T04_n75_a0`	type I	0.044	0.05	0.05100	sim, 2×10^6
Monte Carlo side error at the reported n						
BayesCTDesign	`T01_n70`	power	0.81	0.80	0.77118	sim, 10^5
BayesCTDesign	`T02_n100_a0`	power	0.776	0.80	0.80336	sim, 10^5
BayesCTDesign	`T03_n150_a1`	power	0.804	0.80	0.76993	sim, 1.6×10^5
BayesCTDesign	`T03_n175_a0.5`	power	0.816	0.80	0.79710	sim, 10^5
BayesCTDesign	`T04_n125_a0.25`	type I	0.052	0.05	0.04498	sim, 2.0×10^5
BayesCTDesign	`T05_n175_a0.5`	power	0.804	0.80	0.77298	sim, 1.6×10^5
BayesCTDesign	`T10_n450_e0.8`	power	0.812	0.80	0.77463	sim, 10^5
BayesCTDesign	`T07_r1_a1`	power	0.804	0.80	0.76993	inherited
BayesCTDesign	`T08_r1_a0.5`	power	0.816	0.80	0.79710	inherited
BayesCTDesign	`T08_r1.2_a0.5`	power	0.804	0.80	0.77298	inherited
CBHM	Table 1, scen. 1, subgroup 1	type I	0.099	0.10	$[0.10307, 0.10576]$	exact, release
CBHM	Table 1, scen. 1, subgroup 3	type I	0.098	0.10	$[0.10307, 0.10576]$	exact, release
CBHM (BHM)	Table 1, scen. 1, subgroup 2	type I	0.098	0.10	$[0.10091, 0.10167]$	exact, release
CBHM (BHM)	Table 1, scen. 1, subgroup 3	type I	0.098	0.10	$[0.10328, 0.10407]$	exact, release
CBHM (BHM)	Table 1, scen. 1, subgroup 4	type I	0.098	0.10	$[0.10328, 0.10407]$	exact, release
ASSISTant	Design C, `R7C3`	type I	0.048	0.05	0.05002	exact
ASSISTant	Design C, `R7C6`	type I	0.048	0.05	0.05002	exact
ASSISTant	Design A, `R7C4`	type I	0.052	0.05	0.04621	sim, 4.9×10^4

Table 7: The thirty reversals by mechanism: entries unresolved at the reported number of simulated trials n whose recomputed value lies across the threshold b from the reported estimate $\hat{p}$. Entries are located by the design paper's own table and scenario, BayesCTDesign entries by that paper's scenario identifiers, and ASSISTant entries by their census identifiers. Each $\hat{p}$ is within 2.2 Monte Carlo standard errors $\sqrt{\hat{p}(1-\hat{p})/n}$ of the recomputed value; every row is thus within the Monte Carlo noise at that n, and none indicates an error in the paper concerned. Exact values are rationals or quadrature-based intervals (Appendix C), identical where two arms have the same cutoff or the same exact law. Rows marked release evaluate the cutoffs distributed with the design's code, which its paper does not print and which need not equal the paper's own calibrated values; rows marked recal. give the achieved size of the re-derived cutoff under the at-most reading and would invert under the opposite one; inherited rows are reprints of other rows; simulation rows state the repetitions used, at most 2×10^6.

an arm and the code re-places its cutpoints. The seven counts are integers, but the twelve exposures are continuous, because the code draws event times as real numbers and never rounds them. The exact operating characteristic is therefore not a finite sum of the form (2) but a sum, over the lattice of interval event counts in the two arms, of a twelve-dimensional integral over the exposures. It is well defined and deterministic, and it is impractical: at the smallest sample size in the paper's grid (400 per arm) the count lattice alone retains about 10^{20} points after configurations carrying less than about 10^{-6} of the probability in each arm are discarded, about 10^{10} per arm, and every point requires its own integral of the fitted-model indicator. No published analytic approximation has an error bound sharp enough to take its place. The normal approximations with explicit constants for maximum-likelihood statistics are first order [24, 25], and the Edgeworth expansion of the needed order has no computable constants [26]. The Berry–Esseen bound with the best explicit multivariate constant [27] evaluates, at this dimension and sample size, to at least 258 where the decision needs 10^{-4}. Under the Weibull model with unknown shape parameter the data admit no sufficient statistic of fixed dimension, so no such reduction is available at all: the exact operating characteristic of a Weibull cell remains a deterministic integral of the decision indicator over the censored event times of every patient, of dimension twice the per-arm sample size (80 to 350 in the paper's grid), well defined but beyond any quadrature we could apply.

ASSISTant's multi-look Designs A and B, unlike its single-look Design C, have exact laws too large to evaluate by the enumeration we constructed, a dynamic program that inserts patients in rank order and carries the running rank sums at each look. Its state count for the smallest exact sub-problem grows with the number of patients N as about $N^{6.6}$, measured at seven successive problem sizes; this is the cost of that enumeration, not a lower bound over algorithms. It reaches 8×10^{13} states at the first two interim looks, and the full laws are larger by many orders of magnitude. For these statistics an exact computation is thus well defined in every case but lies far beyond the reach of the reductions and algorithms we found, by more than any factor that a larger compute budget would close.

6 Reproducibility from published code and computational cost

The classification of equation (6) uses only each reported estimate, its n, and its threshold, and does not require the authors' code. Whether a paper's tables can be regenerated from its released code, and at what cost, depends instead on what the release contains. Unless the release fixes the random-number seed [3], a rerun can agree with a reported value only within Monte Carlo error.

6.1 Replication at archived seeds

Three of the nine papers publish, with their code, the seeds of their reported simulations or a per-trial seeding convention (Table 8): adaptr [23], BayesCTDesign [19], and Local-MEM [18]. The BOP2 paper [14] gives no seed, but its authors' archived package code sets one internally (`set.seed(1024)`). Local-MEM's released driver reseeds each simulated trial deterministically. A rerun reproduced 144 of 144 proposed-design entries of the main and sensitivity tables at the reported three decimals and recovered the unprinted calibrated cutoffs. The six-basket table (147 entries) was not rerun. BayesCTDesign prints its seeds in its example calls, and reruns at those seeds reproduced all 308 entries to the last digit. For adaptr the seeds and a complete executable specification are published. We did not rerun adaptr's simulations, because by its

measured cost per simulated trial (Table 1) it belongs with the sampler-fitted designs, which were assessed at their reported n only.

Run unmodified, the archived BOP2 code reproduces the scenario-3 and scenario-4 blocks of Table II of Zhou et al. [14] in full and the scenario-1 block in 11 of 12 entries. The rerun also recovers the unprinted boundaries of scenarios 3 and 4. The unreproduced entry lies in one of the two rows of that table whose printed acceptance and early-termination percentages sum to more than 100%; the true percentages cannot, because a trial that accepts has not stopped early. Appendix C gives both rows (the only two of 128 such printed pairs with this property) and the exact operating characteristics of the design whose stopping boundaries are printed in the BOP2 paper's appendix Table DII. The scenario-2 boundaries could not be recovered from the code. Over a widened grid of the design's two tuning parameters the code produces 350 distinct boundary vectors at the archived seed, none of which reproduces the printed block. In that paper's appendix Table AI the one block whose boundaries are printed (scenario 1) is likewise not reproduced at that seed: 3 of its 4 true-rate rows differ in the rerun.

