On marginal and conditional error in adaptive clinical trials

Zijun Gao
Department of Data Sciences and Operations, USC Marshall School of Business

Ting Ye
Department of Biostatistics, University of Washington

Qingyuan Zhao
Statistical Laboratory, University of Cambridge

23 July 2026

Abstract

Adaptive clinical trials can use interim data to select the treatment, patient population, or endpoint for final testing. Existing methods generally control family-wise type I error marginally, averaging over the selections that the trial might have made. However, in the drug development and regulatory process, the ultimate decision generally concerns the adaptively selected hypothesis, raising an important question: *conditional* on a hypothesis being selected, how often is it incorrectly rejected? Marginal error is a weighted average of these selection-conditional error rates, and therefore marginal error control does not imply selection-conditional error control. We formalize this distinction and study a two-stage adaptive trial in which either one or two of five treatments continue after interim selection. A commonly used closed combination test controls marginal family-wise error, but in our one-treatment illustration each zero-effect treatment has selection-conditional error of roughly 9.4% at nominal 5%. In contrast, a selective Gaussian test and a stage-2-only test control selection-conditional family-wise error in both illustrations. Perhaps surprisingly, the selective Gaussian test is also more powerful than the combination test in these examples. We call for more discussion in the clinical-trial community about the appropriate statistical error criterion for adaptive confirmatory claims.

Keywords: adaptive design; family-wise error rate; hypothesis selection; post-selection inference; seamless phase II/III trial; type I error

1 Introduction

Adaptive clinical trials use accumulating information to modify aspects of a study while it is in progress. Possible adaptations include early stopping, sample-size revision, treatment or dose selection, enrichment of a patient subpopulation, and changes to randomization probabilities. For example, a seamless phase II/III treatment-selection design joins a learning stage and a confirmatory stage, commonly discontinuing less promising treatments at an interim analysis (Bretz et al. 2006; Burdon et al. 2026); a platform trial uses a master protocol under which treatments can enter or leave (Park et al. 2019); and response-adaptive randomization changes allocation probabilities among ongoing treatments (Robertson et al. 2023). These adaptations target different

design components and can coexist. They may shorten development, direct resources toward more promising comparisons, and, in suitable settings, allow more participants to receive a better-performing treatment sooner. For these reasons, there is increasing interest in adaptive clinical trials (Bretz et al. 2009; FDA 2019).

In a conventional non-adaptive trial, the treatment, patient population, primary endpoint, and analysis are ordinarily fixed before the trial begins, using external scientific and clinical evidence. The tested hypothesis is therefore fixed relative to the randomization in that trial. Estimand guidance likewise asks trialists to prespecify the treatment effect of interest (ICH 2019). From the broader perspective of a development program, however, hypotheses are continuously refined using laboratory studies and earlier clinical phases, so there is no sharp boundary between learning and confirmation across the program. An adaptive clinical trial makes this continuation from learning to confirmation explicit in the design: interim data can determine which treatment–population hypothesis becomes the final primary question.

Several methods control family-wise error over the entire adaptive trial: closed combination tests for adaptive hypothesis selection (Bretz et al. 2006), group-sequential treatment-selection procedures (Stallard and Todd 2003), adaptive Dunnett tests (König et al. 2008), and generalized Dunnett tests (Magirr et al. 2012). These methods provide strong marginal control under their respective assumptions, and comparative work has generally assessed them through marginal operating characteristics rather than selection-conditional calibration (Friede and Stallard 2008). Confidence intervals after population selection can likewise target overall coverage averaged across selection decisions (Rosenblum 2013).

In reality, the primary conclusion in an adaptive trial often concerns the selected treatment or patient population. This raises a repeated-sampling question distinct from the usual trial-level family-wise error probability: among repetitions in which the same set of hypotheses is selected, how often is at least one true null hypothesis incorrectly rejected? We call this the *selection-conditional family-wise error rate*. Like the usual trial-level marginal family-wise error rate, it is also a frequentist probability, but one defined within a specified selection event. The corresponding literature on selection-conditional inference for adaptive clinical trials is notably smaller but growing. Exact normal-theory tests and intervals have been developed for drop-the-losers designs (Sampson and Sill 2005), with a subsequent extension to binary outcomes (Sill and Sampson 2009). Takahashi et al. (2022) proposed an exact test conditional on the selected treatment for a design in which treatment selection uses a short-term binary endpoint and confirmation uses a long-term binary endpoint. For binary short-term selection and survival confirmation, Sato et al. (2026) constructed a stage-1 randomization p-value conditional on a prespecified treatment-selection event and combined it with independent stage-2 evidence. The construction of Sato et al. can be viewed as a design-specific implementation of the selective randomization principle formalized more generally by Freidling et al. (2026). On the estimation side, Li et al. (2026) construct confidence intervals with exact coverage conditional on adaptive subpopulation selection. These works use selection conditioning centrally, but they focus on methodological construction—whether general or design-specific—rather than on comparing marginal and selection-conditional FWER as general error criteria.

In this short note, we aim to expose the differences between marginal and selection-conditional errors and stimulate discussion about them in adaptive clinical trials. Section 2 defines the two criteria and distinguishes the selection-conditional error considered here from the classical conditional-error function used in adaptive clinical-trial methodology. Section 3 uses two-stage Gaussian examples that continue one or two treatments to make the distinction concrete. Section 4 discusses further considerations from design, analysis, and regulatory perspectives.