The other five papers, BLAST [15], BaCIS [16], CBHM [17], ASSISTant [21], and ScuRMST [20], provide no seed at which their reported values can be regenerated. ScuRMST's released notebooks do set a seed, but the driver for the proposed design's table rows stops with an error (`object 'fil_combs' not found`). The release also contains two code versions with different search procedures (some reported design rows come from neither) and omits the critical values that define each reported design. Hence, as far as we can determine, no single ScuRMST entry can be re-evaluated without rerunning the whole design search.

6.2 Cost of reference simulation for sampler-fitted designs

Of the four designs for which no reference simulation was run (Section 3.4), BLAST, BaCIS, and CBHM fit their model by Markov chain Monte Carlo in every simulated trial. We timed each release and scaled the cost to the paper's full per-trial settings and n. Each total below is therefore an extrapolation from a short timing run: ten to twelve simulated trials for CBHM and BaCIS, and for BLAST a sampler-level timing and one complete trial at a tenth of the paper's chain length. For BLAST the sampler-level timing gives a floor of 0.33 to 0.58 CPU-hours per simulated trial, and the complete trial, extrapolated linearly in chain length, gives 18.3 CPU-hours. One scenario at the paper's $n = 1{,}000$ then costs at least 330 to 580 core-hours by the floor and roughly 18,000 by the direct measurement. Table 8 gives the BaCIS and CBHM totals: about 112 core-hours for the BaCIS design's own rows, and about 456 to 1,644 for the 67 CBHM scenario-method combinations examined.

However, most of the rerun cost may lie in the released implementations rather than in the designs. We reimplemented BLAST for a single graphics processing unit (GPU; an NVIDIA H100), with the authors' own sampler component translated from their released source, and checked the reimplementation against their unmodified code (Appendix C). Against the 0.33 to 0.58 CPU-hour floor, one GPU runs 124 to 217 times faster than one core running the authors' code. At the paper's full chain length the reimplementation takes 2.668 GPU-hours per 1,000-trial scenario row. For the 24 BLAST entries that a reference value could decide, a reference simulation at that speed (the production run) would take at most 300 GPU-hours, about a thousand dollars of rented GPU time. The thousand dollars is an extrapolation, at the rate we paid, from a measurement that used 1.8 GPU-hours, about six dollars. We did not make the production run, because no claim in this paper needs it; BLAST's 24 decidable cells keep their census classifications.

design	seed in the release	entries reproduced at the seed, of entries rerun	comparator entries without released code	CPU time per simulated trial[a]	core-hours to rerun the paper's own table at its n
BOP2	fixed internally by the authors' archived package code	35 of 36 (Table II: 12 of 12, 12 of 12, 11 of 12); scenario-2 design not recoverable; the paper's appendix Table AI not reproduced at the seed	–	exact evaluator	–
BLAST	none mapping to the reported values	not rerun	66 of 99	0.33–0.58 h (floor); 18.3 h (direct)	at least 330–580 per 1,000-trial scenario (floor); roughly 18,000 per scenario (direct)
BaCIS	none mapping to the reported values	not rerun	62 of 93	$\geq$ 5.3 s	about 112 (proposed design's rows); about 336 (full table)
CBHM	none mapping to the reported values	not rerun	60 of 159 (two of three comparators)	4.9–17.7 s[b]	about 456–1,644 (the 67 scenario-method combinations at $n = 5{,}000$); about 1,361–4,908 per combination at 10^6
adaptr	seeds archived; complete executable specification printed	not rerun	–	0.157 s	about 6.3; about 260 for a 10^6-repetition reference per scenario
ASSISTant	none mapping to the reported values	–	–	0.004–0.017 s	–
BayesCTDesign	seeds printed in the example calls	308 of 308 (3 first-example entries; grids of 170, 35, 50 and 50)	–	0.003–0.048 s	about 35 (projected, serial grids)
Local-MEM	per-trial reseeding convention in the released driver	144 of 144 (proposed design, main and sensitivity tables); six-basket table (147 entries) not rerun	comparator columns	0.031–0.063 s	about 1.6 (the 144 entries)
ScuRMST	seed archived; released driver stops with an error; critical values not archived	none reproducible from the release	–	0.0065–0.0072 s	–

Table 8: Reruns at an archived seed reproduce reported entries to the last digit (144 of 144, 308 of 308, 35 of 36); five releases fix no usable seed; the four costlier designs were not rerun, and the last column gives the extrapolated cost of doing so. Column 4: comparator entries with no released implementation. Column 6: core-hours (3,600 CPU-seconds) to regenerate the paper's table at its reported number of simulated trials n (for the designs not rerun, an extrapolation from short timing runs). Dashes: not applicable or not measured. [a] As measured for Table 1 at the paper's full settings. [b] User time alone (4.9) to user plus system time (17.7).

In reimplementing BLAST we also found that, under the null scenario, the decisions of the released two-cluster model depend measurably on which latent cluster receives which label. The assignment is fixed by a random-number draw early in the burn-in. Runs differing only in seed can therefore end with either of the two labelings, whose decision rates differ. In a sixty-seed comparison the authors' code as released ends with each labeling at rates indistinguishable from the reimplementation's, so this dependence on the labeling is a property of the released implementation.

7 Discussion

When the decision rule of a Phase II design is published in full, its operating characteristics can often be computed exactly. The comparison of an exact value with its threshold then involves no Monte Carlo error. Where simulation remains necessary, the Monte Carlo standard error of each reported estimate determines whether it can be distinguished from a nearby threshold.

For the nine designs examined, most reported operating characteristics are finite sums or low-dimensional posterior integrals. We evaluated these exactly wherever the published decision rule was complete. About one in three of the reported type I error rates and powers lies within two Monte Carlo standard errors of its threshold (Table 3). For such an entry the reported number of simulated trials is too small to establish whether the true value is above or below the threshold. More simulated trials would be needed, by a median factor of 17.7 over the entries not reported exactly at their threshold (Table 4). In the classification underlying these findings (the census), we compared each reported value, at its number of simulated trials, with the precision conventions of the FDA and ICH guidance [1, 2]. The census is, to our knowledge, the first measurement of published design tables against those conventions; earlier work measured Monte Carlo error reporting practice generally [3, 7, 8, 10], and we claim no other priority.

We recomputed the unresolved entries wherever published materials allowed, exactly or by re-simulation at the number of repetitions their margins require. Most of the recomputed entries were confirmed: the new value was on the same side of the threshold as the reported estimate. Where it fell on the other side, it lay within about two Monte Carlo standard errors of the reported value (Table 7). These reversals occurred about as often as expected from the standard errors, which suggests that Monte Carlo noise in the reported estimates accounts for them. However, three in five of the unresolved entries cannot be recomputed from what is published, exactly or by simulation. The calibrated cutoffs, critical values, or comparator analysis rules on which those entries depend are given neither in the papers nor in their released code.

The nine papers were included because each publishes its code, a table with stated thresholds, and the number of simulated trials used. Few candidate papers offer all three. The corpus is small and was selected for checkability; its pooled proportions therefore refer to papers that publish enough to be assessed. We do not estimate the rates among papers that publish less. Extraction and classification were done by a single group, with a duplicate extraction and a scripted recount (Appendix A). Moreover, the supplementary tables of CBHM (19 further scenarios) and of BLAST (two further basket configurations) were unobtainable [15, 17], so for those papers the census covers the main-text tables only.

The interval in equation (6) is the Wald interval at the reported number of simulated trials, chosen to match the FDA's worked example. Using the exact Clopper–Pearson interval instead changes fewer than one percent of the classifications, all of them from determined to undetermined (Section 4.3). Each entry was compared with the threshold that its paper states

for it, and identifying that threshold required judgment in three cases: the familywise entries of BaCIS, whose target sentence names a single comparator; the difference scenarios of adaptr, whose threshold sentence is hedged; and two Local-MEM entries added when the attachment rule was applied uniformly.

The count under the ICH-derived indicator (rule E) reflects only how far the reported numbers of simulated trials fall below the illustrative figure of the draft guideline. The four designs whose released code was too costly per simulated trial for a reference run (BLAST, BaCIS, CBHM, and adaptr) were classified at their reported precision but not re-simulated. Among their unresolved entries, those for which no exact evaluator was available remain so. The classification and the recomputation are two stages of a single assessment. Both stages concern the printed tables only, and neither assesses the papers' prose claims, the designs' merits, or any decision made on them. Deciding the true side of a printed cell says nothing about what any design, paper, or approval should have done.

For the design classes examined, the results suggest a practical reporting standard of four items, two of them already in the guidance [1, 2]. First, the number of simulated trials can be chosen in advance, by equation (5), from the smallest distance to a threshold that the authors need to resolve, and reported with that calculation. Second, the Monte Carlo standard error can be printed beside each simulated operating characteristic, as Morris et al. [3] recommend for simulation studies generally. This item is perhaps the most useful to a reviewer, who can apply the condition of equation (6) by comparing the estimate's distance from its threshold with twice the reported standard error. For a value reported exactly at its threshold, the standard error gives the tolerance within which the design attains that threshold.

Third, the decision rule can be published in full, with its calibrated cutoffs, critical values, and comparator analysis rules, together with the random seed. Where both were available, tables could be regenerated from the released code and unresolved entries recomputed (Table 8). When the published rule was incomplete, no such recomputation was possible. A later release of ScuRMST's critical values, documentation of CBHM's Independent comparator, or an accessible copy of the BaCIS supplement would make the corresponding cells computable, and the same rules would apply to them. By contrast, adaptr's complete decision rule and seeds are published with its code, as the guidance describes. Its null scenario is also the only one among those examined that was simulated at the guidance's illustrative 100,000 repetitions [23].

Fourth, where the operating characteristic is a finite sum or a low-dimensional posterior integral, its exact value can be reported in place of an estimate for the cost of the arithmetic alone, using enumeration or the recursion of equation (8). Where simulation is unavoidable, Table 8 indicates what a reference run at an adequate number of repetitions would cost for the designs examined here. We suggest that a reporting item asking for the Monte Carlo standard error beside every simulated operating characteristic, and for the exact value wherever the design admits one, may suffice for any reader to repeat this assessment.

A Selection of the nine papers, the regulatory-document survey, and the study protocol

The nine design papers of Table 1 are all the eligible papers found in four sources of candidates, each screened in a fixed order under a stated stopping rule (Table A1). We present the sources and the screening below, and then the survey of public regulatory documents, which was carried out before the selection. In the last subsection we describe how the rules of the study were written and amended, the computational limits, and the duplicate extraction of the table entries.

A.1 Sources and screening

Each source is a fixed list of candidates, taken in order until a stated number of eligible papers has been found or the list ends. A design already admitted from an earlier source is skipped as a duplicate, and a candidate whose published version was paywalled and not obtained does not count toward the stated number. The eligibility requirement that a paper specify its simulation unambiguously is met whenever equation (6) can be evaluated from the reported values and the reported number of simulated trials. For example, ScuRMST [20] meets this requirement although the critical values of the designs in its tables are not published. A paper can therefore meet it without publishing the decision rule behind each entry in enough detail for the entry to be recomputed. Entries of this kind that are unresolved at the reported n appear in Table 5 as not recomputable from published materials.

The first source was a feasibility study of three designs, chosen in advance to span the range of simulation cost. They were BOP2 [14], taken from a public software catalog (trialdesign.org); a platform-trial analysis package, NCC; and the earliest published eligible Bayesian-borrowing basket design in the same catalog. That design is BLAST [15], published online 2018-01-14; CBHM, published online 2018-03-02, was also eligible but appeared later. The other three sources were the catalog itself, the CRAN *ClinicalTrials* task view, and a PubMed query. These three sources gave BaCIS [16], CBHM [17], adaptr [23], ASSISTant [21], BayesCTDesign [19], Local-MEM [18], and ScuRMST. Table A1 accounts for every other candidate screened, among them six excluded because their operating characteristics are analytic rather than simulated and 12 whose published versions were paywalled and not obtained.

None of the three publications cited for the NCC package was eligible. For this package the selection rule also requires that its released code can generate the candidate publication's simulation setting. One of them describes a historical-controls simulation setting that the package's own data generators cannot produce. A second [28] states its number of simulated trials (100,000) and its one-sided level but reports every simulated operating characteristic in a figure rather than a table. The third [29], which was paywalled when screened and was obtained later through library access, reports a table of simulated operating characteristics with a stated threshold, one-sided $\alpha = 0.025$, at 10,000 simulated trials per scenario. However, the NCC code cannot generate the third paper's patient allocation and exit structure or its response-scale drift. That paper was therefore excluded and NCC contributes no table entries, which is a limit of the selection rule rather than of the paper.

The written rules set a target of 15 designs. They also provided that, if the second, third, and fourth sources yielded fewer than 8 eligible papers, the eligible set would be taken whole under the same rules and the result reported as a note on that set. The three sources yielded 7 $(2 + 3 + 2)$, so the rule applied, and this paper is that note.

source	candidates	order; stopping rule	eligible designs	other candidates screened
Feasibility study	three designs	chosen in advance to span simulation cost	BOP2[a], BLAST	NCC platform package: none of its three cited papers eligible (see text); admitted, 0 entries
Software catalog (trialdesign.org)	42 entries	page order; stop at six eligible (not reached; all 42 screened)	BaCIS, CBHM	19 not traceable to a source paper; 9 paywalled, not obtained; 8 no threshold-carrying table; 2 duplicates (BOP2, BLAST); 2 analytic
CRAN *ClinicalTrials* task view	130 packages	alphabetical; stop at three eligible (reached at package 15)	adaptr (position 4), ASSISTant (11), BayesCTDesign (15)	ahead of package 15: two paywalled, not obtained (adaptIVPT, basksim); three analytic (adoptr, adpss, baskexact); packages 16 to 130 not examined
PubMed query[b]	9 records	ascending date; all 9 screened	Local-MEM, ScuRMST	one duplicate (CBHM); four no threshold-carrying table; one analytic; one paywalled, not obtained (a 2014 guide to Bayesian group-sequential designs)
All four sources			10 designs admitted, 9 with table entries	12 paywalled, not obtained (9 in the catalog, 2 in the task view, 1 in the PubMed query); six analytic (two, three, and one in the same sources)

[a] Paper paywalled at selection and obtained later through institutional library access. [b] For adaptive or Bayesian designs with operating characteristics, simulation, and public code.

Table A1: Every eligible design found in the four sources was included: the three sources screened after the feasibility study yielded 7 eligible designs against a target of 15, and together with BOP2 and BLAST from the feasibility study these give the nine papers of Table 1. A candidate whose published version was paywalled and not obtained does not count toward a source's stopping rule. "Analytic" marks candidates whose operating characteristics are computed by recursion, quadrature, or a closed form, outside the scope of the study; "no threshold-carrying table" means no table of simulated operating characteristics with a stated type I error or power threshold (figures do not qualify).