2 Marginal and selection-conditional error rates

Consider a two-stage trial with first-stage data X_1 and subsequent data X_2. The statistical model is $\{P_\theta : \theta \in \Theta\}$. There are K candidate null hypotheses $H_1, \ldots, H_K$, indexed by $[K] = \{1, \ldots, K\}$. For each θ, let $H_j(\theta) = 0$ if H_j is true under P_θ and $H_j(\theta) = 1$ otherwise. A prespecified rule applied to X_1 returns a random set $S = S(X_1) \subseteq [K]$. After selection, $R_j = R_j(X_1, X_2)$ indicates rejection of H_j, with $R_j = 0$ when $j \notin S$. Thus $R_j\{1 - H_j(\theta)\} = 1$ indicates a false rejection of H_j.

The *marginal family-wise error rate (FWER)* is

$$\mathrm{FWER}(\theta) = P_\theta\left[\sum_{j \in S} R_j\{1 - H_j(\theta)\} \geq 1\right], \tag{1}$$

Strong marginal FWER control requires $\mathrm{FWER}(\theta) \leq \alpha$ for every $\theta \in \Theta$.

For a fixed θ, A_θ collects all subsets of $[K]$ that have nonzero probability of being selected under P_θ. For each $s \in A_\theta$, define the *selection-conditional family-wise error rate* as

$$\mathrm{FWER}(\theta \mid s) = P_\theta\left[\sum_{j \in s} R_j\{1 - H_j(\theta)\} \geq 1 \mid S = s\right]. \tag{2}$$

Strong selection-conditional FWER control requires $\mathrm{FWER}(\theta \mid s) \leq \alpha$ for every $\theta \in \Theta$ and every $s \in A_\theta$. For a singleton selection $S = \{j\}$, this becomes the selected hypothesis's *conditional type I error*, $P_\theta(R_j = 1 \mid S = \{j\})$, whenever H_j is true.

The law of total probability gives

$$\mathrm{FWER}(\theta) = \sum_{s \in A_\theta} P_\theta(S = s)\,\mathrm{FWER}(\theta \mid s) \leq \max_{s \in A_\theta} \mathrm{FWER}(\theta \mid s). \tag{3}$$

The weights in this average are non-negative and sum to one. Hence selection-conditional control immediately implies marginal control. The converse fails because an average can be controlled while one component is large. Thus, even if the marginal $\mathrm{FWER}(\theta) \leq \alpha$, a selected set s can have conditional $\mathrm{FWER}(\theta \mid s)$ much greater than α. (For example, conditional error rates 0.04 and 0.50 in selection strata with probabilities 0.99 and 0.01 yield a marginal error of 0.0446.)

Marginal FWER is therefore the ex ante probability of at least one false rejection in a trial, averaged over adaptation paths, whereas selection-conditional FWER requires calibration within each specified path. Both are repeated-sampling probabilities under a fixed θ; neither is $P_\theta(H_j \text{ true} \mid S = s, R_j = 1)$ or another posterior probability that an observed decision is wrong. Preference between them is a scientific or regulatory choice about calibration, not a logical consequence of reporting a selected claim.

Our terminology should not be confused with the *conditional-error function* in adaptive-design theory. That function is the remaining null rejection probability of an originally planned design after conditioning on the realized interim data. A modification that spends no more than this data-dependent remaining error preserves the original *unconditional* type I error after averaging over the interim data (Friede et al. 2012; Müller and Schäfer 2001). It does not generally control error conditional on which hypothesis was selected.

3 Two-stage Gaussian treatment-selection examples

We now give two concrete examples demonstrating the difference between marginal and selection-conditional errors. Example 1 selects the stage-1 winner and continues the trial with that treatment, whereas Example 2 selects and continues the two best-performing treatments.

3.1 Design and model

There are five experimental treatments, indexed by $j = 1, \ldots, 5$, and a control, indexed by $j = 0$, with $\theta \in \mathbb{R}^5$. Let Y_{ij} denote the sample mean for any group observed at stage i. Each observed group has $N = 100$ participants, and individual outcomes have known unit variance. The treatment effects are the same at both stages. At stage 1,

$$Y_{1j} \overset{\text{ind}}{\sim} N(\theta_j, 1/N), \qquad j = 0, 1, \ldots, 5, \qquad \theta_0 = 0.$$

Thus θ_j is a standardized mean effect. Let $J_{(1)}$ and $J_{(2)}$ index the largest and second-largest values among $Y_{11}, \ldots, Y_{15}$. We consider two examples. In Example 1, we select the best treatment in stage 1, $S_1 = \{J_{(1)}\}$. In Example 2, we select the unordered set of the two best treatments, $S_2 = \{J_{(1)}, J_{(2)}\}$.

At stage 2, another $N = 100$ participants are enrolled in each selected-treatment group and in a new control group.

For each $m \in \{1, 2\}$ and every realizable set s, conditional on $S_m = s$, the variables $(Y_{2j} : j \in s)$ and Y_{20} are mutually independent, with

$$Y_{2j} \sim N(\theta_j, 1/N), \quad j \in s, \qquad Y_{20} \sim N(0, 1/N),$$

and are generated independently of the complete stage-1 data.

The two stages therefore have distinct controls and share no participants.