A.2 Regulatory documents

Before the design papers were selected we asked whether any public regulatory document could serve in place of a design paper. Three criteria were used: (a) a full simulation protocol, (b) an operating characteristic within about 2 MCSE of its regulatory threshold at a stated number of repetitions, and (c) recomputability from public text alone. We read five of the six public FDA complex-innovative-design case studies in full, excluding the one read before the criteria were fixed. We then screened the posted statistical reviews of the Drugs@FDA original NDA and BLA approvals from 2026-07-16 back to 2025-07-02, 153 application records in all. The European public assessment reports of the 70 most recent EMA initial marketing authorizations, from 2026-06-19 back to 2025-09-17, were screened in the same way.

A total of 228 documents were examined, none of which met any of the three criteria. In the two screenings, no statistical review or assessment report both described an adaptive or Bayesian

design and reprinted the sponsor's simulation-based operating characteristics. Only the five case studies could therefore be assessed against the criteria in detail, and all five failed all three. The only numbers of repetitions stated anywhere in these documents are "repeated 2,000 times" and ">1000 times". The first appears in a passage that gives no operating-characteristic values for the design concerned, and the second refers to an interim predictive-probability computation. Hence no Monte Carlo standard error could be computed for any document.

A.3 Protocol, amendments, and computational limits

The rules of the study were written in two documents, one for the feasibility study and one for the full study, each completed before any candidate paper it governed was retrieved. Amendments were restricted to reclassification of an entry or a paper under the stated conventions and to increases in a computational limit. Two were made, both reclassifications (the two added Local-MEM entries and the ScuRMST calibration entries described in Section 3.1), and no limit was increased.

Apart from BOP2, the designs were assigned to the cost classes of Table 1 by their measured cost per simulated trial alone. BOP2 was placed in the exactly enumerable class without being timed, because an exact evaluator could be built for it. For each design selected after the feasibility study we timed at least ten simulated trials with the released code at the paper's full simulation settings. Designs costing under 0.1 CPU-seconds per simulated trial form the cheap-simulation class and those at or above it the MCMC-cost class. BLAST was classed as MCMC-cost from its feasibility-study runs at reduced Markov-chain length, which give a lower bound on its cost per simulated trial (Tables 1 and 8). In the full study we limited computation in advance to 2,000 core-hours in all. The per-design limits were 5, 300, and 25 core-hours for the exactly enumerable, cheap-simulation, and MCMC-cost classes, respectively. When a design reached its limit the computations still outstanding were dropped in a prespecified order.

Every table entry was extracted from the published tables twice, independently, and all counts and classifications were also recomputed from the per-entry data in exact arithmetic. These two steps led to corrections for three designs (Local-MEM, ASSISTant, and ScuRMST). No correction changed a classification, and for every other design the two extractions agreed. All counts in this paper are those obtained after the corrections.

B Per-design notes on the classification

Table 2 gives the classification counts for each design. The calibration column in particular depends on conventions stated in the design papers themselves, such as how a calibrated cutoff was chosen. In this appendix we give those conventions and other per-design notes, and state how the percentages quoted in the main text were rounded. We also list the entries reported as exactly 0 or 1 and those nearest the 2 MCSE line.

B.1 Per-design notes

The 28 entries of BOP2 [14] are type I error rates under null scenarios, 14 for BOP2 itself and 14 for its two-stage comparators, against a 0.10 threshold. All are calibration entries under the convention that paper states: a cutoff giving exactly 0.1 "does not usually exist," and the authors select the cutoff whose type I error is "closest to and not higher than the nominal value." Thirteen lie within 2 MCSE of the threshold and 21 within 3 MCSE, and every entry outside the

2 MCSE window falls below the threshold, as that convention requires. The calibrated cutoffs of the comparator designs are not reported, and without them no reference value could be computed for their entries.

BaCIS [16] states thresholds of 0.10 for the type I error rate (including the familywise rate) and 0.90 for power. Its released package implements the proposed design only and therefore provides no reference value for the 62 comparator entries. The 14 difference-scenario type I error entries of the adaptr paper [23], at 10^4 repetitions, lie outside the 2 MCSE window and are unresolved under rule E through its condition on n. ASSISTant [21] contributes six null-scenario entries against a 5% threshold. All six are unresolved under each of the three rules, the largest margin being 0.66 MCSE, and one is reported as exactly 5%.

BayesCTDesign [19] contributes the most entries, 308, or 273 after reprints are removed. Its tables of type I error and power appear in the running text as verbatim console output, and we treated them as numeric tables. Local-MEM [18] has a conjugate model, which allowed us to enumerate its outcome space exactly. The power entries of its Simon two-stage comparator design are the subject of the first of the three recomputations in Section 5.2.

B.2 Rounding, zero-error entries, and near-window entries

Percentages in the main text are rounded in the direction unfavorable to the finding that many reported values lie within Monte Carlo error of their thresholds. Thus shares of entries lying within 2 MCSE or 3 MCSE of their threshold are rounded down, and 505 of 1,076 (46.93%) is shown as 46%. Shares that qualify that finding (for example, the fraction of all entries contributed by a single paper) are rounded up, and ScuRMST's 150 of 1,076 (13.94%) is shown as 14%.

Nine entries are exactly 0 or 1 as reported: two adaptr type I error rates of 0.0% at 10^4 repetitions and seven BayesCTDesign power values of 1 at $n = 500$. Their Wald MCSE is zero, and by convention each is then taken at its reported value under equation (6) and classified as determined. Their exact 95% Clopper–Pearson intervals, $[0, 0.00037]$ against 0.05 and $[0.9926, 1]$ against 0.80, exclude the threshold in all nine cases, and these classifications therefore do not depend on the zero-MCSE convention. Among the entries whose classification changes under the half-unit perturbation of reported values described in Section 4.3 are those nearest the 2 MCSE line: the four CBHM entries at 0.109, the two BayesCTDesign entries at 0.0458 (2.01 MCSE from their threshold), and the five BOP2 entries reporting 9.4% (2.06 MCSE from 0.10).

C Exact evaluators and the GPU reimplementation

The exact values in Sections 4.3 and 5 come from evaluators written separately for each design family. We compared each evaluator with an independent calculation of the same quantity. Where a design's authors released a sampler or simulation program, we also compared the evaluator with its output. Two exact calculations had to agree as rationals, or the enclosure from one had to contain the value from the other. An exact and a simulated value had to agree within the simulation's Monte Carlo error. Table C1 lists these comparisons; the GPU reimplementation of BLAST [15] was compared with the authors' original code in the same way, as described at the end of this appendix.

Form of the certificates. For the designs evaluated by enumeration (Simon, BOP2, Local-MEM except as noted below for its six-basket table, binary-endpoint BayesCTDesign, and AS-

SISTant Design C) each reported value is a pair $[\mathrm{OC}_{\mathrm{lower}}, \mathrm{OC}_{\mathrm{upper}}]$, with $\mathrm{OC}_{\mathrm{lower}}$ the exact probability of the outcomes certified to lie on the rejection side and $\mathrm{OC}_{\mathrm{upper}} - \mathrm{OC}_{\mathrm{lower}}$ the exact probability of the outcomes left unresolved (outcomes on which the released code errors or separates, or whose decision statistic could not be separated from its cutoff); an entry is decided only if $\mathrm{OC}_{\mathrm{upper}} < b$ or $\mathrm{OC}_{\mathrm{lower}} > b$. Outcome weights are exact rationals, and per-outcome comparisons are exact integer or rational comparisons, or ball-arithmetic comparisons that must exclude the cutoff. The largest unresolved mass among these entries is 5.7×10^{-54} (separation outcomes of one BayesCTDesign entry). For the hierarchical designs (CBHM with its BHM and BHM-HN comparators, and BaCIS) the interval has the same form and the same exact weights; the per-state posterior tail comes from quadrature checked by an independent route, the two agreeing near the cutoffs to a few parts in 10^6, and a state whose computed tail lies within 10^{-4} of its cutoff is carried as unresolved. Multi-stage designs propagate conservatively: a state whose interim decision is unresolved is removed from the recursion and its full weight is added to the width of every arm. Of the 88 enumeration-register entries certified in this form, 79 are decided by the interval and 9, the recalibrated type I error entries, by the calibration convention itself, under which the achieved size is a rational number that cannot equal the target; none straddles its threshold.

C.1 Validation of each evaluator

The dynamic program for CBHM [17], which also evaluates that paper's BHM comparator, agrees with a brute-force exact evaluation of the same decision rule and with an independently written three-dimensional integration. The estimates from two 200,000-trial Monte Carlo runs of the same decision code at different seeds both fall inside the program's enclosures. The program's values also agree with the design authors' own sampler, and its decisions are the same under both of the paper's two published constant conventions. These two conventions are distinct from the two readings of CBHM's calibration rule discussed in Section 5.3. For the Simon-design columns of Local-MEM's two-stage table and for the other directly enumerated entries, the exact rationals were each obtained by two independent methods that agree exactly. As a positive control, we introduced an error deliberately into one method and confirmed that the two methods then disagreed.

The Local-MEM enumeration [18], run over the full 20^4 outcome space for each of the 15 basket partitions, agrees with the paper's released code to 1.4×10^{-14} in the posterior quantities. The enumeration itself is exact, and the operating characteristics are summed from it in exact integer arithmetic; summation in double precision agrees with the exact sums to 4.1×10^{-16}. For the four-basket tables and the two-stage table every per-outcome decision is also certified with the partition posterior in exact rationals and the incomplete-beta tail as a 200-bit ball: no decision differs from the double-precision one and no ball meets its cutoff (smallest distance 1.0×10^{-6}). For the six-basket table the same certification covers every outcome whose tail lies within 10^{-4} of the cutoff (505,891 outcomes, with no change of decision and none unresolved), and the remaining decisions rest on double-precision tails farther than 10^{-4} from the cutoff. A value is placed on one side of its threshold only when it differs from the threshold by more than 10^{-6}, as every Local-MEM value did. The sentence of the Local-MEM paper that specifies the stage-two rule of its two-stage table can be read in two ways. We ran the enumeration under both readings, and each of the 14 entries concerned falls on the same side of its threshold either way.

The BOP2 evaluator [14] computes the recursion of equation (8) in exact rational arithmetic.

The boundary comparisons are made in ball arithmetic for the binary-endpoint variant and between exact integers for the joint-endpoint variant. The evaluator reproduces all 300 points of a grid of stopping boundaries computed with the design authors' archived code. It also agrees with that code's seeded run of 10,000 trials to within 1.52 MCSE (Table C1). A second evaluator, written independently of the first and using only the rules printed in the paper, returns the same values for the 10 BOP2 entries that have an exact reference value.

For BaCIS [16] and for CBHM's BHM-HN comparator, two independently configured quadrature schemes agree to about 10^{-6} or better. Each quadrature evaluator matches the design authors' own sampler to about one percent, with every difference within three standard errors of the sampler estimate. For the BaCIS entries whose enclosure excludes the threshold, the enclosure has zero width: the computed value is an exact rational.

The exact distribution of the rank statistic of ASSISTant's Design C [21] agrees with brute-force enumeration of all 2^N outcomes at small N and with an independent implementation in R to twelve digits. Perturbing each threshold by ± 1 unit in the last place changes no decision. The evaluations for BaCIS, the BHM and BHM-HN comparators, the recalibrated CBHM configurations, and ASSISTant belong to the second part of the recomputation (Appendix E), as do the re-simulation of ASSISTant's Designs A and B and the first of the two extended runs in Appendix D. That part used about 180 core-hours in all, 55 of them for the first extended run.

C.2 Two checks of printed BOP2 tables

The operating characteristics of a BOP2 design follow from its integer stopping boundaries by the recursion of equation (8) [14]. Hence any design whose boundaries are printed in the BOP2 paper can be evaluated exactly and its operating characteristics compared with those reported for it. The design defined by the boundaries printed in that paper's appendix Table DII has, at the printed null, an exact type I error in $[0.078373, 0.078374]$, an early-stop probability in $[0.90812, 0.90813]$, and an expected sample size in $[23.525, 23.526]$. The scenario-1 row of the printed Table DI gives these quantities as 9.7, 85.4, and 27.1 (probabilities in percent), which differ from the exact values. However, when the recursion is applied in the same way to the stopping boundaries of the efficacy–toxicity example in the paper's main text, it reproduces the operating characteristics reported for that example within 1.3 MCSE. The paper does not say which of the two printed tables corresponds to the simulation actually run, so the comparison with the DI row carries that caveat, and the census cell in that row is undetermined in either case. That is, whichever design was simulated, the Table DI entry classified in Table 2 lies within 2 MCSE of its threshold at the reported n, because the classification uses only the reported estimate, the threshold, and n.

In addition, the operating-characteristic tables of the BOP2 paper can be compared with a bound that every BOP2 design satisfies. A BOP2 trial that accepts the treatment has not stopped early, and the probability of acceptance and the probability of early termination (PET) reported for one scenario therefore satisfy accept + PET $\leq 100\%$. Of the 128 printed (accept, PET) pairs in the paper's eight operating-characteristic tables, two violate the bound. Both are rows of Table II, and neither is among the entries classified in Table 2, because no threshold applies to them. With each value printed to one decimal place, the largest sum consistent with the bound is 100.1, and both rows exceed it (Table C2).

compared with	agreement
CBHM and its BHM comparator (exact-rational dynamic program)	
brute-force exact evaluation of the same decision rule	agree
two 200,000-trial Monte Carlo runs of the same decision code, different seeds	both estimates inside the enclosures
design authors' sampler	96 of 96 comparison values agree
independently written three-dimensional integration	to 10^{-16}
the paper's two published constant conventions	same decisions under both
Simon-design columns of Local-MEM's two-stage table; outcome enumerations (exact rationals)	
a second, independent method	equal as rationals; a deliberately introduced error was detected
Local-MEM enumeration (20^4 outcomes for each of 15 partitions)	
the paper's released code	posterior quantities to 1.4×10^{-14}
the two readings of the stage-two rule	14 of 14 entries on the same side of the threshold
BOP2 (exact-rational recursion; binary and joint-endpoint variants)	
grid of stopping boundaries from the authors' archived code	300 of 300 points equal
the archived code's seeded run of 10,000 trials	within 1.52 MCSE
stopping boundaries printed in the paper's Example 1	equal
second evaluator written from the printed rules alone	same 10 reference values
BaCIS (posterior quadrature checked by a second route; singleton tails exact)	
design authors' sampler	84 of 84 finite-error checks within three standard errors
second, independently configured quadrature	worst disagreement 4.3×10^{-6}
BHM-HN comparator (posterior quadrature checked by a second route)	
second quadrature scheme	worst disagreement 3.6×10^{-10}
design authors' sampler, 96 tail values	three beyond two standard errors, none beyond three; largest 2.714
ASSISTant Design C (exact distribution of the rank statistic)	
brute force over all 2^N outcomes at small N	agree
independent R implementation	to twelve digits
thresholds perturbed by ± 1 ulp	no decision changes

Table C1: Each exact evaluator agrees with an independent exact calculation and, where the design's authors released a sampler, with that sampler. Rows are grouped by evaluator (italic); columns give the calculation it was compared with and the agreement found. "Equal" means identical integers or rationals; "agree" means the exact values coincide or one lies inside the other's enclosure. Standard errors are those of the sampler or simulation estimate; MCSE, Monte Carlo standard error of a simulated proportion; ulp, unit in the last place (double precision).