For every observed treatment–control comparison, define

$$Z_{ij} = \frac{Y_{ij} - Y_{i0}}{\sqrt{2/N}}, \qquad i = 1, 2.$$

For $j \neq k$ that are both observed at stage i, $\operatorname{corr}(Z_{ij}, Z_{ik}) = 1/2$ because both comparisons use the same control data. The stage-1 control cancels from treatment rankings, so ranking the treatment means is equivalent to ranking Z_{1j}. We test $H_j : \theta_j \leq 0$ against $\theta_j > 0$ at the illustrative one-sided level $\alpha = 0.05$. Let Φ denote the standard Gaussian distribution function.

3.2 Three testing procedures

We compare three procedures in both examples.

Closed combination test. We use a Gaussian specialization of the closed combination framework of Bretz et al. (2006), with exact Dunnett local tests, equal inverse-normal stage weights, and full closure over the original five hypotheses. For each nonempty $I \subseteq [K]$, let $H_I = \cap_{j \in I} H_j$. Let $F_{d,1/2}$ denote the distribution function of the maximum of d standard Gaussian variables with pairwise

correlation 1/2; in particular, $F_{1,1/2} = \Phi$. Following Dunnett (1955), the stage-1 local Dunnett p-value for H_I is

$$p_{1I} = 1 - F_{|I|,1/2}\left(\max_{j\in I} Z_{1j}\right).$$

If $I \cap S_m$ is nonempty, the fresh stage-2 local p-value is

$$p_{2I} = 1 - F_{|I\cap S_m|,1/2}\left(\max_{j\in I\cap S_m} Z_{2j}\right).$$

The single-arm formula reduces to $1 - \Phi(Z_{2j})$. If $I \cap S_m$ is empty, set $p_{2I} = 1$, so its stage-2 local test never rejects. We combine the two stagewise p-values by the equal-weight inverse-normal, or Stouffer, combination (Bauer and Köhne 1994; Stouffer et al. 1949):

$$p_I = 1 - \Phi\left[2^{-1/2}\Phi^{-1}(1 - p_{1I}) + 2^{-1/2}\Phi^{-1}(1 - p_{2I})\right].$$

A selected H_j is rejected only if $p_I \le \alpha$ for every I containing j, as required by the closure principle (Marcus et al. 1976). Under H_I, p_{1I} is super-uniform and p_{2I} is super-uniform conditional on X_1. Using the law of total probability, it is not hard to show that p_I is also super-uniform. Thus, closure gives strong marginal FWER control, but not generally selection-conditional control.

Selective Gaussian test. We adapt normal-theory ideas for drop-the-losers designs (Sampson and Sill 2005) using polyhedral Gaussian conditioning (Fithian et al. 2014; Lee et al. 2016); related selective inference for adaptive subpopulation selection is developed by Li et al. (2026). Fix a realizable set s and $j \in s$. Define the pooled statistic

$$W_j = \frac{Z_{1j} + Z_{2j}}{\sqrt{2}} \sim N(\sqrt{N}\,\theta_j, 1).$$

Let $a_{jk} = 1/\sqrt{2}$ if $k = j$ and $a_{jk} = 1/(2\sqrt{2})$ otherwise, and set $Q_{jk} = Z_{1k} - a_{jk}W_j$. These coefficients are chosen so that

$$\mathrm{Cov}(Q_{jk}, W_j) = \mathrm{Cov}(Z_{1k}, W_j) - a_{jk}\,\mathrm{Var}(W_j) = 0 \qquad (k = 1, 2, \ldots, K).$$

Thus each Q_{jk} has zero correlation with W_j. Since W_j and $(Q_{j1}, \ldots, Q_{jK})$ are jointly Gaussian, W_j is independent of this vector. Expressing the selection inequalities in terms of W_j and the Q_{jk}, it is not difficult to show that

$$W_j \mid (Q_{j1}, \ldots, Q_{jK}, S_m = s) \sim \mathrm{TN}_{[L_j(s),\infty)}(\sqrt{N}\,\theta_j, 1), \qquad L_j(s) = 2\sqrt{2}\max_{k\notin s}(Q_{jk} - Q_{jj}),$$

where $\mathrm{TN}_{[L,\infty)}(\mu, 1)$ denotes an $N(\mu, 1)$ variable truncated below at L. Under the boundary null $\theta_j = 0$, the selective upper-tail p-value is therefore

$$p_{\mathrm{sel}j} = \frac{1 - \Phi(W_j)}{1 - \Phi\{L_j(s)\}}. \tag{4}$$

This p-value is uniform at $\theta_j = 0$ and super-uniform for $\theta_j < 0$, conditional on selection. In Example 1, we reject when $p_{\mathrm{sel}j} \le \alpha$. In Example 2, we apply Holm's procedure to the two selective p-values, giving finite-sample selection-conditional strong FWER control, although the attained FWER can be conservative.

Stage-2-only test. Because stage 2 is independent of selection, Example 1 rejects H_j when $Z_{2j} > \Phi^{-1}(0.95) \approx 1.64$. In Example 2, suppose $S_2 = \{j, k\}$. The closed step-down Dunnett procedure rejects H_j exactly when

$$Z_{2j} > \Phi^{-1}(0.95) \approx 1.64 \quad \text{and} \quad \max(Z_{2j}, Z_{2k}) > F^{-1}_{2,1/2}(0.95) \approx 1.92,$$

and analogously for H_k (Dunnett and Tamhane 1991; Marcus et al. 1976). This gives selection-conditional strong FWER control.

3.3 Configurations and results

We consider $\theta = (\theta_1, 0, 0, 0, -0.02)$ for $\theta_1 = 0, 0.05, \ldots, 0.40$. Treatments 2, 3, and 4 remain at the boundary null throughout the grid, treatment 5 is the harmful treatment, and treatment 1 is at the boundary only when $\theta_1 = 0$. The main text reports $\theta_1 = 0.25$; results for the other values appear in Appendix A, together with a global-null calibration benchmark. Every probability in Tables 1 and 2 is estimated from two million repetitions. Under $\theta_1 = 0.25$, treatment 1 is selected with probability 0.892 in Example 1 and 0.973 in Example 2; treatment 5 is selected with probabilities 0.020 and 0.208, respectively.

Table 1: Example 1: rejection probabilities at one-sided $\alpha = 0.05$ for $\theta = (0.25, 0, 0, 0, -0.02)$. Marginal FWER is FWER($\theta$), and global-null marginal FWER is FWER$(0, \ldots, 0)$. The row labeled "Conditional FWER: boundary null" estimates FWER($\theta \mid s$) by pooling the three exchangeable selected sets $s = \{2\}, \{3\}, \{4\}$; each fixed-s quantity has the same theoretical value by symmetry. The row labeled "Conditional FWER: harmful treatment" reports FWER($\theta \mid \{5\}$). Selection-conditional power is $P_\theta(R_1 = 1 \mid S_1 = \{1\})$ and joint power is $P_\theta(S_1 = \{1\}, R_1 = 1)$. The pooled boundary-null stratum has probability 0.088.

Measure	Closed combination	Selective Gaussian	Stage-2-only
Marginal FWER	0.0098	0.0051	0.0051
Selection-conditional power	0.6757	0.7094	0.5484
Joint power	0.6025	0.6326	0.4890
Conditional FWER: boundary null	0.0942	0.0494	0.0492
Conditional FWER: harmful treatment	0.0726	0.0363	0.0373
Global-null marginal FWER	0.0499	0.0502	0.0499

Table 2: Example 2: FWER and power at one-sided $\alpha = 0.05$ for $\theta = (0.25, 0, 0, 0, -0.02)$. Marginal FWER is FWER($\theta$), and global-null marginal FWER is FWER$(0, \ldots, 0)$. The one- and two-boundary-null conditional-FWER rows estimate FWER($\theta \mid s$) by pooling, respectively, the selected sets $s = \{1, j\}$ for $j = 2, 3, 4$ and the three selected sets contained in $\{2, 3, 4\}$. The fixed-s quantities pooled within each row are equal by exchangeability. Selection-conditional power is $P_\theta(R_1 = 1 \mid 1 \in S_2)$ and joint power is $P_\theta(1 \in S_2, R_1 = 1)$. The one- and two-boundary-null strata have probabilities 0.776 and 0.016, respectively.

Measure	Closed combination	Selective Gaussian	Stage-2-only
Marginal FWER	0.0449	0.0439	0.0432
Selection-conditional power	0.5839	0.6629	0.4421
Joint power	0.5679	0.6448	0.4300
Conditional FWER: one boundary null	0.0469	0.0468	0.0454
Conditional FWER: two boundary nulls	0.0677	0.0471	0.0472
Global-null marginal FWER	0.0437	0.0470	0.0498

We would like to highlight two observations from these tables. First, the closed combination test controls marginal FWER, yet in Example 1 its conditional false-rejection probability for a selected boundary-null treatment is about 9.4%, above the nominal 5%. In Example 2 its conditional FWER reaches about 6.8% when two boundary-null treatments are selected. The selective and stage-2-only procedures control both marginal and selection-conditional FWER, as established for these procedures in Section 3.2.

Second, the selective Gaussian test has higher selection-conditional power than the closed combination test here while providing the stronger conditional guarantee. The difference is about three percentage points in Example 1 and eight percentage points in Example 2; joint power shows the same pattern. The selective method gains by using the fully specified Gaussian top-m selection rule, whereas combination procedures permit broader adaptation, and alternative local tests or weights may behave differently. The stage-2-only procedure uses no stage-1 confirmation information and has lower power here. These examples do not establish general dominance, but they show that methods targeting selection-conditional error control can be more powerful than standard methods designed to control marginal FWER. Sato et al. (2026) also reported a related, design-specific power advantage for their selective randomization procedure over the combination-test procedure of Friede et al. (2011).

4 Discussion

To recap the main message of this short note, marginal and selection-conditional error rates correspond to different frequentist repetitions. Selection-conditional control implies marginal control, but not vice versa. The Gaussian examples show that the resulting difference can be substantial.

Which criterion should govern an adaptive confirmatory trial is a practical and regulatory question, not a mathematical question alone. In discussing group-sequential methods, FDA guidance says that multiplicity adjustments "ensure the overall Type I error probability of the trial is controlled at 2.5 percent" (FDA 2019, p. 7). The ICH E20 draft similarly describes the common approach as limiting "the probability of false positive efficacy conclusions within a trial" through control of the type I error probability (ICH 2025, p. 6). Neither document discusses selection-conditional

error in the sense defined here. The phrase "type I error control" is therefore incomplete unless the repetitions being averaged, and any conditioning event, are stated. The same distinction applies to estimation: estimand guidance does not determine whether an interval for a selected estimand should have marginal or selection-conditional coverage (ICH 2019; Li et al. 2026; Rosenblum 2013). In our view, selection-conditional error is often the more relevant risk for the ultimate decision because that decision generally concerns the adaptively selected hypothesis. But this is our position, not an established regulatory principle. Further debate among trialists, clinicians, sponsors, patients, and regulators is therefore needed.