Table II row	printed accept + PET (%)	exact accept + PET (%)
scenario 1, ORR 0.30	$55.2 + 46.2 = 101.4$	$52.75 + 45.59 = 98.34$
scenario 2, ORR 0.60	$98.9 + 1.8 = 100.7$	design unrecovered

Table C2: Two printed rows of BOP2's Table II [14] give an acceptance probability and a probability of early termination (PET) that sum to more than 100%, although a trial that accepts cannot have stopped early; the exact values at the first row's printed stopping boundaries satisfy the bound. ORR, objective response rate; "accept", the reported probability that the trial declares the treatment acceptable at the final analysis. For the second row no exact value exists, because the scenario-2 design could not be recovered from the archived code; only the printed sum is shown.

C.3 The GPU reimplementation of BLAST

The cost in Table 8 of rerunning BLAST was estimated from two measurements made with the authors' released code [15]. The first was a timing of the sampler alone, which gives a lower bound on the cost per simulated trial; the second was a direct run of one complete trial. The direct run, extrapolated linearly in the paper's chain length, gives 18.3 CPU-hours of user time per trial and 87.5 CPU-hours when system time is included. Only the user-time figure enters the cost estimate, because the large system-time component reflects the heavy load on the shared 96-core machine during the run (load average several hundred).

Had the 30 core-hours allowed for BLAST in the feasibility study been spent instead on a partial replication of one scenario, they would have covered at most 50–90 simulated trials at the lower (sampler-only) cost per trial. An estimate from that many trials has a Monte Carlo standard error of at least 3.2 percentage points. This standard error exceeds the ± 1.9-point window (two MCSE at the reported n) inside which BLAST's unresolved entries lie, and such a run could not have resolved any of them.

The GPU reimplementation of BLAST includes the authors' own sampler component, translated from their released source, and before timing the reimplementation we compared it with the original code. We ran the reimplementation at full chain length with 24 seeds per entry, and the authors' unmodified code at 100,000 draws per entry. All 152 scalar quantities on which agreement was judged fell within tolerance, as did the 28 further quantities derived from them. One entry (the final-look fit of the two-cluster model) was compared under a different rule. A study with sixty seeds had shown that at this entry the authors' own code ends with either of two labelings of the latent clusters, whose decision rates differ, as described in Section 6. The rule for that entry therefore requires the two implementations to end with each labeling at indistinguishable rates and to agree within tolerance both within each labeling and with the two labelings pooled, and all three conditions held. At the paper's full chain length the reimplementation simulates 0.104 trials per second with 256 trials run in parallel (the batch width). That rate corresponds to the 2.668 GPU-hours per 1,000-trial scenario row used in Section 6.

D Entries unresolved at two million repetitions and the threshold-valued entries

Fourteen entries were still within 2 MCSE of their thresholds when the reference simulations stopped; two of them were later decided by runs past the maximum of two million repetitions,

and a third, a reprint of one of those two, takes its original's outcome. Table D1 gives the fourteen estimates and their distances from the thresholds. The entries reported exactly at their threshold are listed for each design at the end of this appendix.

D.1 The eleven unresolved entries and the two extended runs

Of the eleven unresolved entries, six are type I error entries of BayesCTDesign [19] under its piecewise-exponential model, estimated at 0.0494 to 0.0503 across that paper's sample sizes (Section 5.2), all still within 2 MCSE of 0.05. A seventh, a power entry of the same model reported as 0.796, is estimated at 0.79833 when re-simulated at the number of repetitions its own margin requires. The eighth is the Weibull power entry `T05_n175_a0`, estimated at 0.800021 against 0.80 at the full two million repetitions. At that count the estimate is 0.07 standard errors from the target (95% interval 0.79947 to 0.80058) and the side is unresolved; the count that would resolve it by simulation depends on the unknown true margin, about 1.3×10^8 for an 80% chance if the true margin were 10^{-4}, and the observed margin of 2.1×10^{-5} gives no planning count.

Two of the eleven are ASSISTant Design B entries [21], whose estimates at the counts their margins require were 0.0502 and 0.0513 against 0.05, or 0.24 and 1.3 MCSE above the threshold. Longer runs of 13.0 million and 440,000 repetitions, whose costs were estimated beforehand at 28 and 1 core-hours, gave 0.04990 and 0.04996, or 1.7 and 0.13 MCSE below the threshold. All four estimates are thus within 2 MCSE of the threshold, and both entries remain unresolved. The last, from BayesCTDesign, is a reprint of `T05_n175_a0` and takes its outcome. A second reprint, of `T04_n175_a0`, takes the outcome of its original's extended run, described below, and is counted among the decided entries.

Both entries decided by runs past the maximum of two million repetitions are Weibull type I error entries of BayesCTDesign. To measure the cost of such runs we first continued `T04_n175_a0` from two million to 7.93 million repetitions, which took 55 core-hours. The resulting estimate, 0.050403, lies above 0.05 and confirms the reported side. The second entry, `T04_n150_a0`, had given 0.05027 against 0.05 at 1.15 million repetitions, and the count this margin requires exceeded two million, with an estimated cost of about 73 core-hours at the rate just measured. Its longer run, 10.5 million repetitions in 22.9 core-hours, gave 0.0505, 7.4 MCSE above 0.05, confirming the reported side, while the reported value 0.058 is not reproduced (0.0505 ± 0.0001). The earlier run is reported beside this one, and the two are not pooled. For the piecewise-exponential entries we made no further runs: the counts required by the margins of their new estimates would cost thousands of core-hours each, and no exact computation is feasible for that model.