Regardless of which error criterion is chosen, adaptive trials require careful prospective planning. A marginal procedure must specify the hypothesis family and provide enough detail about the adaptation and testing algorithm to justify its unconditional reference distribution. Some combination methods allow broad data-dependent treatment selection without requiring every possible decision to be encoded as a rigid deterministic map (Bretz et al. 2006; König et al. 2008). Selection-conditional analysis must additionally specify every relevant input and a reproducible deterministic or stochastic decision rule sufficiently well to reconstruct the selection event for counterfactual datasets. Merely recording the rationale for the realized decision does not define that event; any deliberative inputs or randomness must be incorporated into the selection model. This requirement is especially concrete for selective randomization methods: the prespecified selection rule must be reapplied under each alternative treatment assignment in the randomization distribution (Freidling et al. 2026; Sato et al. 2026). Design and analysis should therefore be developed together. Data splitting reserves independent information for confirmation. Data carving uses information not exhausted by selection, and randomized selection can soften the conditioning event and improve power or interval length (Fithian et al. 2014; Tian and Taylor 2018). Selective randomization inference similarly makes explicit which information passes between stages (Freidling et al. 2026). These approaches may allow scientific flexibility while retaining a defined reference distribution. Sample-size calculations must use the intended error criterion and final analysis, because power under marginal calibration need not transfer to a conditionally calibrated procedure, or vice versa.

Methodological development can proceed in parallel with clearer discussion of statistical risk. The Gaussian examples assume known variance, one or two selected treatments, a fixed stage-2 sample size, and a continuous endpoint. Real trials may stop early, revise sample size, select populations, or use different short- and long-term outcomes. Survival outcomes are particularly challenging because participants enrolled before adaptation continue to accrue follow-up afterward. Outside the clinical-trial literature, there is considerable interest in post-selection and post-model-selection inference, and methods such as data carving and randomized selection have been developed for those problems (Fithian et al. 2014; Kuchibhotla et al. 2022; Lee et al. 2016; Tian and Taylor 2018). Recent econometric, operations-research, and bandit literatures study inference after data-dependent decisions, including selected winners (Andrews et al. 2024), policy evaluation under adaptive allocation (Hadad et al. 2021), batched bandits (Zhang et al. 2020), asymptotic representations for adaptive experiments (Hirano and Porter 2025), and optimal inference conditional on the realized design of a batched adaptive experiment (Chen and Andrews 2026). Further work is needed to extend these methods to adaptive clinical trials.

Conditioning is not costless. Rare selection events can produce unstable selective analyses, and conditional confidence intervals can be wide or unbounded (Kivaranovic and Leeb 2021); conditioning on more information than was used for selection can waste power (Fithian et al. 2014). If marginal inference over the whole hypothesis family is the scientific target, a suitably constructed unconditional procedure can dominate a conditional procedure (Goeman and Solari 2024); the selective test's advantage in Tables 1 and 2 does not contradict that result. Nevertheless, the statistical problem

should be defined before methods are ranked. If selection-conditional error is the relevant risk for a decision, its computational, power, and implementation challenges are design problems to be addressed rather than reasons to substitute a different guarantee.

Data and code availability

No patient-level or empirical data were used. Base R code, reproducible simulation outputs, numerical validation records, and figures accompany this manuscript as supplementary materials.

Funding

No specific funding was received for this work.

Conflict of interest

The authors declare that they have no conflicts of interest.

Generative AI disclosure

OpenAI GPT-5.6 Sol (runtime model identifier `openai/gpt-5.6-sol`, maximum reasoning setting) was used on 13–24 July 2026, under human direction, to review meeting transcripts and source materials, synthesize literature, audit and rewrite simulation code, draft the manuscript and earlier response letters, and support technical and editorial review. The accompanying supplementary document, *Transparent record of the human–AI workflow*, provides a detailed account of this process. The authors take full responsibility for the manuscript's claims, citations, results, and disclosures.

References

Andrews, I., T. Kitagawa, and A. McCloskey (2024). "Inference on Winners". In: *The Quarterly Journal of Economics* 139.1, pp. 305–358. DOI: 10.1093/qje/qjad043.

Bauer, P. and K. Köhne (1994). "Evaluation of Experiments with Adaptive Interim Analyses". In: *Biometrics* 50.4, pp. 1029–1041. DOI: 10.2307/2533441.

Bretz, F., F. König, W. Brannath, E. Glimm, and M. Posch (2009). "Adaptive Designs for Confirmatory Clinical Trials". In: *Statistics in Medicine* 28.8, pp. 1181–1217. DOI: 10.1002/sim.3538.

Bretz, F., H. Schmidli, F. König, A. Racine, and W. Maurer (2006). "Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts". In: *Biometrical Journal* 48.4, pp. 623–634. DOI: 10.1002/bimj.200510232.

Burdon, A., T. Jaki, X. Chen, P. Mozgunov, H. Zheng, and R. Baird (2026). "Modern Clinical Trials: Seamless Designs and Master Protocols". In: *Cancer Medicine* 15.5, e71858. DOI: 10.1002/cam4.71858.

Chen, J. and I. Andrews (2026). *Optimal Conditional Inference in Adaptive Experiments*. arXiv preprint arXiv:2309.12162. Version 2; first posted in 2023. DOI: 10.48550/arXiv.2309.12162.

Dunnett, C. W. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control". In: *Journal of the American Statistical Association* 50.272, pp. 1096–1121. DOI: 10.1080/01621459.1955.10501294.

Dunnett, C. W. and A. C. Tamhane (1991). "Step-Down Multiple Tests for Comparing Treatments with a Control in Unbalanced One-Way Layouts". In: *Statistics in Medicine* 10.6, pp. 939–947. DOI: 10.1002/sim.4780100614.

Fithian, W., D. L. Sun, and J. Taylor (2014). *Optimal Inference After Model Selection*. arXiv preprint arXiv:1410.2597. Version 4.

Freidling, T., Q. Zhao, and Z. Gao (2026). "Selective Randomization Inference for Adaptive Experiments". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology*. Advance article, qkag081. DOI: 10.1093/jrsssb/qkag081.

Friede, T., N. Parsons, and N. Stallard (2012). "A Conditional Error Function Approach for Subgroup Selection in Adaptive Clinical Trials". In: *Statistics in Medicine* 31.30, pp. 4309–4320. DOI: 10.1002/sim.5541.

Friede, T., N. Parsons, N. Stallard, S. Todd, E. Valdes Marquez, J. Chataway, and R. Nicholas (2011). "Designing a Seamless Phase II/III Clinical Trial Using Early Outcomes for Treatment Selection: An Application in Multiple Sclerosis". In: *Statistics in Medicine* 30.13, pp. 1528–1540. DOI: 10.1002/sim.4202.

Friede, T. and N. Stallard (2008). "A Comparison of Methods for Adaptive Treatment Selection". In: *Biometrical Journal* 50.5, pp. 767–781. DOI: 10.1002/bimj.200710453.

Goeman, J. J. and A. Solari (2024). "On Selection and Conditioning in Multiple Testing and Selective Inference". In: *Biometrika* 111.2, pp. 393–416. DOI: 10.1093/biomet/asad078.

Hadad, V., D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey (2021). "Confidence Intervals for Policy Evaluation in Adaptive Experiments". In: *Proceedings of the National Academy of Sciences* 118.15, e2014602118. DOI: 10.1073/pnas.2014602118.

Hirano, K. and J. R. Porter (2025). *Asymptotic Representations for Sequential Decisions, Adaptive Experiments, and Batched Bandits*. arXiv preprint arXiv:2302.03117. Version 2. DOI: 10.48550/arXiv.2302.03117.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (Nov. 2019). *Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1)*. Final version adopted 20 November

2019. URL: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (June 2025). *ICH E20: Adaptive Designs for Clinical Trials*. Draft version endorsed 25 June 2025 at Step 2; EMA Step 2b document released for public consultation 30 June 2025. URL: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e20-guideline-adaptive-designs-clinical-trials-step-2b_en.pdf.

Kivaranovic, D. and H. Leeb (2021). "On the Length of Post-Model-Selection Confidence Intervals Conditional on Polyhedral Constraints". In: *Journal of the American Statistical Association* 116.534, pp. 845–857. DOI: 10.1080/01621459.2020.1732989.

König, F., W. Brannath, F. Bretz, and M. Posch (2008). "Adaptive Dunnett Tests for Treatment Selection". In: *Statistics in Medicine* 27.10, pp. 1612–1625. DOI: 10.1002/sim.3048.

Kuchibhotla, A. K., J. E. Kolassa, and T. A. Kuffner (2022). "Post-Selection Inference". In: *Annual Review of Statistics and Its Application* 9, pp. 505–527. DOI: 10.1146/annurev-statistics-100421-044639.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). "Exact Post-Selection Inference, with Application to the Lasso". In: *The Annals of Statistics* 44.3, pp. 907–927. DOI: 10.1214/15-AOS1371.

Li, E., N. Stallard, E. Glimm, D. Magirr, and P. K. Kimani (2026). *Confidence Intervals for Two-Stage Adaptive Designs with Subpopulation Selection*. arXiv preprint arXiv:2603.13583. Version 1. DOI: 10.48550/arXiv.2603.13583.

Magirr, D., T. Jaki, and J. Whitehead (2012). "A Generalized Dunnett Test for Multi-Arm Multi-Stage Clinical Studies with Treatment Selection". In: *Biometrika* 99.2, pp. 494–501. DOI: 10.1093/biomet/ass002.

Marcus, R., E. Peritz, and K. R. Gabriel (1976). "On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance". In: *Biometrika* 63.3, pp. 655–660. DOI: 10.1093/biomet/63.3.655.

Müller, H.-H. and H. Schäfer (2001). "Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches". In: *Biometrics* 57.3, pp. 886–891. DOI: 10.1111/j.0006-341X.2001.00886.x.

Park, J. J. H., E. Siden, M. J. Zoratti, L. Dron, O. Harari, J. Singer, R. T. Lester, K. Thorlund, and E. J. Mills (2019). "Systematic Review of Basket Trials, Umbrella Trials, and Platform Trials: A Landscape Analysis of Master Protocols". In: *Trials* 20.1, p. 572. DOI: 10.1186/s13063-019-3664-1.