D.2 The threshold-valued entries by design

Of the 77 entries reported at exactly their threshold (Table 4), for which the reported margin sets no re-simulation size, eight were placed above or below the threshold by exact computation and one by a reference simulation, as described in Section 5.3. The remaining 68 include 53 ScuRMST entries and seven comparator entries (four BOP2, one BaCIS, one CBHM, one Local-MEM), none of which can be recomputed from published materials (Table 5), and seven BLAST entries, which can be evaluated only by simulation. The other one is the BOP2 entry of Section 4.3, whose exact reference value lies above the threshold conditional on a logged recovery of the design's code. The entry placed by simulation is an ASSISTant Design A entry reported as 5%, whose exact law is too large to evaluate, as shown in Section 5.4; a reference simulation of 200,000 trials of the released code gives 0.0477 with standard error 0.00048, or 4.8 standard

Entry and run	$\hat{p}$	b	Estimate	Repetitions	Distance from b	Outcome
Unresolved when simulation stopped (eleven entries, and one reprint later decided with its original)						
BayesCTDesign, piecewise-exponential, six type I error entries	0.048–0.068	0.05	0.0494–0.0503	margin-set	within 2 MCSE	no side
BayesCTDesign, piecewise-exponential, power entry	0.796	0.80	0.79833	margin-set	within 2 MCSE	no side
BayesCTDesign, Weibull, power entry `T05_n175_a0`	0.798	0.80	0.800021	two million	0.07 above	no side
ASSISTant Design B, first entry, first run	–	0.05	0.0502	45,697	0.24 above	
longer run, est. 28 core-hours			0.04990	13.0 million	1.7 below	no side
ASSISTant Design B, second entry, first run	–	0.05	0.0513	45,697	1.3 above	
longer run, est. 1 core-hour			0.04996	440,000	0.13 below	no side
Reprint of `T05_n175_a0`						as its original
Reprint of `T04_n175_a0`						as its original: reported side confirmed by the run past two million (lower block)
Decided by runs past two million repetitions (two entries)						
`T04_n175_a0`, BayesCTDesign, Weibull, type I error; run past two million, 55 core-hours	–	0.05	0.050403	7.93 million	above	reported side confirmed
`T04_n150_a0`, BayesCTDesign, Weibull, type I error; first run	0.058	0.05	0.05027	1.15 million	within 2 MCSE	margin-set count beyond two million
longer run, 22.9 core-hours (est. about 73 beforehand)			0.0505 ± 0.0001	10.5 million	7.4 above	reported side confirmed; 0.058 not reproduced

Table D1: Eleven entries remain within 2 MCSE of their thresholds at the largest repetition counts run for them; the two entries decided by longer runs both confirm their reported side, and a reprint of one of them takes that outcome. $\hat{p}$ is the value reported in the design paper and b its threshold (nominal type I error rate or target power). Each estimate was obtained by simulation at the repetition count shown, because none of these entries can be computed exactly, and MCSE is its Monte Carlo standard error, $\sqrt{\hat{p}(1-\hat{p})/n}$ evaluated at that estimate and count n; the distance column gives the separation of the estimate from b in units of that MCSE. "Margin-set" denotes the number of repetitions that the entry's own reported margin requires by equation (5). Where an entry was run twice, each run has its own line. Core-hour figures marked "est." were computed before the run from a measured rate. A dash marks a reported value not quoted here.

errors below the threshold.

E The reversal forecast in full

The expected numbers of reversals in Section 5.3 are computed from equation (10), an illustrative model rather than a calibrated one. In that equation q_i is the posterior probability, under a uniform prior on the true value and a binomial likelihood for the reported count $X_i = n_i\hat{p}_i$, that the true value of entry i lies on the other side of its threshold b_i from $\hat{p}_i$; that is, q_i is the tail beyond b_i of $\mathrm{Beta}(X_i + 1, n_i - X_i + 1)$. With the Jeffreys prior $\mathrm{Beta}(X_i + \frac{1}{2}, n_i - X_i + \frac{1}{2})$ every expected count in Table E1 changes by less than 3% (about 1.3 reversals), and no interval endpoint moves by more than two. An entry is forecastable when $\hat{p}_i \neq b_i$. For an entry reported exactly at its threshold no q_i is defined, because the reported value lies on neither side of b_i. When exact computation later places such an entry above or below its threshold, the result is not counted as a reversal. We carried out the recomputation in two parts and wrote the exact evaluators for BaCIS and for the re-derived CBHM cutoffs only in the second. Before each part we computed the expected number of reversals among the entries it covered.

E.1 Pools and expected counts

Of the 351 entries unresolved at their reported n, 281 are forecastable and 70 are reported exactly at their threshold. These 70 are the 77 threshold-valued entries of Table 4 less BLAST's seven, which are not among the 351 because BLAST was not classified in full. Over the 281 the expected number of reversals is 78.3, with 90% interval $[67, 90]$ (Table E1). For the 128 of these that could be computed without writing a new evaluator, the expectation is 29.86, with interval $[22, 37]$. Of the 78.3 expected reversals, 41.25 come from the 126 forecastable entries that cannot be computed from published materials, a component that no calculation can test.

Before the first part of the recomputation we evaluated the forecast over the 219 forecastable entries then thought computable, obtaining an expected count of 58.6 with 90% interval $[49, 69]$. This set did not include the 126 entries already known not to be computable from published materials, and three groups of entries within it were not recomputed in the first part. First, BLAST's forecastable entries [15] are outside the 351, and, because of the measured cost of the released code (Section 6), no reference simulation was run for them. Second, the entries that require the exact evaluators for BaCIS [16] or for the re-derived CBHM cutoffs [17] could not be recomputed until those evaluators were written, in the second part. Third, the comparator entries of several designs were found, on examination of the published materials, not to be computable.

By the end of the first part, 94 of the 219 entries had been recomputed, of which 84 were decided and the others remained too close to call. Table E1 gives the expected count under the same q_i over the 94 and over the 84, together with the 16 reversals observed, which lie inside both 90% intervals. Whether an entry could be computed was settled by the availability of code, cutoffs, and evaluators, without reference to where its recomputed value would fall. Within the computed entries, however, the decided subset is not free of selection: an entry recomputed by simulation is decided only if its recomputed value clears the noise at the maximum count, so entries whose true value lies very near the threshold, for which a reversal is close to an even chance, remain unresolved and drop out, whereas an entry recomputed exactly is always decided. The comparison of observed with expected counts is therefore a consistency check on the q_i and not a validation of them; a validation would require replaying calibration, simulation,

set of entries	entries	expected	90% interval	observed
Forecast written before the first part of the recomputation				
forecastable entries among the 351 unresolved entries	281	78.3	[67, 90]	–
of the 281, computable with evaluators then available	128	29.86	[22, 37]	–
of the 281, not computable from published materials	126	41.25	–	–
entries thought computable before the first part	219	58.6	[49, 69]	–
of the 219, decided or too close to call in that part	94	21.3	[15, 28]	16
of the 94, decided	84	18.8	[13, 25]	16
Variants of the tail probability on the 219-entry set (exact beta tail, as above, 58.6)				
normal approximation to the tail	219	57.22	–	–
beta tail, margins widened by one last-digit unit	219	57.32	–	–
Forecast written before the second part				
forecastable entries the second part was to address	32	8.52	[5, 13]	–
of the 32, decided in the second part	19	5.9	[3, 9]	8

Table E1: Wherever a set was recomputed in full, the observed number of reversals lies inside the forecast's 90% interval; the 281- and 219-entry sets have no observed count, because many of their entries were never recomputed. Each row is a set of forecastable entries (reported value $\hat{p}_i \neq$ threshold b_i) over which the model of equation (10) was evaluated; "expected" is the sum of their q_i, the interval is that of the number of reversals under the model, and "observed" is given only where every entry of the set was recomputed. A dash marks a quantity undefined or not evaluated.

and stopping on designs whose operating characteristics are known exactly, which we have not done. The intervals in Table E1 further assume independent entries. Entries estimated from one batch of simulated trials, reprints, and entries calibrated on the same simulations are dependent; $E[R]$ is unaffected, but if every entry of a simulation scenario is treated as perfectly dependent (one uniform variate deciding all of a scenario's indicators, the widest case for the given q_i) the 90% intervals of Table E1 widen by factors of 1.5 to 2.6, and the printed intervals should be read as lower bounds on the model's spread. Every observed count also lies inside the widened intervals, which is a weaker statement.