Robertson, D. S., K. M. Lee, B. C. López-Kolkovska, and S. S. Villar (2023). "Response-Adaptive Randomization in Clinical Trials: From Myths to Practical Considerations". In: *Statistical Science* 38.2, pp. 185–208. DOI: 10.1214/22-STS865.

Rosenblum, M. (2013). "Confidence Intervals for the Selected Population in Randomized Trials That Adapt the Population Enrolled". In: *Biometrical Journal* 55.3, pp. 322–340. DOI: 10.1002/bimj.201200080.

Sampson, A. R. and M. W. Sill (2005). "Drop-the-Losers Design: Normal Case". In: *Biometrical Journal* 47.3, pp. 257–268. DOI: 10.1002/bimj.200410119.

Sato, F., K. Uemura, J. Mizusawa, Y. Matsuyama, and Y. Morikawa (2026). "Adaptive Seamless Phase II/III Randomization Test Considering Treatment Group Selection Based on Short-Term Binary Outcomes". In: *Statistics in Medicine* 45.3–5, e70400. DOI: 10.1002/sim.70400.

Sill, M. W. and A. R. Sampson (2009). "Drop-the-Losers Design: Binomial Case". In: *Computational Statistics & Data Analysis* 53.3, pp. 586–595. DOI: 10.1016/j.csda.2008.07.031.

Stallard, N. and S. Todd (2003). "Sequential Designs for Phase III Clinical Trials Incorporating Treatment Selection". In: *Statistics in Medicine* 22.5, pp. 689–703. DOI: 10.1002/sim.1362.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. (1949). *The American Soldier, Volume 1: Adjustment During Army Life*. Princeton, NJ: Princeton University Press.

Takahashi, K., R. Ishii, K. Maruo, and M. Gosho (2022). "Statistical Tests for Two-Stage Adaptive Seamless Design Using Short- and Long-Term Binary Outcomes". In: *Statistics in Medicine* 41.21, pp. 4130–4142. DOI: 10.1002/sim.9500.

Tian, X. and J. E. Taylor (2018). "Selective Inference with a Randomized Response". In: *The Annals of Statistics* 46.2, pp. 679–710. DOI: 10.1214/17-AOS1564.

U.S. Food and Drug Administration (Nov. 2019). *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry*. URL: https://www.fda.gov/media/78495/download.

Zhang, K. W., L. Janson, and S. A. Murphy (2020). "Inference for Batched Bandits". In: *Advances in Neural Information Processing Systems 33*, pp. 9818–9829.

A Additional numerical results

The effect grid uses $\theta = (\theta_1, 0, 0, 0, -0.02)$ with $\theta_1 = 0, 0.05, \ldots, 0.40$. The principal value $\theta_1 = 0.25$ uses two million repetitions; each other grid value uses 10,000,000 repetitions. Figure 1 reports selection-conditional and joint power, while Figure 2 contrasts marginal FWER (solid curves) with selection-conditional FWER in the exchangeable boundary-null strata (dotted curves). Across displayed grid points, marginal-FWER Monte Carlo standard errors are at most 0.00049, power standard errors are at most 0.00132, and selection-conditional-FWER standard errors are at most 0.0059. Every displayed pooled conditional stratum contains at least 4,059 repetitions, and no grid point is omitted.

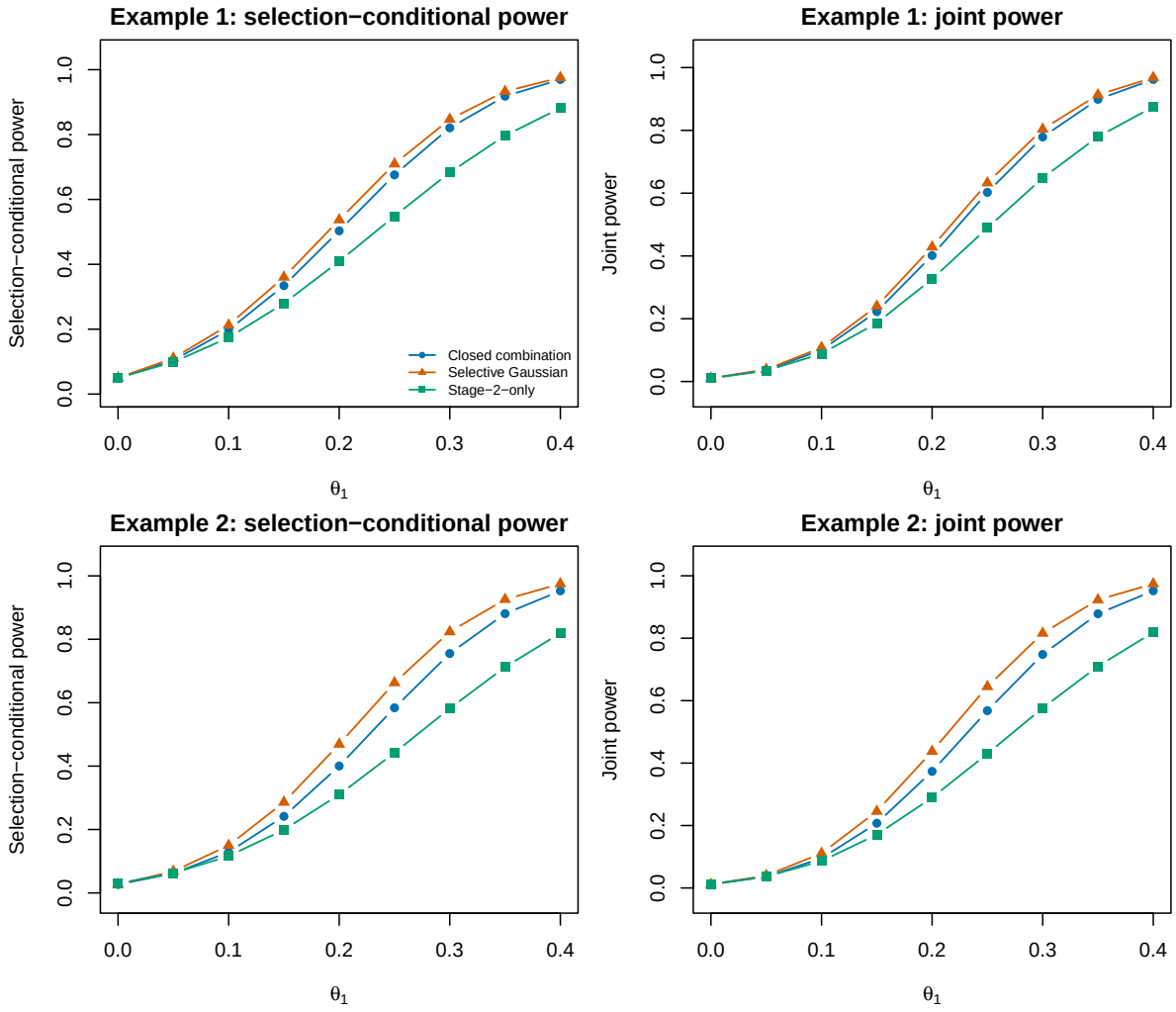

Figure 1: Selection-conditional and joint power across the effect grid for Example 1 (top row) and Example 2 (bottom row). In Example 2, the legend labels "Selective Gaussian" and "Stage-2-only" abbreviate Selective Gaussian–Holm and Stage-2-only Dunnett, respectively. At $\theta_1 = 0$, the arm-1 ordinate is a rejection probability rather than power.

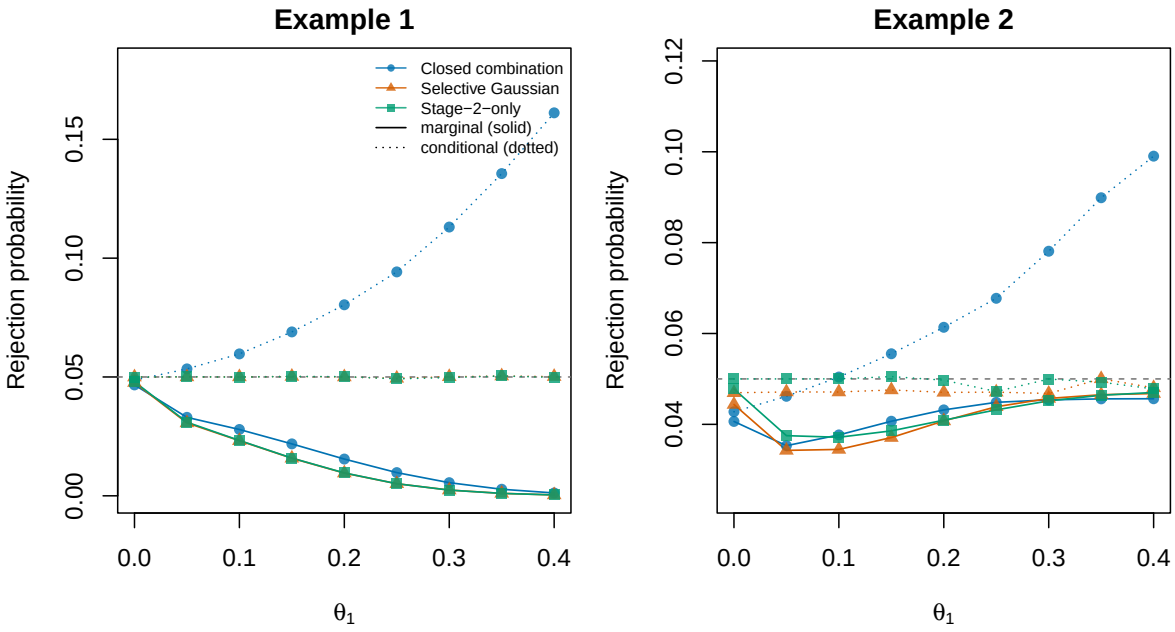

Figure 2: Marginal FWER (solid) and selection-conditional FWER (dotted) across the effect grid. In Example 2, the legend labels "Selective Gaussian" and "Stage-2-only" abbreviate Selective Gaussian–Holm and Stage-2-only Dunnett, respectively. In Example 1, the conditional stratum pools $S_1 = \{2\}, \{3\}, \{4\}$; in Example 2, it pools the three selected sets contained in $\{2, 3, 4\}$. At $\theta_1 = 0$, treatment 1 is a true null, whereas it is an alternative at positive grid values; the connected solid curve therefore crosses a change in the true-null set. The horizontal line marks the illustrative level 0.05. The two panels use different vertical scales.