E.2 Model variants and exclusions

On the 219-entry set the expected count depends little on the form of q_i. The exact beta tail gives 58.6 and a normal approximation to it gives 57.22. Widening each margin $|\hat{p}_i - b_i|$ by one unit in the last printed decimal place, to allow for rounding of the reported values, gives 57.32 with the beta tail.

For the second forecast we applied the same q_i to the 32 forecastable entries that the second part was to address, with the expected counts shown in Table E1. Of these 32, 19 were decided, and the 8 reversals observed among them lie inside the model's 90% interval for those 19. Three groups of entries were excluded from the second forecast because their outcome was already known when that forecast was specified: four BHM comparator entries carried over from the first part and already computed, three BOP2 entries [14] with existing exact reference values, and fourteen reprinted entries, which take the outcome of the entry they reprint.

References

[1] U.S. Food and Drug Administration. Adaptive designs for clinical trials of drugs and biologics: guidance for industry. Technical report, U.S. Department of Health and Human Services, Food and Drug Administration, November 2019. Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research.

[2] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH E20: adaptive designs for clinical trials. Technical report, ICH, 2025. Draft guideline, Step 2 public-consultation text, endorsed 25 June 2025.

[3] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.

[4] Cristiana Mayer, Inna Perevozskaya, Sergei Leonov, Vladimir Dragalin, Yili Pritchett, Alun Bedding, Alan Hartford, Parvin Fardipour, and Greg Cicconetti. Simulation practices for adaptive trial designs in drug and device development. *Statistics in Biopharmaceutical Research*, 11(4):325–335, 2019. doi: 10.1080/19466315.2018.1560359.

[5] Axel Gandy. Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1511, 2009. doi: 10.1198/jasa.2009.tm08368.

[6] Lasse Fischer and Aaditya Ramdas. Sequential Monte Carlo testing by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 87(4):1200–1220, 2025. doi: 10.1093/jrsssb/qkaf014.

[7] Elizabeth Koehler, Elizabeth Brown, and Sebastien J.-P. A. Haneuse. On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162, 2009. doi: 10.1198/tast.2009.0030.

[8] Ian R. White. simsum: analyses of simulation studies including Monte Carlo error. *The Stata Journal*, 10(3):369–385, 2010. doi: 10.1177/1536867X1001000305.

[9] Anne-Laure Boulesteix, Sabine Lauer, and Manuel J. A. Eugster. A plea for neutral comparison studies in computational sciences. *PLoS ONE*, 8(4):e61562, 2013. doi: 10.1371/journal.pone.0061562.

[10] K. Luijken, A. Lohmann, U. Alter, J. Claramunt Gonzalez, F. J. Clouth, J. L. Fossum, L. Hesen, A. H. J. Huizing, J. Ketelaar, A. K. Montoya, L. Nab, R. C. C. Nijman, B. B. L. Penning de Vries, T. D. Tibbe, Y. A. Wang, and R. H. H. Groenwold. Replicability of simulation studies for the investigation of statistical methods: the RepliSims project. *Royal Society Open Science*, 11(1):231003, 2024. doi: 10.1098/rsos.231003.

[11] Richard Simon. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1):1–10, 1989. doi: 10.1016/0197-2456(89)90015-9.

[12] Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, 2000.

[13] Kevin Kunzmann, Maximilian Pilz, Carolin Herrmann, Geraldine Rauch, and Meinhard Kieser. The adoptr package: adaptive optimal designs for clinical trials in R. *Journal of Statistical Software*, 98(9):1–21, 2021. doi: 10.18637/jss.v098.i09.

[14] Heng Zhou, J. Jack Lee, and Ying Yuan. BOP2: Bayesian optimal design for phase II clinical trials with simple and complex endpoints. *Statistics in Medicine*, 36(21):3302–3314, 2017. doi: 10.1002/sim.7338.

[15] Yiyi Chu and Ying Yuan. BLAST: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):723–740, 2018. doi: 10.1111/rssc.12255.

[16] Nan Chen and J. Jack Lee. Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes. *Biometrical Journal*, 61(5):1219–1231, 2019. doi: 10.1002/bimj.201700275.

[17] Yiyi Chu and Ying Yuan. A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials*, 15(2):149–158, 2018. doi: 10.1177/1740774518755122.

[18] Yilin Liu, Michael Kane, Denise Esserman, Ondrej Blaha, Daniel Zelterman, and Wei Wei. Bayesian local exchangeability design for phase II basket trials. *Statistics in Medicine*, 41 (22):4367–4384, 2022. doi: 10.1002/sim.9514.

[19] Barry S. Eggleston, Joseph G. Ibrahim, Becky McNeil, and Diane Catellier. BayesCT-Design: an R package for Bayesian trial design using historical control data. *Journal of Statistical Software*, 100(21):1–51, 2021. doi: 10.18637/jss.v100.i21.

[20] Jiashan He, Ruitao Lin, Yaxian Chen, and K. F. Lam. Two-stage double-arm trial optimal design of restricted mean survival time with sculpted critical region. *Statistics in Medicine*, 45(10-12):e70589, 2026. doi: 10.1002/sim.70589.

[21] Tze Leung Lai, Philip W. Lavori, and Olivia Yueh-Wen Liao. Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials*, 39(2):191–200, 2014. doi: 10.1016/j.cct.2014.09.001.

[22] European Medicines Agency, Committee for Medicinal Products for Human Use. Assessment report: Brinsupri (brensocatib). Technical Report EMA/345552/2025, European Medicines Agency, 2025. Public assessment report, Procedure No. EMEA/H/C/005820/0000, adopted 16 October 2025; initial marketing authorisation 18 November 2025.

[23] Anders Granholm, Aksel Karl Georg Jensen, Theis Lange, Anders Perner, Morten Hylander Møller, and Benjamin Skov Kaas-Hansen. Designing and evaluating Bayesian advanced adaptive randomised clinical trials: a practical guide. *Pharmaceutical Statistics*, 24(6): e70042, 2025. doi: 10.1002/pst.70042.

[24] Iosif Pinelis and Raymond Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063, 2016. doi: 10.1214/16-EJS1133.

[25] Andreas Anastasiou and Gesine Reinert. Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218, 2017. doi: 10.3150/15-BEJ741.

[26] R. N. Bhattacharya and J. K. Ghosh. On the validity of the formal Edgeworth expansion. *Annals of Statistics*, 6(2):434–451, 1978. doi: 10.1214/aos/1176344134.

[27] Martin Raič. A multivariate Berry–Esseen theorem with explicit constants. *Bernoulli*, 25 (4A):2824–2853, 2019. doi: 10.3150/18-BEJ1072.

[28] Marta Bofill Roig, Pavla Krotka, Carl-Fredrik Burman, Ekkehard Glimm, Stefan M. Gold, Katharina Hees, Peter Jacko, Franz Koenig, Dominic Magirr, Peter Mesenbrink, Kert Viele, and Martin Posch. On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC Medical Research Methodology*, 22:228, 2022. doi: 10.1186/s12874-022-01683-w.

[29] Benjamin R. Saville, Donald A. Berry, Nicholas S. Berry, Kert Viele, and Scott M. Berry. The Bayesian time machine: accounting for temporal drift in multi-arm platform trials. *Clinical Trials*, 19(5):490–501, 2022. doi: 10.1177/17407745221112013.